Multilingual Medical Language Models: A Path to Improving Lay Health Worker Effectiveness

Agasthya Gangavarapu
Eastlake High School
Sammamish, Washington 98074 USA
august@safety4xr.org

Abstract

The COVID-19 pandemic has exacerbated the challenges faced by healthcare delivery in developing nations, placing additional strain on already fragile infrastructure and healthcare systems. This has prompted an increased reliance on lay healthcare workers (LHWs) to meet the surging demand for services. Due to limited formal training, many LHWs have resorted to using unreliable sources, such as internet searches, to access medical information.

Large language models (LLMs) offer a promising opportunity to support LHWs by providing accurate, context-sensitive information for improving healthcare delivery, provided they are appropriately fine-tuned on domain-specific multilingual data. This paper delves into critical issues and presents potential solutions for developing LLM-powered virtual assistants tailored to LHWs serving Telugu and Hindi-speaking populations. Key focal points include the customization of language and content to suit local contexts, the integration of feedback mechanisms to continuously enhance assistance quality, and the delicate balance between automation and human oversight.

Introduction

Lay Health Workers (LHWs) have evolved into indispensable pillars of healthcare delivery in developing nations, carrying the weight of responsibility for more than 80% of outpatient and maternal services in rural areas. Their role has grown even more pivotal during the COVID-19 pandemic, driven by global financial constraints that have necessitated significant cutbacks in public health expenditures. Confronted with limited budgets ill-equipped to meet the surging healthcare demands, developing nations have increasingly turned to networks of LHWs, drawn by their cost-effectiveness and remarkable ability to serve communities with precision.

Empowering LHWs with access to high-quality, contextually relevant healthcare information and best practices has risen to the forefront as a strategic imperative for countless rural communities worldwide. Large Language Models (LLMs) stand as promising reservoirs of knowledge and indispensable aides for LHWs, provided they undergo appropriate fine-tuning and customization with pertinent knowledge. Nevertheless, it's crucial to acknowledge that the current efficacy of LLMs as aids to LHWs faces constraints due to their resource-intensive nature and the potential for cultural misalignment.

To enhance the effectiveness of LLM assistants for LHWs, three critical actions are essential:

1. Contextual Training in Native Languages: LLMs must undergo rigorous training using contextual medical data in native languages, ensuring that they can effectively communicate and resonate with local communities.
2. Establish Effective Guardrails: Mechanisms must be in place to prevent overreliance on LLMs, ensuring that LHWs maintain their clinical judgment and decision-making capabilities.
3. Enhance Cost-Effectiveness for Scalability: Efforts should be made to make LLMs more cost-effective for broad adoption, enabling their widespread use as invaluable tools for healthcare support.

By addressing these key considerations, we can unlock the full potential of LLMs as invaluable tools for supporting Lay Health Workers in their critical mission of expanding healthcare access and enhancing its quality for underserved populations.

Contextual Training in Native Langauge

In current LLMs, a notable limitation is their suboptimal performance in low-resource languages such as Telugu, Hindi, and Bangla, primarily due to their predominant training on English-language datasets. This shortcoming becomes particularly evident in specialized domains like healthcare, where these models often render responses that are either incoherent or erroneous. Addressing this gap through localized datasets and individualized training in each local language proves to be impractical, largely owing to the extensive costs associated with data curation and the fine-tuning process for each of the languages and resulting operational costs.

In response to this challenge, I have devised a novel pipeline, leveraging baseline medical dialog datasets in English obtained from a variety of public repositories (Chen et al. 2020). This foundational dataset is then enhanced with elements reflecting local cultural and linguistic idiosyncrasies, thereby ensuring its relevance and applicabil-

ity in diverse linguistic contexts. This innovative approach maintains a uniform baseline dataset while strategically incorporating region-specific and linguistic nuances. Such a methodology not only facilitates scalability and operational efficiency but also ensures the delivery of accurate and contextually nuanced responses across a spectrum of linguistic environments. Additionally, this modular framework affords the flexibility to independently refine each constituent element - ranging from the baseline medical dialog datasets and the integration of local cultural subtleties to the fine-tuning process - thereby enabling comprehensive and targeted enhancements without impinging on the integrity of other components.

