



CAMEL-Bench: A Comprehensive Arabic LMM Benchmark

Sara Ghaboura^{1*} Ahmed Heakl^{1*} Omkar Thawakar¹ Ali Alharthi¹
 Ines Riahi² Abduljalil Saif² Jorma Laaksonen² Fahad Shahbaz Khan^{1,3}
 Salman Khan^{1,4} Rao Muhammad Anwer^{1,2}

¹Mohamed bin Zayed University of AI, ²Aalto University, ³Linköping University, ⁴Australian National University
<https://mbzuai-oryx.github.io/Camel-Bench/>

Abstract

Recent years have witnessed a significant interest in developing large multimodal models (LMMs) capable of performing various visual reasoning and understanding tasks. This has led to the introduction of multiple LMM benchmarks to evaluate LMMs on different tasks. However, most existing LMM evaluation benchmarks are predominantly English-centric. In this work, we develop a comprehensive LMM evaluation benchmark for the Arabic language to represent a large population of over 400 million speakers. The proposed benchmark, named CAMEL-Bench, comprises eight diverse domains and 38 sub-domains including, multi-image understanding, complex visual perception, handwritten document understanding, video understanding, medical imaging, plant diseases, and remote sensing-based land use understanding to evaluate broad scenario generalizability. Our CAMEL-Bench comprises around 29,036 questions that are filtered from a larger pool of samples, where the quality is manually verified by native speakers to ensure reliable model assessment. We conduct evaluations of both closed-source, including GPT-4 series, and open-source LMMs. Our analysis reveals the need for substantial improvement, especially among the best open-source models, with even the closed-source GPT-4o achieving an overall score of 62%. Our benchmark is publicly available on [CAMEL-Bench](https://mbzuai-oryx.github.io/Camel-Bench/) page.

1 Introduction

Large multimodal models (LMMs) have recently achieved significant advancements across a broad spectrum of tasks, including visual reasoning, perception, and multimodal understanding. Closed-source models such as GPT-4V and open-source LMMs, such as LLaVA (Liu et al., 2023) have demonstrated effectiveness in tasks like image captioning (Radford et al., 2021), visual question answering (VQA) (Li et al., 2022, 2023a), and complex visual reasoning (Cho et al., 2021). These



Figure 1: The proposed CAMEL-Bench covers eight diverse and challenging domains: *multimodal understanding and reasoning*, *OCR and document understanding*, *chart and diagram understanding*, *video understanding*, *cultural-specific understanding*, *medical imaging understanding*, *agricultural image understanding*, and *remote sensing understanding* in Arabic. CAMEL-Bench covers 38 sub-domains with over 29K questions carefully curated by native Arabic speakers to rigorously evaluate essential skills desired in Arabic LMMs.

recent developments have led to the introduction of different benchmarks to evaluate the performance of open and closed-source LMMs. Despite these advances, most existing LMM benchmarks are English-centric, limiting their applicability to other languages (Rasheed et al., 2025).

With over 400 million speakers, Arabic is the 5th most widely spoken languages globally. In the context of large language models (LLMs), there exist various attempts in developing Arabic LLMs (Sengupta et al., 2023; Huang et al., 2023) which has also led to the introduction of Arabic LLM benchmarks (Koto et al., 2024). In the context of LMMs, few recent works explore Arabic-centric evaluations in certain areas such as, scientific ex-

ams (Das et al., 2024), cultural aspects (Romero et al., 2024; Alwajih et al., 2024), Arabic question answers and documents (Abdallah et al., 2024; Mahmoud et al., 2014). However, there is still a lack of comprehensive and diverse Arabic LMM evaluation benchmarks (see Tab. 1) in the literature to rigorously evaluate and study LMMs for Arabic.

To address the aforementioned issue, we introduce the first comprehensive Arabic LMM evaluation benchmark, named CAMEL-Bench. CAMEL-Bench is designed to encompass a wide range of tasks and focus on the Arabic-speaking population. It spans eight diverse domains and 38 sub-domains (see Fig. 1). The eight domains are: *Multimodal understanding and reasoning*, *OCR and document understanding*, *chart and diagram understanding*, *video understanding*, *cultural-specific understanding*, *medical image understanding*, *Agricultural image understanding*, and *remote sensing understanding*. Further, the 38 sub-domains (see Fig. 1) covered by our CAMEL-Bench are: visual understanding and reasoning, object hallucination evaluation, math and logic reasoning, scientific reasoning, VQA, infographics VQA, complex visual perception, real-world spatial understanding, multi-image understanding, object-level perception, newsletter, PowerPoint slides, scene text, handwriting, lines, books, documents, charts, diagrams, tables, general video scenes, cultural-specific occasions, countries and landmarks in videos, countries and landmarks in images, food, celebrities, cultural VQA, basic medical science, clinical medicine, public health, pharmacy, diagnosis, medical understanding, plant types, fruit and veggies identification, plant illness, and geospatial imagery subdomains (land, transportation and construction).

