Bridging Language Barriers in Healthcare: A Study on Arabic LLMs.

Nada Saadi, Tathagata Raha, Clément Christophe,
Marco AF Pimentel, Ronnie Rajan, Praveen K Kanithi

M42 Health, Abu Dhabi, UAE

Abstract

This paper investigates the challenges of developing large language models (LLMs) proficient in both multilingual understanding and medical knowledge. We demonstrate that simply translating medical data does not guarantee strong performance on clinical tasks in the target language. Our experiments reveal that the optimal language mix in training data varies significantly across different medical tasks. We find that larger models with carefully calibrated language ratios achieve superior performance on native-language clinical tasks. Furthermore, our results suggest that relying solely on fine-tuning may not be the most effective approach for incorporating new language knowledge into LLMs. Instead, data and computationally intensive pretraining methods may still be necessary to achieve optimal performance in multilingual medical settings. These findings provide valuable guidance for building effective and inclusive medical AI systems for diverse linguistic communities.

Introduction

The evolution of multilingual language models like Llama3 and GPT-4 marks a significant advancement in natural language processing. However, most LLMs are primarily trained on English and other common European languages, often neglecting low-resource languages with different alphabets, such as Arabic. This limitation poses a significant challenge, particularly in specialized domains like healthcare, where accurate language understanding is crucial.

One major obstacle is the scarcity of high-quality, domain-specific data for these languages. In this paper, we address this challenge by evaluating and improving the capabilities of LLMs for clinical tasks in Arabic. We first conduct a comprehensive evaluation of existing open-source LLMs to assess their performance on medical tasks in both English and Arabic. This analysis provides valuable insights into the current state of LLMs in handling clinical information across different languages.

To further enhance the capabilities of these models, we investigate various techniques, including leveraging existing LLMs for translation, paraphrasing, and generating synthetic data to augment Arabic medical datasets. Specifically, we explore the translation capabilities of both Llama and Qwen, highlighting their strengths and weaknesses in handling medical terminology and nuances in Arabic.

Finally, we fine-tune Llama 3.1 using different mixtures of original and synthetic Arabic medical data. This allows us to analyze the impact of different data sources and augmentation techniques on the model's performance across various clinical tasks. Our findings reveal that the optimal data mixture varies depending on the specific task, emphasizing the importance of careful data curation and augmentation strategies for developing effective clinical LLMs.

Our work focuses on the Llama 3.1 model, but the proposed methodology can be extended to other LLMs, domains, and low-resource languages. We believe this research contributes to developing more inclusive and robust LLMs that can effectively serve diverse linguistic communities and specialized domains.

Related Work

The emergence of large language models (LLMs) has marked a significant advancement in artificial intelligence, demonstrating impressive capabilities in natural language understanding and generation. Initially developed for general use-cases such as text summarization, translation, and dialogue generation, LLMs have quickly been adopted across diverse industries, including finance, law, and education.

One domain where LLMs have shown considerable promise is healthcare (Zhang, Wang, and Chen 2023). Recent studies have explored the application of LLMs to a variety of medical tasks, including clinical decision support, medical question answering, and diagnosis assistance. For instance, GPT-4 has demonstrated proficiency in medical knowledge evaluation, achieving scores comparable to human experts on standardized medical exams (Nori et al. 2023). Other models like Meditron (Chen et al. 2023), OpenBioLLM (Ankit Pal 2024) and Med42 (Christophe et al. 2024) have further advanced the field, with many surpassing GPT-4's performance on specific medical tasks and releasing open-source models usable by the research community and facilitating further advancements in the field.

Evaluating the performance of clinical LLMs, however, presents unique challenges. Most current models are primarily evaluated on question answering tasks using datasets like USMLE, MedQA, and PubMedQA. New benchmarks such

as MEDIC (Kanithi et al. 2024) have emerged to provide a more comprehensive and standardized evaluation of medical LLMs, encompassing tasks like clinical diagnosis, treatment recommendation, and patient education.

