Dallah: A Dialect-Aware Multimodal Large Language Model for Arabic

Fakhraddin Alwajih   **Gagan Bhatia**
Muhammad Abdul-Mageed
The University of British Columbia & Invertible AI
{muhammad.mageed@}ubc.ca;invertible.ai

Abstract

Recent advancements have significantly enhanced the capabilities of Multimodal Large Language Models (MLLMs) in generating and understanding image-to-text content. Despite these successes, progress is predominantly limited to English due to the scarcity of high-quality multimodal resources in other languages. This limitation impedes the development of competitive models in languages such as Arabic. To alleviate this situation, we introduce an efficient Arabic multimodal assistant, dubbed *Dallah*, that utilizes an advanced language model based on LLaMA-2 to facilitate multimodal interactions. *Dallah* demonstrates state-of-the-art performance in Arabic MLLMs. Through fine-tuning six Arabic dialects, *Dallah* showcases its capability to handle complex dialectal interactions incorporating both textual and visual elements. The model excels in two benchmark tests: one evaluating its performance on Modern Standard Arabic (MSA) and another specifically designed to assess dialectal responses. Beyond its robust performance in multimodal interaction tasks, *Dallah* has the potential to pave the way for further development of dialect-aware Arabic MLLMs.

1   Introduction

Large language models (LLMs) have revolutionized how machines understand and generate human language. Recent developments have significantly expanded the scope of these models by integrating multimodal data, enabling sophisticated interaction with both textual and visual information. Despite these advances, applying NLP in linguistically diverse environments presents unique challenges, particularly in processing dialectal variations of languages and the integration of these variations within multimodal contexts. These challenges are especially pronounced in Arabic, a collection of languages and varieties characterized by a rich tapestry of dialects that vary significantly

Figure 1: Map highlighting the countries targeted by *Dallah* for dialectal Arabic dataset construction.

across different regions.

Arabic dialects enrich the cultural landscape and present complex linguistic variations that standard NLP models, primarily designed for MSA, often fail to solve. This linguistic diversity requires the development of specialized models that can navigate the rich world of dialectal Arabic and its integration with visual data. Addressing these needs is crucial for improving user interaction and preserving linguistic heritage, as many dialects are underrepresented. Some may be at risk of diminishing in the face of globalization and cultural homogenization.

Multi-cultural and multi-modal LLMs (Alwajih et al., 2024; Huang et al., 2023; Sengupta et al., 2023) are vital against cultural homogenization, where globalization tends to favour dominant languages like Arabic. This can potentially lead to the marginalization or even extinction of less widely spoken dialects (Barnet and Cavanagh, 2014; Ahmedov and Shaxrizoda, 2024; Fahmi and Liska, 2024). By including these low-resource dialects, we can ensure their continued relevance and preserve the rich linguistic diversity of the Arabic-speaking world.

To this end, we introduce a powerful multimodal language model that specifically targets the unique aspects of Arabic dialects. Our model, dubbed *Dallah*, is built on the foundations of LLaVA (Liu et al., 2023b), an advanced multimodal language model

framework. We enhance LLaVA with the linguistic capabilities of AraLLaMA (Alwajih et al., 2024), an LLM proficient in Arabic and English. *Dallah* is designed to understand and generate content in Arabic, navigating the complex interplay between different dialects and visual information effectively.

The following are the main contributions of our work:

1. We present *Dallah*, which combines the robust multimodal processing power of LLaVA with the dialectal versatility of AraLLaMA, creating a model uniquely equipped to handle the linguistic and visual challenges presented by Arabic dialects.

2. We introduce a novel data filtering method that optimizes the selection and usage of training data, ensuring that *Dallah* is fine-tuned with high-quality, relevant multimodal datasets that reflect the linguistic diversity found within the Arab world.

3. *Dallah* supports wide dialectal coverage, successfully fine-tuning over six major Arabic dialects using limited but highly representative dialectal data.

4. We introduce *Dallah-Bench* evaluation benchmark for Arabic dialects tailored to assess the efficacy of multimodal language models in real-world applications that require an understanding of dialectal variations.

5. We have also built an understanding of which model from the set *{GPT4, GPT4-Turbo, Command-R+}* is best suited for evaluating MSA and dialectal data compared to Human evaluation.

The remainder of this paper is structured as follows: In Section 2, we provide an overview of related work. Section 3 details our methodology, including the processes for Arabic dataset translation and filtering, construction of dialectal Arabic datasets, the architecture of *Dallah*, and the training procedures employed. In Section 4, we describe our implementation details and the benchmarks used for evaluation. Section 5 presents our experimental results, including both quantitative and qualitative analyses. We conclude in Section 6 with a discussion of our findings and future work.

2 Related Work

2.1 Large Language Models

Recent progress in NLP has been driven by advances in LLMs, starting with the foundational Transformer model (Vaswani et al., 2017). This innovation paved the way for language models like the encoder-based BERT (Devlin et al., 2018), and the decoder-based Generative Pre-trained Transformer (GPT) (Brown et al., 2020), as well as encoder-decoder-based models like T5 (Raffel et al., 2020), which have significantly improved linguistic understanding and performance in complex language processing tasks. The development of models such as OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023a), LLaMA-2 (Touvron et al., 2023b), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), and ChatGPT (OpenAI, 2023) Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), Phi-2 (Javaheripi et al., 2023), Phi-3 (Abdin et al., 2024), and instruction-tuned variants of LLaMA-2 like Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023), have demonstrated the rapid evolution of the field. These models benefit from extensive training on large datasets and tailored instruction sets, enhancing their effectiveness.

Arabic LLMs Building on the global momentum, the scope of LLMs has extended into Arabic language processing. The introduction of Jasmine (Nagoudi et al., 2023) marked a significant milestone, followed by AceGPT (Huang et al., 2023) and Jais (Sengupta et al., 2023), which have enhanced Arabic conversational AI. Recently, AraLLaMA (Alwajih et al., 2024) set a new standard with its proficiency in the Egyptian Arabic dialect, showcasing the flexibility of LLMs in handling linguistically diverse data.

2.2 Multimodal Large Language Models

The integration of computer vision and natural language processing has given rise to Visual Language Models (VLMs). These models merge visual and linguistic data, enhancing tasks that require visual perception and language abilities. Models like CLIP (Radford et al., 2021) bridge the gap between visual recognition and language tasks, demonstrating the effectiveness of cross-modal applications.