Our CAMEL-Bench comprises 29,036 questions and follows an extensive manual verification process by native-speakers to ensure the resulting benchmark is of high-quality. We conduct extensive experiments using open and closed-source LMMs. Our results reveal the need for substantial improvement in handling of Arabic multimodal data, shedding light on the areas requiring further Arabic LMM improvements.

2 CAMEL-Bench

2.1 Data Collection

Our dataset encompasses eight diverse domains to ensure a versatile multi-task Arabic LMM benchmark for different real-world scenarios. Each

Domain/Characteristics	Exams-V*	CVQA*	Henna	KHATT	CAMEL-Bench (ours)
Multimodal Und. & Reasoning	✓	✗	✓	✗	✓
OCR & Docs Und.	✗	✗	✗	✓	✓
Charts & Diagrams Und.	✓	✗	✗	✗	✓
Video Und.	✗	✗	✗	✗	✓
Medical Image Und.	✗	✗	✗	✗	✓
Agricultural Image Und.	✗	✗	✗	✗	✓
Remote-Sensing Und.	✗	✗	✗	✗	✓
Cultural-Specific Und.	✗	✓	✓	✗	✓
Open Source	✓	✓	✗	✓	✓
Question Numbers	823	200	1.1K	5K	29K

Table 1: Comparison of our CAMEL-Bench with existing Arabic LMM benchmarks: Exams-V (Das et al., 2024), CVQA (Romero et al., 2024), Henna (Alwajih et al., 2024), and KHATT (Mahmoud et al., 2014). Here * denotes that only the Arabic part of the benchmark is counted.

domain is further subdivided into different sub-domains, each focusing on a distinct aspect. During the data collection process, we either utilize available Arabic multimodal data samples or employ samples from existing English-centric LMM benchmarks. These English samples are then translated to Arabic via GPT-4o and verified. Alternatively, we manually collect and generate the Arabic samples for the remaining sub-domains from the internet. Tab. 3 in the Appendix presents the details of different data sources used for data collection for the 38 sub-domains corresponding to eight domains, with around 29k questions in total.

2.2 Question-Answers Pairs Generation

We note that a major part of our original Arabic data is not derived from ready-made VQA datasets. Some sub-domains, such as celebrities and food, consist of image-only data, while others, like Pexel’s countries and landmarks, contain image-caption pairs. To create a rich and diverse VQA corpus, we first ensure that each image is accompanied by detailed contextual information. This context is sourced from a combination of Wikipedia (e.g., for food-related data), manual curation (e.g., for countries and landmarks in videos), and AI-generated content based on a manually provided context (e.g., for diagrams and infographics). Next, we generate multiple-choice questions (MCQs) for each sample using the GPT-4o model. The prompt is meticulously crafted to adhere to key criteria: each sample generates three multiple-choice questions (MCQs), with four distinct, non-synonymous options per question, only one of which is correct. The questions contain no embedded hints, ensuring that answers are derived exclusively from the image, without requiring prior knowledge. Additionally, the image must provide