Despite the rapid progress, a significant limitation of most existing clinical LLMs is their reliance on English data for training and evaluation. This raises concerns about their applicability and fairness in multilingual healthcare settings, where a significant portion of the population may not be proficient in English. While multilingual LLMs like Llama, Qwen or Mistral (Dubey et al. 2024; Yang et al. 2024; Jiang et al. 2023) have demonstrated strong performance across multiple languages, adapting these models for both a new language and a specialized domain like healthcare remains an active area of research. Although some efforts have been made to develop multilingual clinical LLMs (Lopez et al. 2023; Wang et al. 2024), broad studies on their cross-lingual performance and generalizability are limited.

Developing LLMs for languages with unique linguistic characteristics and limited digital resources presents additional challenges. Arabic, with its complex morphology, dialectal variations, and relatively scarce medical corpora, exemplifies these challenges. While general-purpose Arabic LLMs like Jais (Sengupta et al. 2023) and Silma (Silma-AI 2024) have emerged, specialized medical models for Arabic remain scarce. To our knowledge, BiMedix (Pieri et al. 2024a) is the only model that specifically focuses on building an LLM with multilingual capabilities (Arabic + English) for the healthcare domain.

Evaluating Large Language Models' Capabilities in Arabic Medical Applications

While many large language models claim to work well in multiple languages and medical tasks, most testing focuses on either general language skills or English medical knowledge separately. Few studies look at how well these models handle medical content in specific non-English languages (Jin et al. 2024). In this study, we test several popular models of different sizes on Arabic medical benchmarks. Our results show that these models still have a long way to go to match their English performance levels.

Arabic Evaluation Datasets

We utilize a set of Arabic-translated medical datasets for our zero-shot and fine-tuning evaluations. These datasets, originally developed for training and evaluating question-answering systems in the medical domain, include PubMedQA, MedMCQA, MedQA, and Medical MMLU. **PubMedQA**: This dataset is derived from biomedical research articles (Jin et al. 2019). **MedMCQA** (Pal, Umapathi, and Sankarasubbu 2022) and **MedQA** (Jin et al. 2021) consists of multiple-choice questions coming from Indian and United States medical license exams. Medical **MMLU** is derived from the MMLU benchmark, specifically focusing on biomedical subsets including Clinical Knowledge, College Biology, College Medicine, Medical Genetics, Professional Medicine, and Anatomy (Hendrycks et al. 2021).

The translation of these datasets into Arabic was conducted using a semi-automated iterative translation pipeline, as detailed in BiMediX (Pieri et al. 2024a). This process involves initial translations using language models, followed by human refinement to ensure the accuracy and quality of the translations. The translated datasets maintain the original format of questions and answers, allowing for consistent evaluation across languages.

Modifications to Harness Pipeline

Our research utilizes the Harness evaluation framework (Gao et al. 2024), which calculates log-likelihood scores to evaluate predictive model performance. To accommodate the Arabic language, we made significant modifications to handle its unique attributes, including its distinctive script, complex morphology, and syntactic structure, ensuring accurate processing of Arabic data.

Arabic text runs right-to-left (RTL), unlike English (LTR). We updated the framework to display and process Arabic text correctly. This meant reformatting our dataset while keeping its structure intact. To conduct zero-shot evaluations in Arabic, we adapted the entire framework for zero-shot testing, converting all prompts, responses, and multiple-choice options to Arabic, ensuring an accurate display and functionality of the Arabic script. Arabic's linguistic complexity makes context crucial for accurate understanding. A single word can have multiple meanings depending on its grammatical form and context. For instance, the root word جرب (j-r-b) can transform into various derivatives that range from "to try" to "to test" to other nuanced meanings, highlighting why machine learning models must carefully consider contextual cues when processing Arabic text.

Additionally, rather than just calculating the probability of generating answer choice labels (e.g., *a*, *b*, *c*, or *d*), we calculate the probability of generating the full answer text. This modification provides a more detailed understanding of the model's performance by taking into account the entire answer generation process.

