Arabic Medical Question Answering using LLaMa 3.2 3B

1st Abdelaziz Amr
Nile University
Cairo, Egypt

2nd Hussien Ahmed
Nile University
Cairo, Egypt

3rd Mamdouh Koritam
Nile University
Alexandria, Egypt

4th Mohamed Youssef
Nile University
Cairo, Egypt

***Abstract*—Large language models have been performing very well on natural language processing tasks; however, they are still not well explored for low-resource languages like Arabic, especially domain-specific areas such as healthcare. This work is directed at fine-tuning the LLaMA 3.2-3B model on Arabic medical question-answering using the Arabic Healthcare Dataset, comprising 3,000 samples. Three methods were utilized in assessing the performance of the model: zero-shot, few-shot, and fine-tuning. Zero-shot responses were less wordy and more superficial, occasionally even lapsing into English. The few-shot learning gave more information, although occasionally veered off course. The fine-tuning was more contextually relevant, mostly accurate, and conveyed tone and content similar to the dataset. To address the limitations of traditional NLP evaluation metrics like BLEU and ROUGE, GPT-4 was employed as a benchmark for assessing accuracy, contextual understanding, and clarity. The findings highlight the potential of fine-tuning LLMs for low-resource languages while underscoring the importance of advanced evaluation methods and prompt engineering. This work lays the groundwork for future research exploring larger datasets, multilingual capabilities, and domain-specific applications to further enhance the utility of LLMs in healthcare and other critical areas.**

***Index Terms*—Natural language processing, LLM, Contextual, Question-Answering, LLaMa, Healthcare**

I. Introduction

Large Language Models have increased the success of NLP in recent years by applications such as, but not limited to, text generation, machine translation, and question answering. While opensource LLMs are getting popular, with many instances outperforming their closed source peers-like LLaMA-3 outperforming ChatGPT-their performance on multilingual settings is rather less explored, especially for low-resource and morphologically complex languages like Arabic. This under representation is very significant in critical domains such as healthcare, where accuracy and contextually relevant information are a matter of life [1].

Arabic is a very complex language from the NLP point of view because of its morphology, orthographic ambiguity, diglossia, and various dialects. Besides, the limited availability of Arabic linguistic resources compared to high-resource languages, like English, makes all further complications develop robust NLP solutions. That is further exacerbated in specialized domains like medicine, where very small mistakes in retrieving information or understanding might be catastrophic [2].

This paper fills these gaps by exploring the performance of the LLaMA-3-2B model on an Arabic medical question-answering task using the Arabic Healthcare Dataset (AHD). We will review the model in zero-shot, few-shot, and fine-tuned settings; all aim to enhance generation capabilities for correct and contextually understandable Arabic medical answers. The aim of this analysis is to contribute to the development of Arabic-focused LLMs and propose some effective strategies for model performance evaluation in domain-specific low-resource languages[3].

II. Related Work

This section presents the critical review of the existing literature related to LLMs, puts much focus on their application in Arabic NLP-particularly question answering tasks-and also presents key challenges and recent advances. Although LLMs have achieved remarkable results in the benchmarks for the English language, their effectiveness in other languages, including Arabic, is still an active research area. Several studies have indicated that multilingually trained models face challenges when carrying out highly demanding NLP tasks even outside the English-speaking area, let alone Arabic due to its unique properties as a language. It had been noticed, for instance, that while ChatGPT performed well in the English language, it was consistently outperformed by its fine-tuned Arabic variants. Consequently, the requirement for exclusive Arabic-oriented models is clearly recognized. Recent benchmarking of LLaMA-3-70B on Arabic Natural Language Generation tasks has revealed its dismal performance compared to closed-source models like ChatGPT and smaller fine-tuned models. These underpin the potential benefit of an Arabic-optimized model, therefore underlining the existing lack in terms of specialized Arabic NLP tasks within the evaluation of open-source LLMs, including medical question answering[6][7].