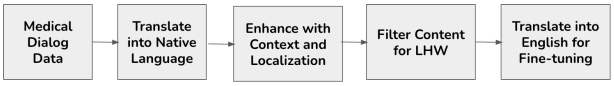

Figure 1: LHW focused data curation

Enhancing LHW assistant ecosystem through this approach offers multifaceted benefits. Firstly, it enables the leveraging of high-performing, open-sourced medical Large Language Models (LLMs) as opposed to constructing bespoke models from the ground up. This utilization of pre-existing, rigorously developed LLMs not only taps into their advanced capabilities but also ensures a foundation of extensive, diverse data, crucial for effective healthcare delivery.

Secondly, the flexibility of this system allows for the periodic replacement and upgrading of LLMs with more advanced iterations as they become available. This dynamic adaptability is key in maintaining the cutting-edge relevance of the LHW assistant tools, ensuring that healthcare professionals have access to the most current and efficient AI-driven resources.

Lastly, the architecture of this approach is inherently cost-effective. By integrating open-source LLMs and employing a scalable model that adds local nuances to a standard English-language baseline, the need for extensive resource allocation towards individual model development for each language is significantly reduced. This economical deployment strategy not only makes the technology more accessible, especially in resource-limited settings but also ensures that the focus remains on enhancing healthcare delivery rather than on the prohibitive costs of technology development.

Essential Guardrails

While advanced techniques like Red-teaming (Perez et al. 2022) and Reinforcement Learning with Human Feedback (RLHF) have significantly bolstered the resilience of core LLMs, mitigating issues such as hallucination and guarding against adversarial attacks like jailbreaking (Deng et al. 2023), it's crucial to adopt a programmatic approach to establish dynamic safeguards. Moreover, the conventional warnings about hallucination prove less effective for LHWs due to the potential psychological impacts they can have on users.

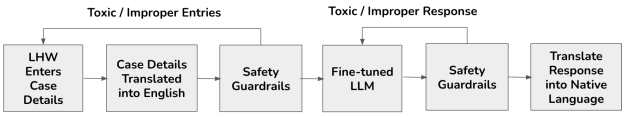

Figure 2: Ineference Pipeline with guardrails

In this evolving landscape, I have employed NVidia NeMo Guardrails (Saunders 2023) to implement multiple protective measures. These measures aim to prevent prompts from users that could lead to adverse outcomes, reject toxic user inputs, and maintain focus on relevant topics. Leveraging these guardrails in a programmatic manner has shown promise in delivering enhanced safety; however, it comes at a cost of a 15% to 20% impact on inference performance. Additionally, it's important to note that NeMo and similar guardrail libraries currently primarily support the English language, limiting their applicability primarily within the core LLM context.

Cost-effective Deployment & Scaling

Cost-effectiveness is crucial for healthcare delivery in remote areas of developing countries, where infrastructural costs vary significantly. The 'Fine-tune Once, Deploy Many' strategy is pivotal in reducing these expenses and enabling scalable healthcare solutions.

Optimization techniques like model compression, quantization, pruning, and knowledge distillation are key in minimizing model size and computational demands, thereby lowering deployment and inference costs. Among these, my research focused on knowledge distillation, applied to the test model, called *Ayurllama*, using a base medical dataset. This approach significantly improved performance and enabled deployment on a single GPU.

By integrating data curation, contextual adaptation, and optimization strategies, especially knowledge distillation, I reduced the deployment and inference cost to $0.002 per interaction. This marked reduction positions the approach as a viable option in resource-limited settings.

Method	Model Size	PubMedQA(ID)
Human (expert)	-	78.0
ChatGPT	175B	63.9
LLaMA-2	13B	68.0
Ayurllama	13B	57.2

Table 1: Comparative performance of AyurLLM model.

Conclusion

In short, this paper underscores the significance of LHWs in developing nations and explores how LLMs can aid them. It discusses language challenges, the need for guardrails, and cost-effective deployment.

References

Chen, S.; Ju, Z.; Dong, X.; Fang, H.; Wang, S.; Yang, Y.; Zeng, J.; Zhang, R.; Zhang, R.; Zhou, M.; Zhu, P.; and Xie, P. 2020. MedDialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*.

Deng, G.; Liu, Y.; Li, Y.; Wang, K.; Zhang, Y.; Li, Z.; Wang, H.; Zhang, T.; and Liu, Y. 2023. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *ArXiv*, abs/2307.08715.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. *ArXiv*, abs/2202.03286.

Saunders, A. 2023. Unlocking the Power of Enterprise-Ready LLMs with NVIDIA NeMo. *NVIDIA Technical Blog*.