Results

As shown in Table 1, large language models in all model families exhibit limited performance on Arabic medical benchmarks. While leading models like Llama3.1 achieve high accuracy in English (62.0 and 78.2 on MedQA), their performance significantly degrades when applied to Arabic (29.5 and 56.6). Although Qwen2.5 models demonstrate relatively better performance in Arabic, accuracy remains suboptimal.

We will focus on improving Arabic performance, using Llama3.1 as a case study to explore strategies to achieve English-language proficiency on Arabic medical benchmarks.

LLM Adaptation Through Translation Pipeline

A straightforward approach to enhance large language model performance across languages is to implement a translation pipeline: convert the Arabic input to English, process it, and translate the output back to Arabic. This

Model/Dataset	PubMedQA		MedMCQA		MedQA		MMLU	
	En	Ar	En	Ar	En	Ar	En	Ar
Qwen2.5-3B-Instruct (Yang et al. 2024)	29.2	61.2	49.2	35.5	48.8	41.7	68.0	28.0
Qwen2.5-7B-Instruct (Yang et al. 2024)	45.2	**74.4**	56.8	**39.5**	60.2	53.9	**76.7**	**34.9**
Pangea-7B (Yue et al. 2024)	57.0	61.0	50.2	37.5	53.0	49.6	68.3	32.4
Mistral-7B-Instruct_v0.3 (Jiang et al. 2023)	45.8	46.6	46.3	28.0	49.3	33.8	65.1	21.6
Llama3.1-8B-Instruct (Dubey et al. 2024)	**76.2**	73.2	**58.4**	35.8	**62.0**	29.5	73.4	46.4
Silma-9B-Instruct-v1.0 (Silma-AI 2024)	75.6	64.0	54.9	38.9	61.6	**54.7**	76.1	31.5
Llama-3.1-70B-Instruct (Dubey et al. 2024)	73.6	**79.4**	71.8	52.2	78.2	56.6	**87.6**	70.0
Qwen2.5-72B-Instruct (Yang et al. 2024)	63.2	76.6	68.4	**56.9**	76.1	**76.1**	87.4	**76.1**
Med42-Llama3.1-70B (Christophe et al. 2024)	77.6	75.0	**72.4**	49.3	**80.4**	53.5	86.8	67.7
Meditron3-70B (Chen et al. 2023)	**80.6**	75.8	70.9	51.2	79.3	72.0	87.0	56.6
BiMedix(Bilingual) (Pieri et al. 2024b)	77.2	78.4	61.6	49.1	65.2	47.3	73.2	56.9

Table 1: Accuracy of publicly available models on different Medical QA benchmarks. Even though Llama3.1 models are performing better in English, Qwen2.5 models show a stronger performance in Arabic.

Translation Model	PubMedQA	MedMCQA	MedQA	MMLU
LlamaX (Lu et al. 2024)	74.6	53.1	55.8	59.5
Helsinki (Helsinki-NLP 2024)	72.0	48.9	40.8	56.6
Flores 101 (seyoungsong 2024)	72.0	36.6	31.2	34.0
Llama3.1-70B-Instruct (Dubey et al. 2024)	75.8	54.8	70.5	70.7
Qwen2.5-72B-Instruct (Yang et al. 2024)	**75.9**	**55.2**	**71.3**	**71.5**

Table 2: Performance comparison of various translation models on Arabic medical benchmarks, translated into English and evaluated using Llama3.1-70B-Instruct for accuracy (%).

method leverages the models' strong English capabilities. However, this approach introduces significant computational overhead, which requires at least three separate model calls instead of one.

We evaluated this pipeline's effectiveness by testing various translation models and comparing Llama-3.1-70B's performance on translated content against its native English capabilities. Using our established evaluation benchmarks, we translated both questions and multiple-choice options before processing them through Llama-3.1-70B, maintaining consistent evaluation methods.