Arabic QA has been approached with rule-based, machine learning, and deep learning methods. Most of the early systems were basically based on RB approaches; therefore, they used the internal structure of words and sentences, stemming, and part-of-speech tagging, while usually combining IR and NLP techniques to extract relevant passages and extract the answers from structured or semi-structured knowledge bases. Recently, most the studies shift their focus to ML and DL approaches. Most of the recent approaches especially rely on pre-trained language models such as AraBERT, MARBERT, and AraELECTRA. Although these are promising models for Arabic QA, they were proposed for general NLP tasks, not application-specific ones such as medical QA. Several Arabic QA datasets include ARCD, TyDiQA, and AQAD, among others; these datasets usually have limitations to the development of accurate models in their size and quality. Most of these datasets were translated from English, and as such, their quality, when used for training, may be affected. Secondly, most studies in Arabic QA do not make use of benchmark datasets, making it difficult to compare different research works. This also limits the reproducibility of findings and comparative analyses because most systems are not publicly available[5].

Several challenges in Arabic QA are identified from the literature: the complicated morphology and ambiguity of the Arabic language; dialectal variations. Further, there is also limited availability of Arabic resources both in datasets and pre-trained models compared to the number in English. Many of these Arabic datasets are not big, and that impairs the performance of deep learning models. Besides, the quality of existing datasets is also in question. Some datasets originally are translations that may provide inconsistencies and inaccuracies so easily. The Arabic QA domain needs more dedicated research; state-of-the-art performance, barely achieved in general domains, far falls behind that in many more specialized domains, including Medicine. Well, designing effective prompts for LLMs is a crucial activity since the design determines how models interact with users and the quality of the responses through the establishment of operational rules that should remain within established frameworks. The process, therefore, needs to optimize prompts with the aim of enhancing model performance[9].

This review has pointed out that more comprehensive studies are required on the efficiencies of open-source LLMs like LLaMA-3 for specialized Arabic NLP tasks such as medical QA. Indeed, your research will go a long way in adding value to the existing literature on this sphere by testing the efficiency of LLaMA-3-2B across variable conditions and hopefully will yield even better models for Arabic medical question answering and provide recommendations on how Arabic-centric future LLMs could be shaped[8].

III. Methodology

A. Dataset Preparation

The Arabic Healthcare Dataset (AHD) consists of over 800,000 expert-annotated QA pairs across 90 categories. For this study, 3,000 samples were randomly selected by stratified randomg sampling for training, ensuring balanced representation across categories. Preprocessing was applied to preserve the contextual richness of the data by using the AraBERT model, with tokenization and encoding performed using LLaMA's tokenizer for seamless model integration.

B. Zero and Few shot learning

In this work, zero-shot and few-shot learning paradigms were implemented to check the model's generalization and adaptation capabilities with minimal supervision. Few-shot learning has been helpful in providing contextual examples for which the model is supposed to generate accurate and relevant QA pairs.

C. Model Fine-Tuning

The fine-tuning process was conducted on the LLaMA 3.2-3B model, a large language model with 3 billion parameters. This process leveraged the Hugging Face "transformers" library and the PEFT (Parameter-Efficient Fine-Tuning) framework, optimizing the model's performance while maintaining computational efficiency. Tokenization was handled using LLaMA's tokenizer to ensure compatibility with the Arabic script. Fine-tuning was performed over multiple epochs, focusing on enhancing the model's ability to generate precise and contextually relevant answers. Example prompts from the few-shot learning phase were integrated into the training process to further refine the model's understanding of nuanced medical questions.

D. Evaluation Metrics

Traditional metrics like BLEU and ROUGE, often used in NLP tasks, were tested but was found inadequate for evaluating the nuanced requirements of medical QA. As an alternative, GPT-4 was used as an evaluation benchmark, focusing on accuracy, contextual Understanding and clarity, it also compared the answers of the dataset with the generated answers[4].

IV. Results

A. Fine-Tuned Answers

The fine-tuned answers are remarkable in staying as close to the original answers in terms of tone, context, and relevance. This might be indicative of the fine-tuning process, where the model is better placed to understand the context of the dataset it was trained on and adapt to it. These answers are full and reliable; hence, they can be a strong candidate for tasks that require high contextual fidelity. One possible weakness, however, lies in the tendency to sometimes add more than is necessary or which might be superfluous. Such additions may not detract from the overall quality of the response significantly but can water down the conciseness and clarity, especially in straightforward questions. .