Recent advancements show that LLMs improve VLMs. Innovations such as Flamingo (Alayrac et al., 2022), Blip-2 (Li et al., 2023), and

LLaVA (Liu et al., 2023b) have leveraged large image-text pair datasets, enhancing cross-modal coordination and learning efficiency. These models also employ specialized architectural features for better integration. For instance, Flamingo utilizes a perceiver resampler to integrate visual data, and Blip-2 introduces a Q-Former (Li et al., 2023) for aligning visual and language modalities. LLaVA (Liu et al., 2023b) adjusts a linear projection layer to synchronize vision and language modalities. Meanwhile, LLaVA1.6 (Liu et al., 2023a) incorporates extensive instruction tuning and a high-resolution vision encoder, achieving outstanding results across multiple benchmarks.

Arabic Multimodal LLMs In Arabic NLP, Peacock (Alwajih et al., 2024) represents the first work in Arabic-centric MLLM capable of handling Arabic multimodal interaction effectively. Additionally, the multilingual PALO (Maaz et al., 2024) has demonstrated the ability to process and integrate multiple languages, including Arabic, into multimodal contexts.

2.3 Multimodal Instruction Tuning Datasets

Development of MLLMs typically involves two phases. The first phase focuses on aligning visual and linguistic features, utilizing datasets such as COCO (Lin et al., 2014), LLaVA-Pretrain (Liu et al., 2023b), and Laion (Schuhmann et al., 2022). The subsequent visual instruction fine-tuning phase enhances the models' capabilities to follow complex multimodal instructions. This phase often involves transforming existing datasets into more conversationally relevant formats using advanced LLMs such as GPT-4, as seen in models like LLaVA-Instruct (Liu et al., 2023b) and SVIT (Liu et al., 2024). Recent works utilized GPT-4V (OpenAI, 2023) to generate new captions and question-answers, such as in ShareGPT4V (Chen et al., 2023), LVIS-instruct4v (Wang et al., 2023) and Allava (Chen et al., 2024). Arabic MLLMs utilized translated versions of LLaVA-Instruct (Alwajih et al., 2024; Maaz et al., 2024) using different tools for translation.

3 Methodology

3.1 Arabic Dataset Translation and Filtering

In the first step, we aimed to build an Arabic MLLM using Arabic datasets. A major obstacle facing Arabic MLLMs is the lack of resources.

This lack is largely due to the challenges of sourcing relevant Arabic image-text pairs on a large scale. To bridge this resource gap, we have implemented a careful translate-and-filter pipeline consisting of a translation stage and a filtering stage inspired by (Mohamed et al., 2023; Alwajih et al., 2024). This pipeline converts publicly available, English-centric image-text and visual instruction datasets into Arabic while maintaining data quality and preventing error propagation due to translation.

We utilize the latest version of the Google Translate API (Google Cloud) for **the translation stage** which is the best translation method as shown by (Zhu et al., 2023). We also conducted back translation as required by the subsequent filtering stage. During **the filtering stage**, we ensure the quality of our translations by employing a sentence embedding model (Meng et al., 2024; Wang et al., 2024). We assess the quality by calculating the similarity of embedding between the original and back-translated sentences for both question and answer pairs, retaining only those translations that meet our quality standards. Essentially, we keep examples with questions and answers above a predefined threshold, which we have empirically set at 80%. Figure 3 illustrates the translation and filtering process. Unlike the methods used in (Mohamed et al., 2023; Alwajih et al., 2024), we employ an English sentence embedding model based on Mistral-7b (Wang et al., 2024) to calculate the similarities. This last model is more powerful than embedding models used in Mohamed et al. (2023); Alwajih et al. (2024) as shown by MTEB leaderboard (Muennighoff et al., 2022). Refer to Figure 2 for examples illustrating our pipeline's effectiveness.

3.2 Dialectal Arabic Dataset Construction

Due to the absence of Arabic dialectal data tailored for vision tasks, we randomly select six subsets from our translated LLaVA-instruct 150k dataset. We aim to ensure that each subset included diverse content types, such as conversations, complex reasoning, and detailed descriptions, from the LLaVA-instruct dataset. Our goal is to capture the dialects of Egypt, Mauritania, Morocco, Palestine, Saudi Arabia, and Yemen. These dialects represent a broad spectrum of Arabic dialects.

These subsets are then assigned to native professional translators from the aforementioned countries in order to translate them from MSA into their respective dialects. Table 1 displays the number of

Original	Arabic Translation	English Back Translation	Similarity
What are the surfers holding?	ماذا يحمل متصفحي؟	What does my browser load?	0.513672
The jumper in the image is a snowboarder.	الطائر الموجود في الصورة هو متزلج على الجليد.	The bird in the photo is an ice skater.	0.549805
Where is the surfer located?	أين يقع مقر سيرفر؟	Where is Server's headquarters?	0.580566
What is the condition of the street where the stop sign is located?	ما هي حالة الشارع الذي توجد فيه إشارة التوقف؟	What is the condition of the street where the stop sign is located?	1.00000
The kitchen counter is cluttered with various items, contributing to the messy appearance of the kitchen.	ما هي حالة الشارع الذي توجد فيه إشارة التوقف؟	What is the condition of the street where the stop sign is located?	0.900391
The kitchen counter is cluttered with various items, contributing to the messy appearance of the kitchen.	نعم، توجد تعليقات معلقة على الهاتف القابل للطي، مما يضيف لمسة من التخصيص والأناقة إلى الجهاز.	Yes, there are suspensions on the foldable phone, adding a touch of personalization and elegance to the device.	0.800293

Figure 2: This figure illustrates the translation and filtering process used in constructing the Arabic dataset for *Dallah*. The red rows represent examples that were removed due to low similarity scores between the original English text and the back-translated English text. The green rows show the retained examples that met the similarity threshold, ensuring high-quality translations for effective model training.
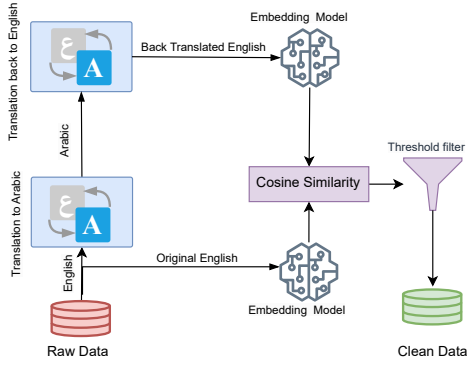

Figure 3: Illustration of the translation and filtering process for constructing high-quality Arabic multi-modal datasets. Examples illustrating the results of this pipeline are in Figure 2.

samples per country, while Figure 1 illustrates the targeted countries on the map. Refer to A.2 for more details.