Our investigation included two categories of translation systems: specialized translation models designed for precise Arabic-English conversion, and general-purpose large language models trained on both languages. Though not specifically optimized for translation, these general-purpose models offer broader language understanding.

Our results in Table 2 reveal that despite their reputation for accuracy, specialized translation models faced considerable challenges with medical content. The nuanced nature of medical terminology makes literal translations problematic, often resulting in technically accurate but contextually inappropriate translations. General-purpose models such as Llama and Qwen demonstrated superior performance in this domain, producing translations that better preserved both technical accuracy and medical context. Although translation pipelines can improve the performance of LLMs in Arabic medical content, the results still lag significantly behind their native English capabilities, highlighting the imperfections of current translation methods. This discrepancy raises serious concerns in the healthcare domain, where accurate understanding and generation of medical information is crucial. Furthermore, the added computational overhead of translation limits the feasibility of deploying such models in resource-constrained environments, hindering their accessibility in regions where they are most needed.

Language Specific Finetuning

We aim to improve large language models' performance by finetuning on bilingual domain-specific data not encountered during pretraining. In this section, we detail our finetuning pipeline, present the datasets utilized, and analyze the optimal balance between English and Arabic training data.

Finetuning Datasets Due to the scarcity of high-quality Arabic clinical data, we developed a comprehensive data preparation pipeline. This section details our methodology for cleaning existing datasets, generating new data, and performing translations to ensure robust data quality. The size of each dataset is described in Table 3.

- **Arabic Health Questions & Answers Dataset (AHQAD)**: The AHQAD dataset, with its 90 richly diverse categories, offers a comprehensive landscape of medical and healthcare-related themes tailored specifically for the Arabic-speaking region. This collection spans an extensive array of topics, from general medicine

Dataset	Original	Description	Final	# of Tokens
1. AHQAD	Arabic	100K sampled based on completeness and paraphrased with Qwen-72B-Instruct	Arabic	8.33 M
2. Translated MED42 Dataset	English	500K sampled randomly, cleaned and translated with Qwen-72B-Instruct	Arabic	230.69 M
3. CIDAR	Arabic	10K Instruction-Output good quality dataset	Arabic	1.34 M
4. Med42 Dataset	English	Full English FT dataset	English	464.97 M
5. Synthetic Open-Ended	English	˜200K sampled based on 1-5 rating and translated with Qwen-72B-Instruct	Arabic	240 M
Total # of Arabic Tokens				**480.36 M**
Total # of English Tokens				**469.97 M**

Table 3: Dataset Overview with Language Distribution

to specialized fields such as cardiology, pediatrics, and oncology, as well as practical areas like pharmacology and patient care. It also includes emergent fields such as telemedicine and health informatics.

For the AHQAD dataset, which comprises medical queries, we applied a more selective filtering process. Out of its total 298,000 entries, we chose 100,000 that featured the most complete questions. This was crucial as it ensured the data's clarity and relevance, enhancing the quality of the training material. Responses were standardized using Qwen2.5-72B model for paraphrasing. We instructed the model to rewrite responses while maintaining the original meaning, removing typos and complex abbreviations, and improving clarity. The model was explicitly prompted to preserve the original information without adding any new content.

- **Translated Med42 Dataset**: We leverage the finetuning dataset used for finetuning Med42-v2. The Med42 dataset is curated from various medical and biomedical resources and it also features chat interactions and chain-of-thought reasoning apart beyond simple question-answering.