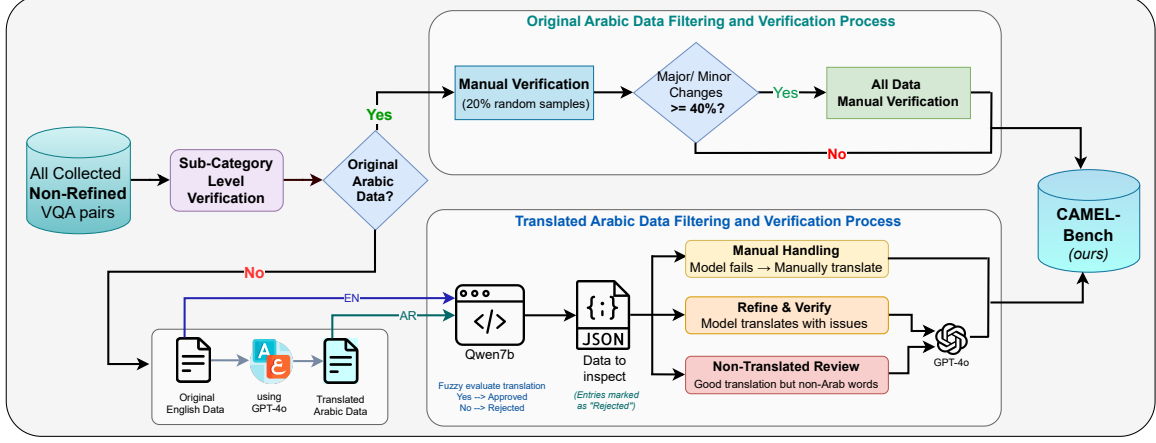


Figure 2: The CAMEL-Bench Filtering and Verification Pipeline consists of two paths: Original Arabic and translated Arabic. For original Arabic (top row), a 20% random sample undergoes manual verification; if errors are below 40%, the data passes; otherwise, the entire sub-category is reviewed. For Translated Arabic (bottom row), We employ the Qwen7B model (Bai et al., 2023a) to assess semantic similarity between the original and translated question-answer pairs on fuzzy-basis evaluation. Pairs passing the evaluation proceed, while those that fail undergo manual review. Based on this, data may require *Manual Handling* for manual re-translation, *Refine & Verify* for refinement through the model, or *Non-Translated Review* where the data is re-sent for translation due to the absence of an Arabic version.

enough information to fully support the correct answer, eliminating the need for guesswork. In total, this process produces a corpus of 4.4K generated questions with 17.7K answers, enabling a comprehensive set of questions for evaluation.

2.3 Data Filtering and Verification

The data collection and question-answer pair generation process lead to over 41k questions in total which then undergoes to filtering and verification process. The CAMEL-Bench filtering and verification process (see Fig. 2) is carefully conducted based on whether the QA text is originally Arabic or translated into Arabic from English language. For all sub-domains derived from original Arabic context, we take a 20% randomly sampled subset for manual verification. In case the error remains less than a 40% threshold, the sub-category is accepted into CAMEL-Bench. Alternatively, the entire sub-category undergoes manual review.

In the case of the translated Arabic data from English, the original English context is also incorporated into the filtering and verification process. Here, Qwen7B (Bai et al., 2023a) is used to compare the semantic similarity between the English and the English-translated data at the QA-pair level using fuzzy evaluation. To ensure the model understands semantic similarity in Arabic, we provided 5 few-shot prompts. Subsequently, QA-pairs rejected by Qwen7B (Bai et al., 2023a) are manu-

ally reviewed, resulting in one of three outcomes. *Manual Handling* implying that data requires full re-translation. *Refine and Verify* referring that the translation can be refined using the model. *Non-Translated Review* implying that the non-translated data is re-sent to the model for translation. Consequently, we obtain 29,036 high-quality questions.

3 CAMEL-Bench Benchmark Evaluation

Evaluation Metrics: Our evaluation framework is designed with three specialized metrics, each carefully aligned to different types of datasets and tasks. For MCQ datasets like MMT (Ying et al., 2024) and MMMU (Yue et al., 2024a), we utilize exact match accuracy to ensure precise evaluation. For optical character recognition (OCR) datasets, such as PATS (Al-Muhtaseb, 2010) and Evarest (Hassan et al., 2021), where accurate text extraction is critical, we adopt edit distance (Ristad and Yianilos, 1998) as the key metric. For more flexible datasets like VQAv2 (Goyal et al., 2017), MathVista (Lu et al., 2023), and GeoChat (Kuckreja et al., 2024), where multiple synonymous answers can be considered correct, we implement a fuzzy evaluation method for all such datasets. This approach uses GPT-4o to compare the predicted answer with the ground truth while accounting for the context of the question. By incorporating these diverse metrics, our evaluation provides a robust and comprehensive assessment that adapts to the unique demands