Stage	Source	#Sample
Pretraining	LLaVA-Pretrain LCS (Arabic+English)	800k
Inst. Tuning	LLaVA-Instruct English	150k
	LLaVA-Instruct Arabic	139k
Dialectal Tuning	Egypt	738
	Mauritania	495
	Morocco	505
	Palestine	853
	Saudi Arabia	784
	Yemen	604
	Total	1.1 M

Table 1: Number of samples used for each stage in the training of *Dallah*.

3.3 Architecture

The *Dallah* model follows the structure of LLaVA1.5 (Liu et al., 2023a) and comprises three key elements:

1. **Vision Encoder:** The vision encoder (V_φ) employs the CLIP-Large model (Radford et al., 2021) to process input images (X) into 576 visual tokens at a resolution of 336x336 with a patch size of 14, producing a sequence of patch features $V = \{v_j \in \mathbb{R}^{d_x}\}_{i=j}^M$.

2. **Projector:** A connector (P_ϕ), designed as a two-layer multi-layer perceptron (MLP), maps the visual patch sequences $\{v_j\}_{j=1}^M$ to the text embedding space $\{h_j\}_{j=1}^M$, allowing for effective integration between the pre-trained LLM and the vision encoder.

3. **Language Model (LLM):** The Arabic LLM (F_θ), based on AraLLaMA (Alwajih et al., 2024)[1] processes sequences of text embeddings $\{h_i\}_{i=0}^{N-1}$ in the d-dimensional space, outputting corresponding next predictions $\{h_i\}_{i=1}^N$. A tokenizer and embedding module maps text sequences $\{y_i\}_{i=0}^{N-1}$ to the embedding space and back to output text sequences $\{y_i\}_{i=1}^N$.

This structure equips the model to handle various multimodal understanding tasks, taking an image and instruction text sequence as input and generating a text sequence as output. Figure 4 illustrates *Dallah* architecture.

---

[1]AraLLaMA is an enhanced version of LLaMA-2 (Touvron et al., 2023a).
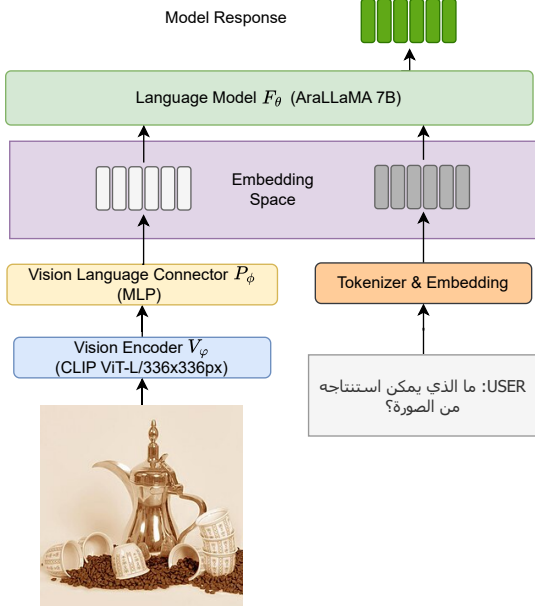
Figure 4: *Dallah* model architecture, showcasing the integration of the vision encoder, projector, and language model.

3.4 Training

Training of **Dallah** consists of three stages: (i) pre-training using data LLaVA-Pretrain (MAS Arabic and English), (ii) visual instruction supervised fine-tuning using LLaVA-Instruct (Arabic MSA and English), and (iii) further visual instruction supervised fine-tuning using dialectal data. Table 1 details the datasets used in each stage. Training data comprises pairs of images and text (X, Y), with the text sequence Y formatted as a single-turn in the pre-training phase and multi-turn conversation in the visual Instruction Supervised Fine-tuning stage. $Y = (Y_q^1, Y_a^1, \ldots, Y_q^T, Y_a^T)$. Here, T represents the number of conversation turns, Y_q^t the user's prompt, and Y_a^t the model's response.

3.4.1 Pre-training

During this phase, the goal is to enhance the alignment between the vision and text data within the embedding space. For this, image-caption style data (X, Y_a) is extracted from the conversation, where X is the image and Y_a is a corresponding text description. The probability of generating Y_a given the image is calculated as:

$$p(Y_a|X) = \prod_{i=1}^{N_a} F_\theta(y_i | P_\phi \circ V_\varphi(X)),$$

where N_a is the length of Y_a. The training objective is to maximize the log-likelihood of Y_a autoregres-
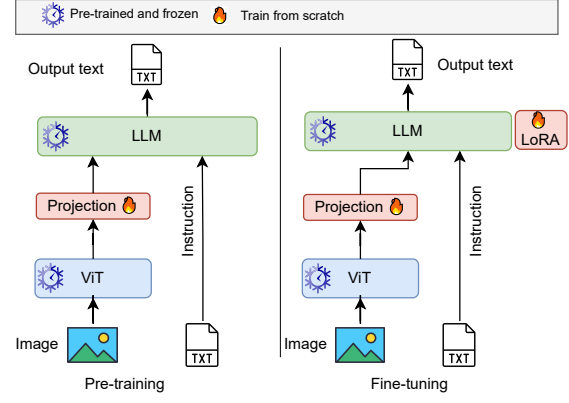
sively:

$$\max_\phi \sum_{i=1}^{N_a} \log F_\theta(y_i | P_\phi \circ V_\varphi(X)),$$

This framework permits the adjustment of learnable parameters of projector layers during pre-training. LLM and vision encoder learnable parameters are frozen during pre-training, as shown in Figure 5.

3.4.2 Visual Instruction Supervised Fine-tuning

The full image-text pairs (X, Y) in their multi-turn conversation format are used for fine-tuning. The set of tokens corresponding to the model's responses is denoted as $A = \{y | y \in Y_a^t, \text{for any } t = 1, \ldots, T\}$. The training objective is to maximize the log-likelihood of the model's responses autoregressively:

$$\max_{\acute{\theta},\phi} \sum_{i=1}^{N} I(y_i \in A) \log F_\theta(y_i | P_\phi \circ V_\varphi(X)),$$

where N is the total number of tokens in Y, $\acute{\theta}$ is a subset of θ, and $I(y_i \in A)$ is 1 if y_i belongs to A, and 0 otherwise. Training the projector layers' learnable parameters while the vision encoder is kept frozen during this phase. For LLM, we utilize LoRA (Hu et al., 2021) to train the LLM. Figure 5 visualizes both pre-training and visual instruction supervised fine-tuning stages.