  For the Med42 dataset, we randomly selected 500,000 records from the dataset used to train Med42 (Christophe et al. 2024). This dataset was then translated to Arabic using the Qwen2.5-72B-Instruct model, chosen for its strong performance on Arabic benchmarks, as demonstrated in Table 1. Our manual evaluation further confirmed that its translations are of higher quality, with superior context preservation, both in medical and non-medical scenarios. Following translation, we meticulously cleaned the data to ensure only Arabic samples were retained, discarding any residual English phrases, ultimately preserving approximately 90% of the dataset.
- **Culturally Relevant Instruction Dataset For Arabic (CIDAR)**: The CIDAR dataset (Alyafeai et al. 2024) on Hugging Face is an instruction-output pair dataset explicitly crafted for Arabic NLP tasks, making it particularly valuable for training models in zero-shot and few-shot learning scenarios. Each record in the dataset features a distinct instruction—a prompt, question, or directive in Arabic—and a corresponding output that provides a precise, contextually relevant response. This structure is intended to help models interpret and generate responses across various types of tasks, such as factual questions, conversational exchanges, and directive-based commands, thereby enhancing the model's instruction-following capabilities.
- **Synthetic Open-Ended QA**: The preparation of healthcare-specific synthetic question-answer pairs employs a systematic multi-stage approach. The process begins by randomly selecting seed questions from HealthSearchQA, ExpertQA, and MedicationQA datasets. These seed questions serve as the foundational examples for iteratively building a synthetic instruction dataset. The iterative process involves using the seed instructions as few-shot examples to generate new synthetic instructions. With each iteration, the pool of instructions expands, ensuring diversity and coverage across healthcare topics. Importantly, to preserve data independence, the final synthetic instruction dataset excludes the original seed instructions. The generated questions are then processed using Llama-3.1-70B-Instruct to create comprehensive responses. To ensure quality, these responses undergo evaluation using the same model on a scale of 1 to 5, with only pairs rated 5 being retained in the final dataset to ensure high quality. The entire dataset is subsequently translated into Arabic using Qwen2.5-72B-Instruct.

Fine-tuning Pipeline The bilingual medical fine-tuning pipeline explores the optimal combination of Arabic and English medical datasets to enhance model performance on both English and Arabic medical tasks. The pipeline incorporates two distinct data streams: high-quality Arabic medical content obtained through rigorous cleaning and filtering of native Arabic medical datasets, and carefully translated English medical datasets that maintain clinical accuracy in both languages. For different ratios, we maintain a constant number of 469.97M tokens, randomly sampling from our dataset presented in Table 3.

We finetuned both Llama3.1 models. We employ the classic auto-regressive loss for finetuning. Loss is backpropagated only on output tokens. This approach ensures that the

Model Configuration	Dataset Ratio (Arabic-English)	Accuracy							
		PubMedQA		MedMCQA		MedQA		MMLU	
		En	Ar	En	Ar	En	Ar	En	Ar
Llama 3.1 8B	Baseline	38.0	34.4	50.0	32.2	55.1	27.3	64.8	39.8
1. Arabic Only	100%-0%	68.6	**71.2**	56.6	32.2	55.8	27.5	72.5	40.4
2. Strong Arabic Majority	80%-20%	**73.6**	68.0	56.7	**35.1**	57.4	27.5	71.1	40.2
3. Arabic Majority	60%-40%	69.2	63.0	56.6	32.9	57.9	28.3	**72.9**	40.9
4. Balanced Distribution	50%-50%	70.0	61.2	56.6	34.7	**58.7**	29.5	71.9	42.3
5. English Majority	40%-60%	59.0	53.8	57.0	33.4	57.7	**29.8**	70.8	42.0
6. Strong English Majority	20%-80%	67.0	53.8	57.6	33.4	57.3	28.7	71.7	42.0
7. English Only	0%-100%	72.8	61.2	**58.8**	34.7	58.5	29.5	71.62	**42.4**
Llama 3.1 8B-Instruct	Baseline	76.2	**73.2**	58.4	35.8	**62.0**	29.5	73.4	**46.4**
1. Arabic Only	100%-0%	72.0	72.0	58.7	31.9	57.9	29.2	73.6	30.0
2. Strong Arabic Majority	80%-20%	76.6	69.0	58.4	34.7	59.6	29.7	72.8	42.1
3. Arabic Majority	60%-40%	**76.8**	68.4	58.1	33.7	60.3	30.9	73.1	43.0
4. Balanced Distribution	50%-50%	71.2	66.2	58.7	34.6	60.2	**33.5**	**73.7**	42.5
5. English Majority	40%-60%	74.8	66.2	59.1	**36.8**	61.9	33.5	72.3	42.1
6. Strong English Majority	20%-80%	73.4	64.0	**59.5**	35.1	60.8	31.6	73.3	42.8
7. English Only	0%-100%	68.4	60.0	**59.5**	36.4	60.6	30.2	73.6	46.3
Llama 3.1 70B	Baseline	15.6	51.8	65.1	41.6	75.9	48.2	82.4	55.8
1. Arabic Only	100%-0%	70.6	**76.8**	68.8	**51.7**	75.3	**53.3**	84.3	**67.3**
2. Balanced Distribution	50%-50%	**78.0**	55.0	68.9	50.5	**76.9**	50.9	84.0	59.3
3. English Only	0%-100%	75.2	48.6	**70.2**	47.5	76.1	48.6	**85.7**	61.9
Llama 3.1 70B-Instruct	Baseline	73.6	**79.4**	**71.8**	52.2	**78.2**	**56.6**	**87.6**	**70.0**
1. Arabic Only	100%-0%	**79.8**	78.2	70.6	**52.8**	77.1	55.8	86.3	67.1
2. Balanced Distribution	50%-50%	77.2	61.8	71.4	50.8	76.2	55.6	86.9	67.0
3. English Only	0%-100%	75.0	48.2	71.7	49.7	77.8	52.9	87.5	66.5