Method	MM Understanding & Reasoning	OCR & Document Understanding	Charts & Diagram Understanding	Video Understanding	Cultural Specific Understanding	Medical Imaging	Agro Specific	Remote Sensing Understanding
GPT-4o	57.90	59.11	73.57	74.27	80.86	49.90	80.75	22.85
GPT-4o-mini	48.82	42.89	64.98	68.11	65.92	47.37	79.58	16.93
Gemini-1.5-Pro	46.67	36.59	47.06	42.94	56.24	33.77	72.12	17.07
Gemini-1.5-Flash	45.58	33.59	48.25	53.31	46.54	42.86	76.06	14.95
Pangea-7B	40.09	26.47	38.87	49.01	20.34	31.99	74.51	6.67
Qwen2-VL-2B	40.59	25.68	27.83	38.90	34.27	29.12	52.02	12.56
InternVL2-8B	30.41	15.91	30.27	51.42	20.88	29.48	44.47	5.36
LLaVa-NeXt-7B	26.33	19.12	27.56	44.90	28.30	22.54	42.00	8.33

Table 2: **Performance comparison of different closed-and open-source LMMs on CAMEL-Bench.** We present per-domain results of seven LMMs: GPT-4o (OpenAI, 2024), GPT-4o-mini (OpenAI, 2024), Gemini-1.5-Pro (AI, 2023), Gemini-1.5-Flash (AI, 2023), Pangea-7B (Yue et al., 2024b), Qwen2-VL (Bai et al., 2023b), InternVL2-8B (Chen et al., 2023), and LLaVaNeXt-7B (Liu et al., 2024). GPT-4o excels in most domains, while GPT-4o-mini offers an impressive balance of performance and model size. All models struggle with remote sensing, medical imaging, OCR & document understanding, and general multimodal understanding and reasoning domains. Open-source models like InternVL2-8B and LLaVaNeXt-7B show a decline in performance across domains, with their best results in video understanding.

	Multimodal Und. & Reasoning	Remote Sensing Understanding	Chart & Diagram Und. / Tables
Ground Truth	س: ما هو شكل الطائرة الورقية؟ ج: سداسي الاضلاع	س: كم عدد ملاعب كرة السلة الموجودة في الصورة؟ ج: يوجد ٢ ملاعب كرة سلة في الصورة	س: ما هي قيمة السكان بالملايين في عام 2021 المقطرة في الرسم البياني؟ ج: 434
GPT-4o	ج: مستطيل	ج: أربعة ملاعب لكرة السلة	ج: 440
Gemini-1.5-Pro	ج: خماسي	ج: لا يوجد ملاعب	ج: 422
Gemini-1.5-Flash	ج: Diamond	ج: ٤ ملاعب	ج: 457

Figure 3: Qualitative example of various models on CAMEL-Bench sample. The green box means correct and red means incorrect.

and response formats of each dataset.

Tab. 2 presents a comparative evaluation of five different models on a range of multimodal (MM) understanding tasks, each assessing the capabilities of the models in distinct domains. The models include GPT-4o, GPT-4o-mini, Gemini-1.5-Pro, Gemini-1.5-Flash, and Qwen2-VL-2B, evaluated on key tasks such as multimodal reasoning, OCR & document understanding, chart & diagram interpretation, video analysis, and several domain-specific tasks like cultural understanding, medical imaging, agricultural (agro) understanding, and remote sensing. GPT-4o excels across tasks, leading in MM reasoning (57.90), chart/diagram understanding (73.57), video analysis (74.27), cultural (80.86), and agro-specific understanding (80.75). Models perform well on MCQs and binary-option tasks due to guessing probability and context. Infographics, designed for easy interpretation, also see high accuracy across all models. In contrast, Arabic OCR tasks, particularly in datasets like Khatt, historical documents prove exceptionally challenging. This difficulty stems from the complex nature of Arabic script, which uses ligatures and diacritics (small markings that alter pronunciation and meaning). Remote sensing understanding also remains difficult, with scores like 22.85 (GPT-4o) and 16.93

(Qwen2-VL-2B), highlighting the complexities of interpreting satellite imagery.

The Fig. 3 highlights a critical challenge in Arabic multimodal understanding, where all models, failed to accurately interpret the linguistic context in the provided CAMEL-Bench samples. This underscores the complexity of Arabic linguistics, especially in multimodal tasks, and the need for more robust language models that can effectively integrate both visual and textual information in Arabic contexts.

4 Conclusion, Limitations and Societal Impact

We present a comprehensive and diverse benchmark, named CAMEL-Bench, for Arabic LMM evaluation. To the best of our knowledge, CAMEL-Bench is the first comprehensive Arabic LMM benchmark comprising eight diverse domains and 38 sub-domains with around 29k questions that are filtered from a larger pool of 41k samples with the quality verified by native speakers. We conduct extensive evaluations of open- and closed-source LMMs, highlighting the need for substantial improvements in different areas for future Arabic LMM development. Although our CAMEL-Bench

strives to significantly contribute towards developing sophisticated Arabic LMMs, we note that it mainly covers modern standard Arabic and does not fully explore other Arabic dialects. As the data samples are either based on existing datasets or new data that is crawled from the internet, it is possible that CAMEL-Bench exhibits biases already existing in the benchmarks. Nevertheless, we believe CAMEL-Bench is a step towards the inclusion of Arabic language and Arabic-speaking populations in accessing the benefits of LMMs.

