Research Focus Summary *

A strong submission will:

- Demonstrate a clear understanding of the research challenge and its aims for societal impact.
- Describe relevant work/accomplishments, demonstrate the alignment of one's motivation to the research challenge, and outline the potential impact achieved through collaboration.
- Be clearly written, in alignment with program guidelines.
- Consider that readability and clarity are valued in the review process.

Statement of Interest guidelines

- Describe the motivation, or "why", behind a candidate's interest in research exploration in this research challenge. We are specifically looking for candidates to describe how their research goals and aspirations align with this challenge.
- Description of relevant qualifications, expertise, and perspectives the candidate would bring to the research challenge and how this experience would impact the outcomes of this collaboration.
- Detail any relevant areas that you would like to collaborate on with the Microsoft researchers and applied scientists listed as involved in the research challenge and how collaboration could create impact in those areas.
- Where applicable, include details of any relevant, preliminary research or work that shows progress and investment in this space.

Reducing the Digital Divide of Generative AI in the Global South

Research Focus Summary *

This proposal combines research and practice to define, measure, and mitigate the inequities for information access arising from Gen AI-infused tools that are deployed in the (so-called) Global South. The research focus is on evaluating systems such as retrieval augmented generation (RAG) tools, often pitched as "grounded" AI in consumer-facing products, in order to measure their performance on low-resource Indic languages that directly affects end-users. The goal is to provide a shared language to inform the reliable development of Gen AI within products meant for non-English speaking audiences, specifically healthcare-focused RAG chatbots deployed by nonprofits in low and middle-income countries.

The Statement of Interest is an opportunity to describe your motivations, qualifications, and relevant interests in this research challenge

Minimum 10-point font with a maximum of 3 pages, margins should be 1” or wider to ensure readability

Title: Evaluating Multilingual Retrieval Augmented Generation Products for Healthcare

The advent of generative AI technologies has empowered non-technical audiences to engage with technology more effectively. It has reduced the barriers to accomplish complex tasks involving advanced information retrieval and multimodal content generation. There have been significant communication and accessibility advances made possible by the deployment of such systems in multilingual environments, and we are actively witnessing the burgeoning growth in the ‘AI creator ecosystem’, with the advent of consumer-facing AI-infused products, especially large language models (LLMs). This comes with its own

challenges, a root cause of which is the lack of effective safeguards around undesirable outcomes including the generation of explicit, inaccurate, insensitive, or otherwise harmful content that unwitting end-users are directly exposed to.

Safety measures to mitigate these outcomes include model alignment efforts such as reinforcement learning on human feedback, model steering and safety-specific control tokens, fine-tuning and pre-training dataset pruning, post-hoc restrictions, synthetic fingerprinting, and other ad-hoc solutions. However, there is a vacuum in terms of comprehensive multilingual safeguards or general guidelines to deliver safe and trustworthy AI systems to end-users. The issue is worse in the case of low-resource languages because of:

1. Well-documented historical issues related to bias and fairness AI models intended for Global North audiences when they are 'retrofitted' and deployed to serve the Global South; an extension of out-of-distribution problems in addition to a lack of adequate evaluation of these models before releasing them.
2. The resources dedicated to the deployment of AI tools in the Global South often exceed the resources available for appropriate monitoring, measurement, and mitigation of harms in these environments resulting in lack of awareness of inadvertent harms that affect end-users in the first place.
3. Lack of sociocultural awareness around the usage of systems by end-users and missing context around their interpretation of the generated outcome(s).

While many collaborative research efforts provide multitask benchmarks on diverse datasets, and policy efforts are attempting to create a national 'AI governance' framework, we have a ways to go before we arrive at a shared language to measure harms across different types of AI systems in an effort to provide a globally-applicable solution to such challenges. In a similar vein, top-down regulatory efforts for social media transparency have demonstrated that it is impossible to provide sufficient coverage for all manner of harms in the first iteration of such multi-party efforts. Instead, I suggest adopting a bottom-up approach starting with the evaluation of an awareness-focused healthcare chatbot available to deploy in a regional setting for non-English speaking audiences. Following this, the scope of this research can be expanded to adjacent topics in healthcare including two different languages and audiences from different regions. Finally, we can use a deployed system as a testbed for our research with nonprofit external partners who can provide anonymized user experience data for supporting this research.

Key Contributions:

1. Provide a qualitative framework encapsulating the pitfalls for 'retrieval augmented generation' (RAG) systems deployed in multilingual environments. This will examine the choices of embedding architecture, knowledge graphs, and other information retrieval mechanisms for RAG tasks.
2. Provide quantitative evidence to support the qualitative framework through an evaluation of popular proprietary and open-source LLMs on a set of tasks related to a specific domain of application and in a specific context e.g. healthcare: menstrual health and hygiene.

3. Create an automated suite[1] to evaluate different types of RAG systems for healthcare, specific to the language considered, and confirm that this results in strictly better performance of the RAG chatbots deployed to pilot users.
4. Expand the set of domains and applications to other areas of healthcare including primary care, mental health, sexual health and wellness (or beyond healthcare) to evaluate the validity of the recommendations and incorporate updates to bridge the generative AI divide.