4 Experiments

4.1 Implementation Details

Model Configurations. In this work, we develop the *Dallah* model based on the LLaVA1.5 framework. Specifically, we employ the CLIP-ViT-L/14

Figure 5: Training schema for *Dallah*, detailing the pre-training and visual instruction supervised fine-tuning phases.

as the visual encoder with a standard resolution of 336×336. We also use AraLLaMA, a language model tailored specifically for Arabic, and employ a two-layer MLP as the interface to connect the visual encoder with the LLM.

For the construction of *Dallah*, a three-stages training process is implemented. Initially, we establish a base MSA MLLM in two stages pretraining and MSA fine-tuning, followed by an adaptation stage for Arabic dialects. The following are details of these different stages:

Stage 1: Pre-training stage. During this initial stage, training is conducted for a single epoch only on the projector as described in 3.4.1 using the translated LCS-558 (Liu et al., 2023b) dataset. This dataset includes data filtered for Arabic and 300K samples of English samples. The optimization is done using the AdamW optimizer with a learning rate of 1×10^{-3}, combined with a cosine learning rate schedule. The overall batch size is maintained at 32. This phase required approximately 28 hours of training on a single A100 GPU.

Stage 2: Instruction-tuning stage. In this satge, the visual encoder is frozen, tuning is applied to the visual projector, and the LLM is fine-tuned using LoRA as described in 3.4.2. Here we employ the 150K LLaVA-Instruct dataset in English alongside a translated and filtered Arabic version. The learning rate is set to 2×10^{-4} with a batch size of 8, maintaining the same settings as in the first stage. This training phase took around 58 hours using a single A100 GPU.

Stage 3: Dialectal instruction-tuning stage. This stage is similar to stage 2, but is focused on dialectal data for six different Arabic dialects and parallel data for MSA. The settings remain the same as in the second stage with learning rate 2×10^{-5}, over five epochs. This training phase took approximately 2 hours using a single A100 GPU. Table 1 details the data used in the aforementioned stages.

4.2 Benchmarks

We evaluate our model using two benchmarks: LLaVA-Bench for MSA evaluation and comparison with counterpart Arabic MLLMs and *Dallah-Bench* to assess the model's capabilities in six Arabic dialects.

4.2.1 LLaVA-Bench for MSA

The Ara-LLaVA-Bench is the Arabic version of the LLaVA-Bench, translated using Google API
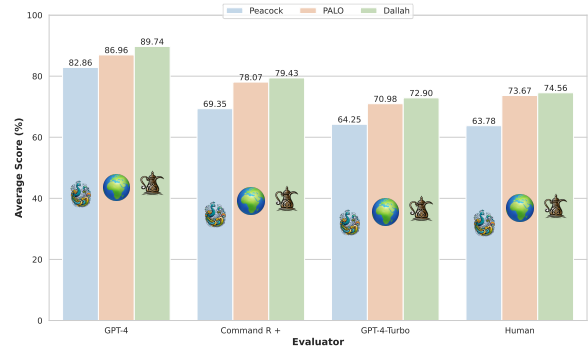

Figure 6: Average score comparison by evaluator and model.

and reviewed by Arabic native annotator. The LLaVA-Bench includes 30 images selected from the COCO2014 validation dataset. Each image is accompanied by three types of questions: conversion, detailed descriptions, and complex reasoning, amounting to 90 questions.

4.2.2 Dallah-Bench for Dialects

We select a subset of 20 questions from the Henna (Alwajih et al., 2024) dataset to assess the model's response to dialects, naming it *Dallah-Bench*. We task native professionals from each dialect to translate these from MSA into the aforementioned six Arabic dialects.

5 Results

5.1 LLaVA-Bench for MSA

Evaluator	Model	Arch.	CC	DD	CR	Avg
GPT-4	Peacock		85.66	80.25	82.52	82.86
	PALO		87.30	86.31	87.24	86.96
	Dallah		**89.64**	**90.79**	**88.80**	**89.74**
Command R +	Peacock		74.49	66.54	66.78	69.35
	PALO		**76.61**	75.32	82.15	78.07
	Dallah		75.37	**77.1**	**85.78**	**79.43**
GPT-4-Turbo	Peacock		67.35	62.5	62.86	64.25
	PALO		69.2	70.02	73.76	70.98
	Dallah		**71.47**	**70.75**	**76.52**	**72.9**
Human	Peacock		68.56	61.00	61.78	63.78
	PALO		75.11	67.78	**78.11**	73.67
	Dallah		**78.56**	**69.44**	75.67	**74.56**

Table 2: Evaluation of Arabic LLaVA-Bench in MSA using four different evaluators: GPT-4, GPT-4-Turbo, Cohere Command R+, and Human. We consider Human Evaluation to be the gold standard. CC: Conversations, DD: Details, CR: Complex Reasoning.

We evaluate *Dallah* using the LLaVA-Bench

benchmark described in 4.2.1, specifically designed for multimodal models. We compare *Dallah* against two baselines: Peacock, an Arabic MLLM based on the InstructBlip architecture integrated with AraLLaMA, an LLM based on LLaMA-2; and PALO, a multilingual MLLM based on the LLaVA architecture integrated with Vicuna, an LLM based on LLaMA-2. PALO supports Arabic, along with nine other languages. We utilize both model-based evaluation and human evaluation. We describe each of these next.

Model Evaluation. We follow the original methodology of LLaVA-Bench (Liu et al., 2023b) by calling the APIs of three different models: GPT-4, Cohere Command R+ [2], and GPT-4-Turbo. We slightly modify the prompts to accommodate Arabic instead of English.

Human Evaluation. We conduct a human evaluation of three models using LLaVA-Bench. We present three well-educated annotators with images and questions-answer pairs generated by *Dallah* and two other baseline models, Peacock and PALO. To ensure integrity, the names of the models are hidden throughout the evaluation process. Annotators are asked to rate the models' responses on a scale from 1 to 10, based on *correctness*[3], *helpfulness*[4] and question-answer *consistency*[5].

Figure 6 presents the average scores (%) of the models Peacock, PALO, and Dallah as evaluated by four different evaluators: GPT-4, Command R+, GPT-4-Turbo, and Human.