References

- Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024. Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2049–2059.
- Google AI. 2023. *Gemini: A family of highly capable multimodal models*.
- Husni A Al-Muhtaseb. 2010. *Arabic text recognition of printed manuscripts. Efficient recognition of off-line printed Arabic text using Hidden Markov Models, Bigram Statistical Language Model, and post-processing*. Ph.D. thesis, University of Bradford.
- Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.
- Anthropic. 2024. *Claude*. AI assistant.
- Arar Tawil. 2023. Arabic food 101. <https://www.kaggle.com/datasets/araraltawil/arabic-food-101>.
- Muhammad Awais, Ali Husain Salem Abdulla Alharthi, Amandeep Kumar, Hisham Cholakkal, and Rao M. Anwer. 2024. *Agrogpt: Efficient agricultural vision-language model with expert tuning*. *arXiv*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng-guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.
- Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.
- Heba Hassan, Ahmed El-Mahdy, and Mohamed E Hussein. 2021. Arabic scene text recognition in the deep learning era: Analysis on a novel dataset. *IEEE Access*.
- 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. [Acept, localizing large language models in arabic](#).
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.
- Kartik Kuckreja, Muhammad S. Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad S. Khan. 2024. Geochat: Grounded large vision-language model for remote sensing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- 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 *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Taffjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.
- Sabri A Mahmoud, Irfan Ahmad, Wasfi G Al-Khatib, Mohammad Alshayeb, Mohammad Tanvir Parvez, Volker Märgner, and Gernot A Fink. 2014. Khatt: An open arabic offline handwritten text database. *Pattern Recognition*, 47(3):1096–1112.
- Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [ChartQA: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- Mohammad-Alfaifi. n.d. Github - mohammad-alfaifi/arab-celeb-dataset. <https://github.com/mohammad-alfaifi/arab-celeb-dataset>. Accessed: 2024-10-15.
- Rayyan Najam and Safiullah Faizullah. 2024. [Historical arabic handwritten text recognition dataset](#).
- OpenAI. 2024. Gpt-4o model. <https://openai.com>. Accessed: 2024-10-14.
- Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3170–3180.
- PATD. 2024. Printed arabic text database for recognition systems. <http://www.inf.u-szeged.hu/patd/>.
- Pexel. 2024. Pexel: The best free stock photos, royalty-free images and videos shared by creators. <https://www.pexels.com/>.

- Pinterest. 2024. Pinterest platform. <https://www.pinterest.com/>.
- 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.
- Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A large multilingual multimodal language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.
- Rana SM Saad, Randa I Elanwar, NS Abdel Kader, Samia Mashali, and Margrit Betke. 2016. Bce-arabic-v1 dataset: Towards interpreting arabic document images for people with visual impairments. In *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 1–8.
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2024. *Mtvqa: Benchmarking multilingual text-centric visual question answering*.
- Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.
- Wikipedia. 2024. Wikipedia the free encyclopedia. <https://www.wikipedia.org/>.
- Yue Wu and Prem Natarajan. 2017. Self-organized text detection with minimal post-processing via border learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- xAI. 2024. xai. grok-1.5 vision preview. <https://x.ai/blog/grok-1.5v>.
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.
- Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.
- Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi lei, Quanfeng Lu, Peng Gao, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. *MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI*. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- YouTube. 2024. <https://www.youtube.com/>. Accessed: 2024-10-01.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoyi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantaruban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024b. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2024c. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

A Appendix

B More on Dataset Curation

The dataset utilized in this work was carefully curated with a rigorous focus on data quality, relevance, and diversity. Our curation process involved selecting multimodal data from various domains, including images, text, videos, and specialized fields such as medical imaging, agriculture, and remote sensing. To ensure the integrity and accuracy of the dataset, we employed multiple stages of data verification. This process involved cross-validation, thorough verification procedures for Arabic content, and the integration of standardized data sources where applicable.