Note: Results show performance on English (En) and Arabic (Ar) evaluations for each metric.

Table 4: Accuracy of Finetuned Llama3.1-8b and 70b models with Different Arabic-English Dataset Ratios on medical QA benchmarks. Fine-tuning Llama 3.1 models on varying Arabic-English dataset ratios yields inconsistent results across medical QA tasks. Even large instruct models show limited improvement on Arabic benchmarks after fine-tuning.

model learns to generate appropriate responses and not learn to generate the prompts. Our training samples are concatenated into chunks of 8192 tokens. Each model was finetuned for two epochs over our curated dataset using a cosine learning rate schedule between 1×10^{-5} and 1×10^{-6}. All experiments are performed on a cluster of 4 H100 nodes.

Results Our results in Table 4 show that different ratio of Arabic-English data yield to different performance levels depending on the evaluation task. For PubMedQA, training with exclusively Arabic data produces the best accuracy (71.2). While for MedMCQA and MedQA, the models perform best with strong Arabic majority and English Majority, respectively (35.1 and 29.8). Surprisingly, for the MMLU datasets, which focuses on testing direct knowledge application, using only English data, achieves 42.4 compared to the 39.8 zero-shot accuracy.

These patterns remain consistent across both the base and instruct models. These results highlight the fact that the relationship between the language distribution used for finetuning and performance is fundamentally linked to the nature of each task. For instance, PubMedQA requires complex analytical reasoning within medical contexts, while MMLU focuses on structured knowledge assessment through multiple-choice questions. This difference suggests that tasks requiring a deeper understanding of context need stronger language-specific training.

Interestingly, our findings on the larger 70B parameter models show more consistent behavior across all Arabic tasks, with Arabic-only training data consistently achieving the best results. This suggests that larger models may handle language-specific tasks more uniformly than their smaller counterparts, given their greater capacity to abstract and generalize linguistic features across different training distributions. Llama3.1-70B base model exhibits poor performance on the PubMedQA test set due to its inability to follow the chat-template for a highly context-based task. Intriguingly, our fine-tuned version of Llama3.1-70B-Instruct rarely outperforms the original model, suggesting that it has reached its maximum capabilities after subsequent pretraining, supervised fine-tuning, and alignment stages.

Conclusion

Our research into Arabic-English medical AI reveals critical insights for developing truly effective multilingual language models.