B. Few-Shot Answers

The few-shot answers add depth and detail to the generated responses. Though this might be an oversimplification, due to the inclusions of contextual examples during training itself, the model learns to generate answers that, in many cases, are far more informed. The extra context given can make the answers more information-carrying, sometimes even complementing the original answers quite well. While this is a strength, there is also the tendency to digress from the core question or include information irrelevant to the question. This might affect the precision of the answers, especially for those questions that require the answer to be more to the point. The richness of the information provided by the few-shot approach, however, makes it a worthy method for applications where supplementary details are desirable

C. Zero-Shot Answers

Zero-shot answers are either concise or brief, superb in situations where few words might be sufficient. However, they perform worse on elaborative questions that may rely on complex thoughts or local cultures. Not to mention, sometimes the model answers in English instead of Arabic, which diminishes more possibilities of using this model with sophisticated or language-related questions.

D. Performance

Best Overall Performance: Fine-Tuned Answers consistently align most closely with the Answers column. Most Informative: Few-Shot Answers provide additional insights in many cases but require refinement for focus and relevance. Least Reliable: Zero-Shot Answers, while concise, frequently lack the depth needed to fully address the questions.

V. Conclusion

This work showcases the excellent prospects of fine-tuning large language models, like LLaMA 3.2-3B, on Arabic medical question-answering for closing the gap between the language and domain-specific NLP applications. We have implemented both zero-shot and few-shot testing methods in order to assess the model's adaptability and performance. Zero-shot testing showed that the model could provide short and straightforward answers, though at a shallow level, with some responses in English. On the other hand, few-shot learning had more informative responses with enrichment of context, though there was some variation in focus and relevance. Further fine-tuning of the model resulted in even better performance; the responses had a tone, context, and relevance similar to that of the original dataset.

Our results indicate that with minimal available resources, significant strides can be made in adapting large-language models to low-resource languages as a means of promoting health care accessibility and improving NLP research. These results bring to the fore the importance of fine-tuning, prompt engineering, and comparative testing in improving model performance for domain-specific tasks. The work hereby lays a foundation for future research to explore larger datasets, broader evaluation metrics, and multilingual capabilities to further refine the application of LLMs in healthcare and other critical domains.

References

[1] Khondaker, M. T. I., Naeem, N., Khan, F., Elmadany, A., Abdul-Mageed, M. (2024). Benchmarking LLaMA-3 on Arabic Language Generation Tasks. In Proceedings of The Second Arabic Natural Language Processing Conference (pp. 283-297).

[2] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A Pretrained Biomedical Language Model for Biomedical Text Mining. Bioinformatics.

[3] Touvron, H., Lavril, M., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

[4] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the ACL Workshop.

[5] Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.

[6] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

[7] Bhat, M. M., Meng, R., Liu, Y., Zhou, Y., Yavuz, S. (2023). Investigating Answerability of LLMs for Long-Form Question Answering. arXiv preprint arXiv:2309.08210.

[8] Alrayzah, A., Alsolami, F., Saleh, M. (2023). Challenges and opportunities for Arabic question-answering systems: current techniques and future directions. PeerJ Computer Science, 9, e1633.

[9] Ray, S. K., Shaalan, K. (2016). A review and future perspectives of Arabic question answering systems. IEEE Transactions on Knowledge and Data Engineering, 28(12), 3169-3190.

VI. appendix

Question1: انا عندما تأتيني الدورة الشهرية ... هل هذا يؤثر على الغشاء؟ Fine-Tuned Answer: Correctly emphasizes caution but adds unnecessary advice about water temperature. Zero-Shot Answer: Accurate and concise but oversimplified. Few-Shot Answer: Completely irrelevant and talks about hepatitis symptoms. Best: Fine-Tuned Answer (more relevant).

Question2: أخي ربما مصاب بمتلازمة داون ... هل هي حالة نفسية ؟

Fine-Tuned Answer: Attempts to contextualize the situation but lacks actionable advice. Zero-Shot Answer: Makes assumptions about the situation, leading to a generic response. Few-Shot Answer: Irrelevant, mentioning diabetes and arthritis. Best: Fine-Tuned Answer.

Question3: أانا اعاني من اكتئاب وقلق شديد ... دواء فيرين للقولون.

Fine-Tuned Answer: Provides a plausible explanation but lacks direct relevance. Zero-Shot Answer: Misleading advice to stop medication. Few-Shot Answer: Adds value by suggesting vitamin D deficiency but not directly tied to the original Answer. Best: Fine-Tuned Answer (most aligned).