5.1.1 Analysis

We report the main results in Table 2. In the GPT-4 evaluation, the scale of scores is higher than in other evaluations. The overall scores for Cohere Command R+ and GPT-4-Turbo are close to those of the human evaluation, with GPT-4-Turbo being the closest numerically to human evaluation.

From Table 2, it is observed that the *Dallah* model outperforms the baseline models in most dimensions of the LLaVa-Bench across all evaluation methods. Peacock generally showed the lowest performance, which could be attributed to multiple factors, including the scale of training data and

---

[3]**Correctness**: The accuracy and factual correctness of the model's response to the given question.

[4]**Helpfulness**: The degree to which the model's response provides useful and informative assistance to the user.

[5]**Consistency**: The coherence and logical flow within the model's responses, ensuring they are free from contradictions.

---

the model architecture, where the best model in the Peacock suite is based on InstructBlip and was trained with frozen Q-former components.

Dallah and PALO show close results across all evaluations, with *Dallah* having a slight advantage. *Dallah* surpasses PALO by an average of 1.74% across all evaluations. Both models share the same LLaVA architecture but differ in their LLMs; *Dallah* uses AraLLaMA, an Arabic LLM, giving it an advantage, whereas PALO utilizes Vicuna, a multilingual LLM based on LLaMa-2 (Touvron et al., 2023a). The training data are almost identical, though *Dallah*'s data went through a careful filtering method to ensure quality. Also, *Dallah* is trained on high-quality human-translated dialectal data.

These results demonstrate *Dallah*'s effectiveness in MSA, exhibiting strong reasoning capabilities and substantial knowledge.

5.2 Dallah-Bench for Dialects

Evalutator	Command R+		GPT-4-Turbo		Human	
Country	DA	CA	DA	CA	DA	CA
Egypt	7.82	8.04	7.00	5.96	6.59	7.22
Mauritania	7.11	7.86	3.59	5.04	4.41	6.36
Morocco	8.82	8.59	7.54	6.68	6.50	5.27
Palestine	8.00	8.32	5.32	6.36	8.73	7.68
Saudi	8.50	8.91	5.77	6.46	7.50	8.27
Yemen	8.36	9.00	5.04	4.77	7.49	7.73
Average	8.10	8.45	5.71	5.88	6.87	7.09

Table 3: Evaluation of *Dallah* model on dialect bench using three evaluators: Cohere Command R+, GPT4-Turbo, and Humans from respective countries. **DA**: Dialect Authenticity, **CA**: Content Accuracy. Prompts for the evaluation can be found in Figure 7.

We assess *Dallah*'s performance in dialects using *Dallah-Bench* described in 4.2.2. We employ two types of evaluations: human evaluation and model-based evaluation.

Human Evaluation. To evaluate the model responses related to dialect questions about the image, we ask native speakers from each respective country to score the models on a scale from 1 to 10 using the following criteria:

- **Context Accuracy Score**: This criterion focuses on the accuracy of the model's response in relation to the question posed, irrespective of the dialect or language used. It assesses how well the response addresses the content and context of the question.

- **Dialect Authenticity Score**: This criterion assesses the authenticity of the dialect used in the response, independent of the content's accuracy. It evaluates how authentically the response represents the specific dialect in question.

Model Evaluation. We craft a prompt to assess *Dallah*'s responses and subsequently call the APIs of two different models, Cohere Command R+ and GPT-4 Turbo. In the prompt, we request that the evaluator models rate *Dallah*'s response on a scale from 1 to 10 based on the Dialect Authenticity Score and Content Accuracy Score. We utilize GPT4V to extract a detailed description of the image content and include this description in the prompt to give context to the model. Figure 7 illustrates the prompt used to instruct the models for evaluation.

5.2.1 Model vs. Human Evaluation

In our analysis in Table 3, we compare the performance of *Dallah* based on two evaluators, Cohere Command R+ and GPT-4-Turbo, against human evaluations across several Arabic dialects. The mean absolute differences in scores for dialect authenticity and content accuracy are calculated to quantify the closeness of model evaluations to human judgments. Cohere Command R+ consistently shows a smaller deviation from human scores, with an average difference of 1.47 in dialect authenticity and 1.36 in content accuracy, compared to GPT-4-Turbo's 1.64 and 1.68, respectively. This suggests that Cohere Command R+ is better aligned with human evaluations, offering a more accurate reflection of human perception in dialect authenticity and content accuracy assessments.

5.2.2 Dallah Performance on Dialects

The evaluation of *Dallah*'s performance on various Arabic dialects using model-based and human evaluators in Table 3 provides crucial insights into its linguistic capabilities. The dialect authenticity and content accuracy scores indicate that *Dallah* can generate generally well-received outputs, with some variations across different evaluators and dialects.

The higher ratings from Cohere Command R+ indicate that *Dallah* excels in producing authentic dialect responses that align well with Command R+'s evaluation framework. However, the lower scores from GPT-4-Turbo reveal that some dialects are underrepresented in its training data, leading to
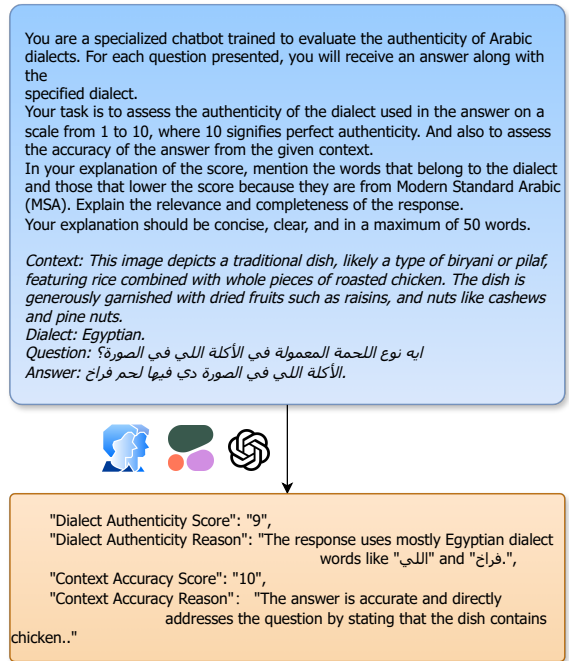

Figure 7: Method employed to evaluate dialect authenticity and content accuracy using both model and human evaluators. Human evaluators are also provided with an image for each question to facilitate better judgment.

misunderstandings of dialectal responses and lower scores in content accuracy.