C Dataset Overview and Task Splits

This section provides a comprehensive breakdown of the datasets used across eight distinct categories, illustrating the diversity and depth of our evaluation framework. Each category is further divided into sub-domains, ensuring that the multimodal models are rigorously tested on a wide range of tasks and datasets. This structure guarantees comprehensive coverage and introduces varied challenges to thoroughly assess model performance. Refer to Tab. 3 for a detailed breakdown of the data categories with their statistics.

C.1 Multimodal Understanding and Reasoning

This category encompasses various sub-domains such as visual understanding, object hallucination evaluation, and complex visual perception. Key datasets include MME, MMBench, ScienceQA-IMG, and VQA2. These datasets test the model’s ability to handle intricate reasoning tasks across both visual and textual inputs, with a total of 3,971 questions under the visual understanding sub-domain, and significant representation from other tasks like scientific reasoning (1,624 questions) and object-level perception (60 questions).

C.2 OCR and Document Understanding

Document understanding covers scanned documents, scene understanding, text extraction, and more. This category emphasizes precise OCR and textual recognition from images and scanned materials. Datasets like ArabicDatasetOCR and ISI-PPT-Dataset challenge the model to process a diverse range of document types. A substantial

number of questions come from Handwritten Text datasets (1,400 questions) and PPT OCR (2,354 questions), ensuring the model is evaluated across both structured and unstructured document types.

C.3 Chart and Diagram Understanding

In chart and diagram interpretation, models are tested on understanding visual representations of data, such as charts, diagrams, and tables. This includes datasets like ChartQA, MMMU, and BCE-Arabic. The evaluation focuses on tasks such as understanding diagrammatic reasoning and tabular data with 1,994 questions from diagram datasets and 745 questions involving charts, providing a robust examination of the model’s ability to interpret visual data efficiently.

C.4 Video Understanding

This category assesses the model’s ability to process and comprehend video data, focusing on tasks like recognizing countries, landmarks, and occasions. Video-MME is a prominent dataset, contributing 654 questions to the evaluation. The inclusion of diverse sub-domains, such as recognizing cultural aspects through video, highlights the importance of temporal and visual information synthesis in multimodal reasoning.

C.5 Cultural Specific Understanding

The cultural understanding domain tests the model’s capacity to handle tasks specific to certain cultures, including food, landmarks, and celebrities. Datasets like arabic-food-101 and Pexel challenge the model to recognize culturally significant items, with 444 questions focused on celebrities and 494 on countries/landmarks. These tasks highlight the model’s ability to adapt and generalize across different cultural contexts.

C.6 Medical Imaging

Covering a range of sub-domains in the medical field, this category includes tasks related to basic medical science, clinical medicine, and public health, using datasets like MMMU and MMT-MI-Bench. These datasets assess the model’s potential in specialized medical contexts, with over 1,200 questions spanning diagnosis, medical understanding, and pharmacy, ensuring a rigorous evaluation of the model’s performance in handling critical medical information.

Domains	Sub-Domains	Source	Number of Questions
Multimodal Understanding and Reasoning	Visual understanding/reasoning	MME, MMBench, MMT-Bench-MI, SEED, MMMU	3,971
	Object hallucination evaluation	CountBench, MMT-Bench-MI, POPE	997
	Math and logic reasoning	MathVista	531
	Scientific reasoning	ScienceQA-IMG, Exams-V	1,624
	Visual Question Answering	GQA, VizWiz, VQAv2	3,840
	Infogrphics VQA	AI-Generated (GPT-4o), Pinterest	120
	Complex Visual Perception	BLINK	1,422
	Real-world Spatial Understanding	RealWorldQA	624
	Multi-image Understanding	MMT-Bench-MI, MuirBench	1,062
	Object-level Perception	COCO, ImageNet, Mocheg, Snli-Ve	60
OCR and Document Understanding	Scanned Documents	ArabicDatasetOCR	480
	TextVQA	MTVQA	703
	Scene	EvArEST	1,217
	Books	Historical Arabic Handwritten Text Recognition Dataset	40
	PPTs (OCR)	ISI-PPT-Dataset	2,354
	PPTs (VQA)	ISI-PPT-Dataset	711
	Handwritten	KHATT Line	1,400
	Scanned Newsletters (VQA)	PATD	506
	OCR Lines	PATS-01	520
	Charts	ChartQA	745
Chart and Diagram Understanding	Diagrams Understanding	MMMU (diagrams), ICON-QA, AI-Generated, Pinterest, BCE-Arabic	1,994
	Tables	BCE-Arabic, Excel	81
Video Understanding	Countries/landmarks	Pexel	111
	Occasions	Pexel	
	Video-MME	Video-MME	654
Cultural Specific Understanding	Celebrities	arab-celeb-dataset	444
	Food	arabic-food-101, Pexel	347
	Countries/landmarks	Pexel	494
Medical Imaging	Basic Medical Science	MMMU	154
	Clinical Medicine	MMMU	
	Public Health	MMMU	
	Pharmacy	MMMU	
	Diagnosis	MMMU	
	Medical Understanding	MMT-MI-Bench	78
	Basic Medical Science	MMMU Pro	274
	Clinical Medicine	MMMU Pro	
Agricultural Image Understanding	Agriculture VQA	AgroGPT	769
Remote Sensing Understanding	GeoData VQA	GeoChat	709
Total			29,036