Methodology:
We will define, monitor, and measure the challenges arising from deploying RAG chatbots in multilingual contexts. This is because the method is widely being used (or at least searched for)[2] in 'grounded' AI systems and (maybe more concretely) deployed in many consumer-facing chatbots already, accelerated by the availability of developer ecosystems like OpenAI's GPTs[3], Amazon's Sagemaker, and various enterprise products with hundreds of millions of end users placed globally. We will start by generating a manual test suite of prompts that can be used to reliably evaluate the chatbot for a specific, multilingual use-case, informed by data on popular questions[4] available from prior research into the specific application. In this case, we evaluate the performance of an RAG system that answers popular questions about menstrual hygiene management (MHM) raised by Bengali females who suffer from 'period poverty'[5], a lack of access to such information. The questions will be informed by conversations with a nonprofit actively working in this area with various local communities, WaterAid Bangladesh. Thereafter, we will quantitatively evaluate the performance of various prompt-tuned RAG systems using popular proprietary and open-source chatbots using conventional metrics on information retrieval as well as human evaluation by analysts that speak the language provided by partnerships with the relevant nonprofit. We will intervene on the dataset quality, prompting strategy, and fine-tune relevant open-source models to evaluate post-hoc performance on the prompts. Lastly, we will evaluate the user experience comparing results from an early pilot with the updated RAG system to identify if our suggested strategies were able to adequately improve the user experience, and boost corresponding utilization of the product.

Evaluation in a Real-world Context: The Sakhi chatbot[6] is a WhatsApp chatbot for disseminating menstrual health and hygiene information including retrieval augmented generation and prompt-tuned Bengali responses provided to end users, based on a custom dataset of menstrual health and hygiene documentation that was compiled from public documentation including the UK-based charity WaterAid's[7] capacity building efforts in Bangladesh, WHO, and UNICEF reports on menstrual health management for women and adolescent girls. Sakhi, which means 'female friend' is a passion project built by a research collective I founded, called SimPPL[8], with the mission to build free, open-access trust and safety tools for nonprofits to amplify their impact using technology-assisted interventions. In fact, this project was built and scaled by a team of Indian and Bangladeshi students and has led to us winning a small grant from the German govt. towards developing and deploying the system.

---

[1] Zhou et al., 2023

[2] Google Trends data comparing the search popularity for variants of 'chatbot', 'RAG', 'grounded'

[3] There are already early complaints of privacy violations made possible by reverse engineering / adversarial prompting / 'jailbreaking' of OpenAI's GPTs, which certainly does not inspire confidence for these tools to be propagated in the Global South.

[4] Mehjabeen et al., 2022

[5] Period Poverty in Bangladesh - The Borgen Project

[6] https://sakhi.simppl.org

[7] https://www.wateraid.org/uk/where-we-work/bangladesh

[8] https://simppl.org

Potential Impact: We have the opportunity to deploy Sakhi for local communities in Bangladesh in partnership with WaterAid, which provides a valuable opportunity to collect real-world data, measure user ratings, complaints, and feature requests to improve the Gen AI-infused tool for them. But even beyond this, the architectural design for the chatbot allows us to replicate and deploy the system in a varied set of use-cases. For example, we have been requested to implement this system for three registered Indian nonprofits (DigiSwasthya[9], Heal Station Foundation[10], Aadhar Bahuddeshiya Sanstha) that we have previously conversed with, focused on public health, mental health, and sexual health and wellness, in rural Maharashtra, Pune, and Uttar Pradesh, for Marathi and Hindi speakers. While no commitments have been made on our end given a lack of funding to scale this system, I am sharing this information to explain that there are broader opportunities available to run small studies in different sociocultural contexts if it benefits this research to measure the experience of external users of RAG systems in a secure and safe setting i.e. using only publicly available data compiled in collaboration with the nonprofits and to a limited set of consenting users, with appropriate safeguards and monitoring in place. As explained, this is strictly in a non-commercial context, with the open-access tool being made available to nonprofits interested in deploying it safely.

Background: As a postdoctoral associate jointly at Boston University and MIT, I research platform governance and free speech. I hold a Ph.D. from NYU's Center for Data Science, specializing in machine learning, causal inference, and their applications in social media and politics at the Center for Social Media and Politics. In 2021, I founded and lead SimPPL, a research collective focused on creating civic integrity tools for media development organizations and journalists. We have won awards, fellowships, and grants from Google, Mozilla, Amazon, the Wikimedia Foundation., the Goethe Institute, the Anti-Defamation League, the NYC Media Lab, and others. I am passionate about empowering researchers from the global south to participate in mitigating online harms and building responsible AI tools for global audiences. I have previously worked on machine learning products and research at Slack, Adobe, Twitter, Oxford, CERN, and various startups catering to Fortune 50 clients in the domains of artificial intelligence and cybersecurity.

I run SimPPL, a research collective that is fiscally sponsored by the One Fact Foundation, a 501(c)(3) nonprofit in the US. For the past several years, I've mentored and trained students (pro bono) from non-premier institutes and underserved communities, paying them to develop scalable trust and safety tools to protect the public from online harms, initially at Unicode[11], then the NYU AI School. I formally started SimPPL as a volunteer-driven student organization in 2021 and we have since built technology products like Parrot and Sakhi analyzing hundreds of millions of online posts for The Sunday Times in the UK, Deutsche Welle in Germany, Tech Global Institute, and WaterAid in Bangladesh. Our work has been featured by Google[12] (separate blog post coming out soon), Deutsche Welle, in local and national news outlets in India and we recently launched a fellowship program to allow even more students from Tier-II and Tier-III institutes in India and abroad to contribute to our goal. We're empowering communities in the global south to participate in integrity work and contribute to tech and policy discourse and I hope that I receive an opportunity to do so with MSR.

---

[9] https://digiswasthya.org
[10] https://www.facebook.com/healstationfoundation
[11] https://djunicode.in
[12] Google Cloud Research Innovators program has 37 new participants