Moreover, the variation within individual dialects and the representation of each dialect in the LLM, along with the limited data used in the fine-tuning, affect the performance of these systems. It is important to note that spoken dialects often differ from written forms. Written dialects are typically closer to MSA as they lack acoustic cues, which may influence human evaluators. When reading the written model responses, evaluators might perceive them as MSA. When spoken aloud, however, the text may sound dialectal.

Furthermore, we attempt to compare our results with GPT-4V. However, GPT-4V consistently responded in MSA even when prompted with dialects. This highlights our model's superiority over GPT-4V in handling dialectal variation as our model responds in dialect.

Additionally, the human evaluation highlights the importance of incorporating cultural and linguistic subtleties in model assessments. Future work could explore more sophisticated evaluation metrics or integrate additional human feedback to refine *Dallah*'s performance further. We provide qualitative analysis in A.3. This analysis offers

insights into the model's handling of MSA and various dialects.

5.2.3 Feasibility of Model Evaluations for Arabic Dialects

Given the findings from the comparative analysis, model evaluations, particularly using Cohere Command R+, demonstrate potential as useful tools for assessing dialect authenticity in Arabic dialects. While these evaluations do not completely align with human judgments, they offer a sufficiently close approximation that can be valuable in scenarios where rapid or large-scale evaluations are necessary. However, for applications requiring accurate understanding and cultural sensitivity, human assessments should ideally complement these model evaluations to ensure accuracy and relevance in the context of specific Arabic dialects.

6 Conclusion

The paper introduces *Dallah*, an advanced multimodal large language model tailored for Arabic dialects, which demonstrates superior performance in processing both standard Arabic and regional dialects. Developed through innovative data filtering and training, *Dallah* achieves state-of-the-art performance in the LLaVA-benchmark. *Dallah* maintains dialect authenticity and content accuracy, showing promising results in benchmarks and evaluations. Extensive testing shows the model's robustness in MSA and across various dialects and contexts. This model marks a significant step in enhancing Arabic NLP, with future goals to expand dialect coverage and refine evaluation metrics for better user interaction insights.

Limitations

We identify a number of limitations for our work, as follows:

- **Representation of Arabic Culture**: Vision models, LLMs, and datasets used in building MLLMs inadequately represent Arabic figures, places, and culture. As shown in Figure 10, *Dallah* struggles with recognizing Arabic figures, unlike those from the USA. This highlights the need for more diverse cultural datasets.

- **Hallucination Control**: *Dallah*, like many LLMs, is prone to hallucinations, generating inaccurate information. Advanced techniques

and more robust datasets are needed to mitigate this issue and ensure reliability.

- **Dialect Variation and Mixing**: The model sometimes mixes similar dialects, such as Yemeni and Saudi, and struggles with dialects close to MSA. This can be improved with more extensive data collection and fine-tuning.

- **Arabic Text Recognition in Images**: *Dallah* cannot effectively recognize Arabic text within images due to the lack of annotated datasets. Developing such datasets is essential to enhance the model's multimodal capabilities.

Ethics Statement

Acknowledgments

Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[6] and UBC ARC-Sockeye.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Azimjon Ahmedov and Xabibullayeva Shaxrizoda. 2024. The english effect. *Ta'limning zamonaviy transformatsiyasi*, 7(3):18–23.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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. *Preprint*, arXiv:2403.01031.

Richard Barnet and John Cavanagh. 2014. Homogenization of global culture. In *The case against the global economy*, pages 169–174. Routledge.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An opensource chatbot impressing gpt-4 with 90%* chatgpt quality.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zakaria Fahmi and Dacota Liska. 2024. Promoting arabic as a foreign language in the middle east and north africa: Host-grounded study abroad discourses. *Frontiers: The Interdisciplinary Journal of Study Abroad*, 36(1):384–417.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Abdulmohsen Alharthi, Bang An, Ziche Liu, Zhiyi Zhang, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2023. Acegpt, localizing large language models in arabic. *Preprint*, arXiv:2309.12053.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. 2023. Phi-2: The surprising power of small language models. Microsoft Research Blog.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2024. Palo: A polyglot large multimodal model for 5b people. *Preprint*, arXiv:2402.14818.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for Arabic image captioning with gemini decoder. In *Proceedings of ArabicNLP 2023*, pages 1–11, Singapore (Hybrid). Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Alcoba Inciarte, and Md Tawkat Islam Khondaker. 2023. Jasmine: Arabic gpt models for few-shot learning. *Preprint*, arXiv:2212.10755.

OpenAI. 2023. Chatgpt. AI language model. Accessed: 2024-04-30.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

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, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A   Appendices

We organize content here as follows:

- Translation and Filtering Details (A.1)

- Dialect Translation Examples (A.2)

- Qualitative Analysis (A.3)

A.1   Translation and Filtering Details

As described in Section 3.1, we employed a careful translation and filtering process to ensure the quality of the Arabic dataset used for training the *Dallah* model. Figure 2 demonstrates this process, highlighting the importance of maintaining high translation accuracy. Examples with low similarity scores between the original English text and back-translated English text were removed, as shown in the red rows. These include mistranslations such as "What does my browser load?" and "The bird in the photo is an ice skater," which were incorrectly back-translated from the original prompts. Conversely, examples with high similarity scores, as shown in the green rows, were retained, ensuring that both the questions and answers remained consistent and accurate. This careful filtering process was crucial in developing a robust and reliable Arabic multimodal language model capable of handling complex dialectal interactions.

A.2   Dialect Translation Examples

Figure 8 showcases examples of translations from MSA to regional dialects from six Arabic-speaking countries: Egypt, Mauritania, Morocco, Palestine, Saudi Arabia, and Yemen. These translations were performed by native speakers, ensuring cultural and contextual accuracy. Such examples highlight the complexities involved in developing a dialect-aware multimodal language model like *Dallah*.

A.3   Qualitative Analysis

The qualitative evaluation of *Dallah*'s responses showcases its effectiveness in generating accurate and contextually relevant answers across various Arabic dialects. This evaluation is based on examples illustrated in Figures 9, 10, 11, and 12 , highlighting *Dallah*'s capability to handle MSA and diverse dialectal interactions in both textual and visual contexts.

In the context of food descriptions in Figure 9, *Dallah* was asked to describe a traditional dish, including the preparation steps. The response provided a detailed and coherent step-by-step guide, demonstrating the model's understanding of culinary terms. This example highlights *Dallah*'s proficiency in generating detailed content. Additionally, in the same figure, *Dallah* demonstrated the ability to generate detailed and accurate descriptions of an image containing a group of children and was capable of providing potential risks about the activity in the image when asked about potential risks.