Table 3: Different data sources used for 38 sub-domains corresponding to eight domains, with around 29k questions in total. The different data sources are: MME (Fu et al., 2023), MMBench (Liu et al., 2025), MMT-Bench-MI (Ying et al., 2024), SEED (Li et al., 2024), MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024c), CountBench (Paiss et al., 2023), POPE (Li et al., 2023b), MathVista (Lu et al., 2023), Exams-V (Arabic portion) (Das et al., 2024), ScienceQA-IMG (Lu et al., 2022), GQA (Hudson and Manning, 2019), VizWiz (Bigham et al., 2010), VQAv2 (Goyal et al., 2017), BLINK (Fu et al., 2024), MuirBench (Wang et al., 2024), COCO (Lin et al., 2014), Imagenet (Deng et al., 2009), Mocheg (Yao et al., 2023), Snli-Ve (Xie et al., 2019), Pinterest (Pinterest, 2024), RealWorldQA (xAI, 2024), PATS-01 (Al-Muhtaseb, 2010), KHATT (Mahmoud et al., 2014), PATD (PATD, 2024), Historical Arabic Handwritten Text Recognition Dataset (Najam and Faizullah, 2024), ISI-PPT-Dataset (Wu and Natarajan, 2017), EvArEST (Hassan et al., 2021), MTVQA (Tang et al., 2024), ChartQA (Masry et al., 2022), IconQA (Lu et al., 2021), BEC-Arabic (Saad et al., 2016), Claude-3.5 (Anthropic, 2024), arab-celeb-dataset (Mohammad-Alfaifi, n.d.), arabic-food-101 (Arar Tawil, 2023), Countries and landmarks (Wikipedia, 2024; Pexel, 2024; YouTube, 2024), Pexel (Pexel, 2024), AgroGPT (Awais et al., 2024), GeoChat (Kuckreja et al., 2024).

C.7 Agricultural Image Understanding

The agricultural domain is represented through datasets like AgroGPT, with 769 questions focused on agricultural understanding tasks. These tasks test the model’s capacity to process and interpret images related to agricultural settings, reinforcing the model’s ability to work with real-world scenarios in agriculture and environment-based challenges.

C.8 Remote Sensing Understanding

This category evaluates the model’s ability to handle remote sensing data, specifically focusing on geographical data interpretation through datasets

like GeoData VQA and GeoChat. With 709 questions in this domain, the model is tested on its spatial reasoning and understanding of complex remote-sensing imagery, crucial for applications in fields like environmental monitoring and geography.

In total, the dataset includes 29,036 questions across all categories, providing a comprehensive and diverse benchmark for evaluating the multimodal model’s performance across a wide spectrum of tasks. This balanced distribution ensures that the model is tested thoroughly, with each domain offering unique challenges and insights into the model’s strengths and areas for improvement.

D CAMEL-Bench Data Samples

The Fig. 4 illustrates a rich variety of data samples from CAMEL-Bench, highlighting its versatility across eight distinct domains. These categories span diverse tasks, including Chart & Diagram Understanding, Multimodal Reasoning, OCR & Document Understanding, Cultural-Specific Knowledge (with separate focuses on celebrities and food), Agricultural Image Analysis, and Video Understanding of Landmarks. Each category showcases a unique challenge, from interpreting medical X-rays and reading handwritten Arabic text to recognizing plant diseases, famous landmarks, or culturally significant items like food and celebrities.

This diversity emphasizes CAMEL-Bench's strength in addressing both linguistic and visual understanding, particularly in contexts rich with cultural nuance. Whether it's identifying objects in images, interpreting technical documents, or recognizing cultural symbols, the dataset is designed to foster well-rounded, context-aware AI models. These capabilities are essential for real-world applications, such as healthcare diagnostics, agriculture, tourism, and cross-cultural interactions, making CAMEL-Bench a powerful tool for advancing multimodal AI systems.