First, we highlight the significant performance gap between English and Arabic, especially pronounced in smaller models. This disparity underscores the need for models deeply trained in specific languages to achieve genuine language understanding and complex medical reasoning. Smaller models, with their limited capacity, struggle to capture the nuances of different languages and medical terminology, resulting in a substantial performance gap between languages like English and Arabic.

Second, while some general language models demonstrate superior translation capabilities compared to specialized translation models, they come with high computational costs and are not perfect. General language models, despite their broader training data, still have limitations in accurately translating medical terminology and complex linguistic structures, highlighting the need for further research in this area.

Third, fine-tuning models do not always guarantee improved performance compared to the baseline, and the results are highly dependent on the distribution of languages in the training data. The effectiveness of fine-tuning can vary significantly depending on the specific language mix used in the training data, suggesting that a careful balance of languages is crucial for optimal performance.

We acknowledge that our evaluation primarily focuses on close-ended question benchmarks, which, while valuable for assessing domain knowledge, do not fully capture the generation capabilities, safety, and bias aspects of a model. These aspects are crucially important for any healthcare model. Therefore, we advocate for new benchmarks, such as MEDIC (Kanithi et al. 2024), to include multilingual capabilities tests to address these critical dimensions.

It is important to note that the impact of language mixing is particularly significant when dealing with languages that have vastly different alphabets, such as Arabic, Chinese, or Latin-based languages. The non-overlapping nature of their tokens can lead to unique challenges in training and optimization. Moreover, the performance of a model in one domain should ideally transfer seamlessly across multiple languages. It should be easier for a model to learn technical vocabularies in a new language if it is already trained on that domain and possesses a good understanding of the language. Therefore, the transfer capabilities of a model for a specific domain from one language to another should be high.

Thus, we need to continue relying on extensive pretraining for models to learn a new language effectively. At the same time, exploring the transfer capabilities of models for specific domains across languages is crucial. The ultimate goal extends beyond technical achievement: we aim to create AI systems that can break down language barriers, provide accurate medical insights, and expand healthcare access, especially in underserved and linguistically diverse communities. This means developing models that do not just translate words, but truly comprehend the intricate cultural and linguistic subtleties of medical communication. Achieving this goal will require models that can not only translate medical information accurately but also understand the cultural context and linguistic nuances associated with different languages, ensuring effective communication and healthcare access for diverse populations.

References

Alyafeai, Z.; Almubarak, K.; Ashraf, A.; Alnuhait, D.; Alshahrani, S.; Abdulrahman, G. A. Q.; Ahmed, G.; Gawah, Q.; Saleh, Z.; Ghaleb, M.; Ali, Y.; and Al-Shaibani, M. S. 2024. CIDAR: Culturally Relevant Instruction Dataset For Arabic. arXiv:2402.03177.

Ankit Pal, M. S. 2024. OpenBioLLMs: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; Sallinen, A.; Sakhaeirad, A.; Swamy, V.; Krawczuk, I.; Bayazit, D.; Marmet, A.; Montariol, S.; Hartley, M.-A.; Jaggi, M.; and Bosselut, A. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint arXiv:2311.16079*.

Christophe, C.; Kanithi, P. K.; Raha, T.; Khan, S.; and Pimentel, M. A. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2024. A framework for few-shot language model evaluation.

Helsinki-NLP. 2024. opus-mt-ar-en. https://huggingface.co/Helsinki-NLP/opus-mt-ar-en.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577.

Jin, Y.; Chandra, M.; Verma, G.; Hu, Y.; De Choudhury, M.; and Kumar, S. 2024. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, 2627–2638. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701719.

Kanithi, P. K.; Christophe, C.; Pimentel, M. A. F.; Raha, T.; Saadi, N.; Javed, H.; Maslenkova, S.; Hayat, N.; Rajan, R.; and Khan, S. 2024. MEDIC: Towards a Comprehensive Framework for Evaluating LLMs in Clinical Applications. arXiv:arXiv:2409.07314.

Lopez, M.; Parikh, A.; Yacouby, R.; and et al. 2023. GatorTron-M: Multilingual Clinical Language Models for Medical Natural Language Processing.