As shown in Figure 10, *Dallah* illustrates its ability to describe the appearance of persons and Arabic figures but fails to identify these figures. In contrast, the model was capable of identifying the US president, which is due to the lack of representation for Arabic figures and culture.

Dallah also demonstrated its ability to manage dialectal variations and maintain contextual accuracy as shown in Figure 11 and 12. When addressing questions in a specific dialect, the model accurately reflected local linguistic features and idiomatic expressions. The ability to switch between dialects and maintain contextual accuracy is crucial for multilingual and multicultural applications, highlighting *Dallah*'s comprehensive training on diverse dialectal datasets.

Dallah's qualitative performance underscores its potential as a robust multimodal language model tailored for Arabic dialects. Its capability to generate accurate, contextually relevant, and dialect-specific responses makes it a valuable tool for various applications, from education to cultural preservation. The model's strength in handling diverse dialectal variations and integrating visual and textual information is particularly noteworthy, paving the way for further advancements in Arabic NLP.

Country	MSA	Dialect Translation
Egypt	نعم، يبدو أن المرأتين منخرطتان في محادثة أثناء استخدامهما لهواتفهما، مما يوحي بأنهما ربما يناقشان ما يشاركانه على أجهزتهما أو يستمتعان بمحادثة ودية أثناء تناول العشاء معًا.	ايوا، واضح ان الاتنين ستات مشغولين وهما بيتكلموا في التليفونات بتاعتهم، اللي بيوحي انهم ممكن يكونوا بيتناقشوا اللي في اجهزتهم أو بيستمتعوا بمكاملة ودية وهما بيتعشوا سوا.
Mauritania	هل المرأة واقفة أم جالسة؟	لمرا واقفة وللا قاعدة؟
Morocco	كم عدد الأشخاص الموجودين في الصورة، وماذا يفعلون؟	شحال ديال الناس كاينين في الصورة، وشنو كايديرو؟
Palestine	ما نوع الزهور التي يمكن رؤيتها تحيط بمعالم الحديقة؟	شو نوع الورد الي قادر تشوفه حولين المنتزة؟
Suadi Arabia	نعم، هناك رجل يظهر في الصورة، وهو ينظر إلى الدراجات النارية المتوقفة.	ايه، في رجال يطلّ في الصورة، يطالع في الدبابات واقفة.
Yemen	لا يوجد سوى شخص واحد، وهو راكب أمواج، يظهر في الصورة.	ما بش الا شخص واحد ، وهو راكب أمواج ، باين في الصورة.

Figure 8: Examples of Human Translation of MSA to Dialects Across Six Arabic-Speaking Countries

Figure 9: Three examples demonstrating *Dallah*'s chat capabilities in Modern Standard Arabic (MSA). The first example involves describing a traditional dish and its preparation steps. The second example showcases *Dallah*'s ability to describe an image and identify potential risks for children in the activity depicted. The third example highlights *Dallah*'s proficiency in providing nutritional advice and describing various types of food.

في الصورة، رجل كبير السن، رجل كبير السن ذو لحية طويلة بيضاء ويرتدي رداءً أبيض. يبدو أنه رجل كبير السن، ربما من الشرق الأوسط، ذو شعر أبيض. تشير وضعيته إلى أنه قد يكون في منتصف لحظة تفكير أو تأمل. الصورة بالأبيض والأسود، مما يضفي عليها طابعًا خالدًا وكلاسيكيًا.

صف الصورة بالتفصيل؟

هل من الممكن ان تخبرني بموطن هذا الرجل

ومن الصعب تحديد موطن الرجل بشكل قاطع بناءً على الصورة فقط. ومع ذلك، فإن ملابسه ووجوده صحراء قد يشير إلى أنه في مكان له طابع شرقي أو متوسط شرقي، مثل الشرق الأوسط أو شمال أفريقيا. يمكن أن تكون هذه أيضًا منطقة ذات تأثير ثقافي أوروبي أو متوسط أوسط أو متوسط شرقي. ومع ذلك، بدون سياق إضافي أو تفاصيل أكثر تحديدًا، فمن المستحيل تحديد موطنه بدقة.

انه عمر المختار هل تعرف تاريخه؟

عمر المختار هو رجل ليبي المولد، ولد في عام 1862، وهو قائد عسكري وسياسي ليبي مشهور. اشتهر خلال الحرب العالمية الأولى عندما قاد المقاومة الليبية ضد الاحتلال الإيطالي للبلاد. أصبح رمزًا للاستقلال الليبي وأصبح شخصية مهمة في الكفاح الليبي من أجل الحرية والاستقلال. خلال الحرب العالمية الأولى، قاد عمر المختار القوات الليبية في معارك مختلفة ضد القوات الإيطالية، وكان من أبرز معاركه معركة القردة الشهيرة. في عام 1916، تم أسر عمر المختار من قبل القوات الإيطالية، وتم نقله إلى روما حيث تم اعدامه في العام التالي. على الرغم من وفاته، فإن عمر المختار يظل شخصية تاريخية مهمة ورمزًا للاستقلال الليبي. لقد ألهمت حياته وإنجازاته العديد من الليبيين الذين واصلوا الكفاح من أجل حريتهم واستقلالهم.

ماذا ترى في الصورة؟

في الصورة، يقف رجل يرتدي بدلة وربطة عنق أمام حشد من الناس. إنه يرتدي نظارة ويبدو أنه يبتسم. هناك العديد من الأشخاص الآخرين في الخلفية، مما يخلق جوًا مزدحمًا. يبدو أن الرجل الذي يرتدي البدلة هو الشخصية المركزية في الصورة، حيث يجذب انتباه المشاهد بحضوره.

هل تعرف الشخصية هذه؟

بناءً على الصورة، لا يمكن تحديد هوية الشخصية بشكل قاطع. ومع ذلك، نظرًا لملابسهم وحقيقة أنهم يقفون أمام حشد من الناس، فمن المحتمل أن تكون شخصية بارزة أو مشهورة.

انه الشاعر عبدالله البردوني اخبرني عنه؟

عبد الله البردوني هو شاعر يمني ولد في عام 1910 وتوفي في عام 1986. يُعتبر من أهم الشعراء العرب في القرن العشرين. وُلد في قرية البردوني في اليمن. اشتهر البردوني بأسلوبه الفريد في الشعر وأسلوبه الذي يمزج بين التقاليد الشعرية القديمة والحديثة.