<p>Chart & Diagram Und. / Tables</p> <p>س: ما هو التشخيص المحتمل للمريض الموضح أدناه؟</p>  <p>(أ) التقرص (ب) تليين العظام (ج) متلازمة مارفان (د) اعتلال العظم الكلوي</p>	<p>Multi-modal Und. & Reasoning / Multi-Image Understanding</p> <p>س: أي من الصور التالية تحتوي على نفس الشيء الذي يظهر في الصورة المعطاة ولكن بخاصية اللعنان؟</p>  <p>(أ) </p> <p>(ب) </p> <p>(ج) </p> <p>(د) </p> <p>(هـ) الإجابة ليست ضمن الخيارات المتاحة</p>								
<p>OCR & Document Understanding</p> <p>والسلام والرضى والتبر من أموالهم الإله والربوب المشاعة والصبر على في الإله والمال والكرامة والكرامة الله الصديق الذي لا يذنبه وهو تربية عليه فهو مدح من الله والحمد لله انتصر لنفسه الخذلان وأن الفرع من الكروب بمعنى أنه يعقبه لا محالة لعدم دوامه فعليك أن تصبر على ما أصابك منه محتسبا راجيا وقوع الفرج من ذلك أن ذلك من غم الأمور وأن مع العشر كالكرب وضيق الصدر شبرا كالفرج والشرح فأنظر أوقات الصبر والكرب والعسر أول أوقات النصر والفرج واليسر فكانها مقارنة لها وأعلم أن القلم قد جرى بما هو</p>	<p>Chart & Diagram Und. / Tables</p> <p>س: كم عدد المدخنين الذين أصيبوا بالالتهاب الشعبي المزمن بعد ثلاث سنوات من الملاحظة؟</p> <table border="1"> <thead> <tr> <th>مدخنون</th> <th>غير مدخنين</th> </tr> </thead> <tbody> <tr> <td>٢٥</td> <td>١٥</td> </tr> <tr> <td>٢٤٩٠</td> <td>٧٤٧٠</td> </tr> <tr> <td>٢٥١٥</td> <td>٧٤٨٥</td> </tr> </tbody> </table> <p>أصيبوا بالالتهاب الشعبي المزمن بعد ثلاث سنوات من الملاحظة لم يصابوا بالتهاب شعبي مزمن إجمالي تعداد المجتمع (١٠ آلاف فرد)</p> <p>(أ) ١٥ (ب) ٢٥ (ج) ٧٤٩٠ (د) ٢٤٩٠</p>	مدخنون	غير مدخنين	٢٥	١٥	٢٤٩٠	٧٤٧٠	٢٥١٥	٧٤٨٥
مدخنون	غير مدخنين								
٢٥	١٥								
٢٤٩٠	٧٤٧٠								
٢٥١٥	٧٤٨٥								
<p>Cultural-Specific Understanding/ Celebrities</p> <p>س: ما جنسية الشخص الذي يظهر في الصورة؟</p>  <p>(أ) تونسي (ب) مصري (ج) مغربي (د) جزائري</p>	<p>Cultural-Specific Understanding/ Food</p> <p>س: ما نوع الخبز المستخدم في الصورة؟</p>  <p>(أ) خبز التوست (ب) خبز البيتزا (ج) خبز الصاج (د) خبز الباجيت</p>								
<p>Agricultural Image Understanding/ Plant Illness</p> <p>س: ما هي حالة الأوراق في الصورة؟</p>  <p>(أ) جافة وبنية (ب) خضراء وصحية (ج) ملونة بالأصفر والأحمر (د) ممزقة وغير صحية</p>	<p>Cultural-Specific Understanding/ Celebrities</p> <p>س: كم عدد الطائرات الفضية الموجودة في الصورة؟</p>  <p>ج: يوجد طائرتان فضيتان في الصورة</p>								
<p>Video Understanding/ Countries and Landmarks in videos</p> <p>س: ما هو الهيكل القديم الذي يظهر في المقطع حيث يتجول أشخاص حوله؟</p>  <p>(أ) الهرم الأكبر (ب) هرم سفارة (ج) معبد هاتشيبسوت (د) معبد كارناك</p>									

Figure 4: CAMEL-Bench data samples span eight diverse domains, encompassing a wide range of types and tasks.