Question4: سؤال للدكتور امجد الحداد ... ولا تستطيع النوم بسبب ذلك

Fine-Tuned Answer: Lacks actionable advice and is vague. Zero-Shot Answer: Repeats the question without providing an answer. Few-Shot Answer: Some relevance but still unclear. Best: None excel; all are vague or irrelevant.

Question5: لو عامله عمليه ترقيع ... وقت الولاده

Fine-Tuned Answer: Off-topic and confusing. Zero-Shot Answer: Simplistic and incomplete. Few-Shot Answer: Adds extraneous details unrelated to the question. Best: Zero-Shot Answer (minimal clarity).

Question6: السلام عليكم انا حاطه تقويم ... يأثر علي اسناني Fine-Tuned Answer: Provides a balanced response emphasizing moderation. Zero-Shot Answer: Direct and correct but lacks contextual richness. Few-Shot Answer: Correct but introduces unrelated details about digestion. Best: Fine-Tuned Answer (most comprehensive).

Question7: س ع انا ام مرضعة اذا اخذت ابرة كبد فيروسي يؤثر على حليبي

Fine-Tuned Answer: Offers practical advice relevant to the context. Zero-Shot Answer: Accurate but oversimplified. Few-Shot Answer: Adds valuable dietary advice, complementing the Answer column. Best: Few-Shot Answer (adds helpful detail).

Question8: ماهي التمارين الفعاله لشد منطقة الكتف والزنود ؟

Fine-Tuned Answer: Correct and well-rounded, aligning with the original Answer. Zero-Shot Answer: Accurate but too generic. Few-Shot Answer: Adds extra details about fish consumption, enhancing context. Best: Few-Shot Answer (more informative).

Question9: نوبة عدم القدره على اخذ نفس ... ويتبعه الاسهال ؟!

Fine-Tuned Answer: Matches the tone and intent of the Answer column perfectly. Zero-Shot Answer: Correct but minimalistic. Few-Shot Answer: Adds slightly more context about disease prevention. Best: Fine-Tuned Answer (closest match).

Question10: اجريت تحليل سكر صائم في الدم وكانت النسبة 691

Fine-Tuned Answer: Highlights relevant causes but introduces unnecessary details about habits. Zero-Shot Answer: Accurate and concise, sticking to the context. Few-Shot Answer: Adds extra explanation about overthinking but remains relevant. Best: Few-Shot Answer (more comprehensive).

Question11: انا حمل بشهر السابع ... ماذا افعل

Fine-Tuned Answer: Covers both physical and mental benefits succinctly. Zero-Shot Answer: Accurate but overly brief. Few-Shot Answer: Adds emphasis on mood improvement, enhancing the answer. Best: Few-Shot Answer (adds valuable detail).

Question12: انا اعاني من حب الشباب مع العلم اني عالجته اختفي لكنه عاد مره اخري فماذا افعل

Fine-Tuned Answer: Correct and comprehensive, emphasizing key foods. Zero-Shot Answer: Accurate but lacks richness. Few-Shot Answer: Matches the Answer column in simplicity but misses depth. Best: Fine-Tuned Answer (most aligned).

Question13: دورتي كانت ٩٢ مارس ... اختبار الدم ليعطي

Fine-Tuned Answer: Thorough and covers multiple health risks. Zero-Shot Answer: Concise and relevant but less detailed. Few-Shot Answer: Adds emphasis on specific diseases like cancer. Best: Fine-Tuned Answer (most detailed).

Question14: لدي ألام في الخصية اليسرى هل هو طبيعي و يوجدعرق عليها عمري 61 عاما

Fine-Tuned Answer: Accurate and covers major causes. Zero-Shot Answer: Correct but overly brief. Few-Shot Answer: Adds genetic and dietary context, improving relevance. Best: Few-Shot Answer (most comprehensive).

Question15: كيف اعالج تحسس طفلي ... احمر.ماذا اعطيه.وشكرا

Fine-Tuned Answer: Practical and well-rounded advice. Zero-Shot Answer: Correct but overly simplistic. Few-Shot Answer: Adds valuable dietary tips, enriching the answer. Best: Few-Shot Answer (more insightful).

Final Observations: Fine-Tuned Answers excel in closely matching the tone and context of the original Answers, often being the most reliable option. Few-Shot Answers provide additional context and information, which is helpful but can sometimes stray from the core intent. Zero-Shot Answers are concise and accurate but often lack depth