هل يمكنك ان تتعرف على الشخص الموجود في الصورة؟

نعم يمكن التعرف على الشخص الموجود في الصورة على أنه الرئيس أوباما.

اين يقع المكان الذي تم تصويره فيها؟

تم تصوير الرئيس أوباما في موقع تاريخي، ربما في مصر، كما يدل على ذلك وجود الأهرامات في الخلفية.

Figure 10: *Dallah* is capable of answering questions related to images. However, *Dallah* couldn't identify Arabic figures in the first and second examples, while it was able to recognize the US president. Like many LLMs, *Dallah* is prone to hallucinations. Red color font in the responses highlights examples of *Dallah*'s hallucinations.

MA

فين كاين هاد المعلم التاريخي, عطي جواب بالتفاصيل؟

المعلم التاريخي كاين عند سفح جبل صخري، وكايبان فالصورة معلم حجري قديم، يمكن يكون قصر أو معبد. المعلم كايبان فالخلفية، وهادشي كايخليه النقطة المحورية فالصورة. وجود العديد من الأشخاص كايشير أنها يمكن تكون منطقة جذب سياحي أو موقع تاريخي شهير، مما كايخليها مكان شهير للناس باش يتفرجو على المعلم ويستكشفو.

EG

في أي مواقع سياحية معروفة في الصورة دي؟

الصورة فيها مجموعة من البلدات الصغيرة أو القرى على جبل أو تلة، ودا بيظهر أنها أماكن سياحية شهيرة. السياح بيجيوا هنا عشان يستمتعوا بالمناظر الخلابة والتفاعل مع السكان المحليين والتعرف على الثقافة المحلية.

MR

شنهو نوع اللحم الي مشغل في هذي الوجبة الي في الصورة؟

اللحم المشغل في هذي الوجبة هو الدجاج.

SA

ايش نشوف فوق بالصورة ورا المباني اللي عليها نور؟

في جبل كبير باين في الصورة ورا المباني اللي عليها نور.

PS

شو نوع الثوب الي لابسه الشخص الي بالصورة؟

الشخص الي بالصورة لابس ثوب عربي، الي ممكن يكون جزء من الزي الخليجي.

YE

ايش يعني لك لون السماء بالصورة؟

السماء في الصورة زرقاء مع بعض السحب، وهذا يدل على أن الوقت ليل والناس يتجمعون قدام مبنى كبير.

Figure 11: Examples of Dallah's answers in different dialects from the evaluation datasets.

ايش هو الغريب في هاي الصورة؟

الجانب الغريب في هاي الصورة هو وجود قطة برتقالية وبيضاء فوق شجرة مزهرة. القطط عموماً مش حيوانات بتفرح بالتسلق على الأشجار، خاصة لما تكون فيها زهور أو ثمار. ممكن يكون هذا المنظر غير عادي لان القطط مش عموماً بتتفاعل مع البيئة الزهرية أو بتتفاعل مع الثمرة الموجودة على الشجرة. كمان، القطط عموماً مش حيوانات بتتفاعل مع الأشجار، لانها بتستخدم مخالبها عشان تتسلق بس!

ايه هو الغريب في الصورة دي؟

الجانب الغريب في الصورة دي هو وجود قطة برتقالية وبيضاء فوق شجرة أزهار بتلات زهور وردية اللون. القطط مش بتتعلق بالأشجار عادةً، ودا أمر غير معتاد لأن القطط تفضل غالبًا الجلوس على الأسطح المستوية زي الأسرة أو الكراسي أو الطاولات. وكمان القطط حيوانات مفترسة بطبيعتها، فإنها بتحب تستكشف المناطق المحيطة بها، وكمان ممكن تكون القطة الموجودة في الصورة فضولية أو حتى منجذبة للبتلات الوردية الموجودة على الشجرة.

شو الغريب في هاي الصورة؟

الجانب الغريب في هاي الصورة هو وجود قطة برتقالية وبيضة فوق شجرة مزهرة. القطط عموماً مش حيوانات زهور وبتكون في الأماكن العادية، زي الحمامات أو الأحواض أو الأماكن المغلقة. مش مشهد عادي انك تشوف قطة واقفة على شجرة فيها زهور أو شجرة فيها شجر. وجود القطة في هذا المكان غير معتاد وبيثير الفضول.

آشنو هوّ لْعجيب فهاد لتصويرة؟

آش كريب فهاد الصورة هو تصويرة قطة برتقالية وبيضاء كاتجلس فوق غصن شجرة مليانة زهور. القطة كاينة على طرف الغصن، والكاب ديالها كايان الصافية الفولاذية من تحتها. الظلال والتفاعل بين القطة والخلفية كايخلقو صورة ملفتة للنظر وجذابة بصريًا. القطة كاتاخد مركز الصدارة فالصورة، بينما أوراق الشجر والزهور كاتزيد من الإحساس بالطبيعة والهدوء. هاد التصويرة كاتبين بلي القطة تحب الجلوس فوق الأشجار أو ربما كاتحب تستمتع بالمنظر أو الضوء.

ايش الغريب بالصورة هذي؟

الجانب الغريب بالصورة هذي هو وجود قطة برتقالية وبيضاء فوق شجرة مزهرة. عادة ما تكون القطط حيوانات داخلية، وبتحب تسترخي على الفرش أو الأسرة أو أي سطح داخلي تاني. ما تشوف قطة واقفة على شجرة بشكل منتظم، خاصة شجرة فيها زهور بتزيد من المنظر الغير عادي. القطة اللي في الصورة بتتطلع مباشرة للكاميرا، وهذا بيزيد من التأكيد على المنظر الفريد. الصورة بتبين فضول القطة وانجذابها للمكان الخارجي.

شنهو الغريب في هالصورة؟

الجانب الغريب في هالصورة هو وجود قطة برتقالية وبيضة جالسة على غصن شجرة مزهر. ليس من الشائع أن تشوف قطط جالسة على أغصان الأشجار المزهرة، حيث أن القطط عادة تفضل الجلوس على الأسطح الصلبة مثل الأرضيات أو الأثاث. بالإضافة إلى ذلك، فإن وجود القطة وسط الزهور الجميلة يضيف لمسة غريبة للمشهد.

Figure 12: In this examples, *Dallah* was asked "What is the strange thing in this picture?" using six different dialects. *Dallah* responded to each question in the corresponding dialect, demonstrating a degree of dialectness. The responses conveyed the same meaning with slight variations across the different dialects.