Lu, Y.; Zhu, W.; Li, L.; Qiao, Y.; and Yuan, F. 2024. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages. *arXiv preprint arXiv:2407.05975*.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Flores, G.; Chen, G. H.; Pollard, T.; Ho, J. C.; and Naumann, T., eds., *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, 248–260. PMLR.

Pieri, S.; Mullappilly, S. S.; Khan, F. S.; Anwer, R. M.; Khan, S.; Baldwin, T.; and Cholakkal, H. 2024a. Bimedix: Bilingual medical mixture of experts llm. *arXiv preprint arXiv:2402.13253*.

Pieri, S.; Mullappilly, S. S.; Khan, F. S.; Anwer, R. M.; Khan, S.; Baldwin, T.; and Cholakkal, H. 2024b. BiMediX: Bilingual Medical Mixture of Experts LLM. arXiv:2402.13253.

Sengupta, N.; Sahu, S. K.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Marshall, W.; Gosal, G.; Liu, C.; Chen, Z.; et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

seyoungsong. 2024. flores101mm100175M. https://huggingface.co/seyoungsong/flores101-_mm100_175M.

Silma-AI. 2024. SILMA 1.0. https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0.

Wang, X.; Chen, N.; Chen, J.; Hu, Y.; Wang, Y.; Wu, X.; Gao, A.; Wan, X.; Li, H.; and Wang, B. 2024. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yue, X.; Song, Y.; Asai, A.; Kim, S.; de Dieu Nyandwi, J.; Khanuja, S.; Kantharuban, A.; Sutawika, L.; Ramamoorthy, S.; and Neubig, G. 2024. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages. *arXiv preprint arXiv:2410.16153*.

Zhang, X.; Wang, Y.; and Chen, H. 2023. Large Language Models in Medicine: The Potentials and Pitfalls.

Appendix

Translation Comparison for Medical Terminology

Helsinki Translation

Original Translation:

How big is the bottle you're gonna use in a patient who needs a quick blood transfusion (based on medical knowledge in 2020)?

Options:

- 18 Gg.
- 20 Gs.
- 22 Gig.
- 24 Gg.

Flores-101 Translation

Original Translation:

What size of cannula would you use in a patient who needed a rapid blood transfusion (as of 2020 medical knowledge)?

Options:

- 18 gauge
- 20 gauge
- 22 gauge
- 24 gauge

LlamaX Translation

Original Translation:

What is the volume of the cannula you will use for a patient who needs a rapid blood transfusion (according to medical knowledge in 2020)?

Options:

- 18 qij.
- 20 qij.
- 22 qij.
- 24 qij.

Analysis

Translation Quality Comparison

- **Helsinki:** Uses informal language (”gonna”) and confuses the medical terminology by referring to a ”bottle” instead of a cannula. Unit notation is inconsistent (Gg., Gs., Gig.).
- **Flores-101:** Provides the most accurate medical terminology, using ”cannula” and ”gauge” correctly. Maintains consistent formatting and professional medical language.
- **LlamaX:** Uses correct medical term ”cannula” but focuses on ”volume” rather than size. Unit notation shows consistent but incorrect translation (qij.).

Qwen Translation Example

Input Context:
Original English Text:

Why do some people develop hypothyroidism after radioactive iodine treatment for thyroid cancer? Radioactive iodine treatment for thyroid cancer can sometimes lead to decreased thyroid function in some individuals. This occurs due to the damage caused by radiation to the thyroid gland cells.

Qwen Translation:

لماذا يعاني بعض الأشخاص من نقص نشاط الغدة الدرقية بعد العلاج الإشعاعي لسرطان الغدة الدرقية؟العلاج الإشعاعي لسرطان الغدة الدرقية قد يؤدي في بعض الأحيان إلى فرط نقص الغدة الدرقية في بعض الأفراد. يحدث هذا بسبب الضرر الذي يسببه الإشعاع للخلايا في الغدة الدرقية.