**Motivation:** The dynamic information ecosystem around elections provides a digital landscape that is ripe for information interference globally. With over 80 national elections in 2024, monitoring the integrity of the online information ecosystem presents a significant challenge for Trust and Safety (T&S) researchers and social platforms, especially across diverse linguistic and cultural landscapes. Initiatives run by Meedan (Check) and Times of India used public-facing tiplines where users could submit suspicious messages and posts, and access verified and accurate information about elections. But tiplines continue to rely heavily on human in the loop mechanisms that limit their scalability to large audiences without incurring significant material cost. We used LLMs to build a reliable, community-led solution to address a similar problem with public health information in Bangladesh and create a scalable, efficient, multilingual solution. We are applying to this grant to adapt, deploy, and demonstrate the viability of such systems for combatting election information interference. At SimPPL, are a nonprofit research collective of global majority students and aim to contribute a viable solution that India, and other regions may deploy in the year of elections. Our AI-based tipline in Hindi, English, and Marathi will employ large language models (LLMs) to disseminate accurate multilingual information and fact-checking tools during critical election periods, grounded in a public, governable database of fact-checks. Our AI-based tipline, developed in collaboration with trusted fact-checking nonprofits and local students, is personalized to the unique social norms and languages of each community and their users. We append preventative measures against adversarial use that will guarantee the tool's effectiveness and integrity in advancing Trust and Safety research.

**Problem Area**: Tiplines serve as a mechanism to crowdsource emergent deceptive and misleading narratives on popular channels, directing platform interventions and fact-checking to priority problem areas. Social media platforms and AI tools cater to an international audience, and often struggle with sociocultural nuances that complicates guardrail creation, content moderation, and the sourcing of harms, especially in hotly contested elections. We propose a community-based participatory research prototype allowing platforms and researchers to stay informed of local priorities and operationalize siloed third-party fact-checking efforts. Considering potential abuse of such systems, we implement additional filtering based on machine learning clustering and linguistic analyses to identify priority incidents.

**Proposed Solution:** We build a community-run LLM tipline to counter information interference. This AI-powered tipline will augment human fact-checking efforts, prioritizing actions based on user-submitted queries and tips. Initially focused on Mumbai and Jalgaon, a village in Maharashtra, the tipline will use facts verified by trusted local organizations to respond to local user inquiries, helping fact-checkers target prevalent misinformation narratives effectively. To do this, we aggregate fact-checks into a multilingual database, harnessing generative AI to ground our chatbot. We streamline the fact-checking process to empower voters, and democratize T&S research by creating

community-centric approaches.The transparent tracking of tips and public commentary will serve as a prototype for scaling and operationalizing multi-regional fact-checking endeavors. The governable database produces a first-of-its-kind pan-India collection of multilingual suspicious narratives.

**Research Methodology:** We deploy a WhatsApp Business phone number that users can text to share suspicious information for verification where possible, and receiving context where necessary. The users can have a multilingual conversation with an LLM that is prompt-tuned to respond in Hindi, Marathi, and English, including code-switching in conversations. They can share links to external material that will be processed and analyzed to find matches to an existing database of fact-checks that we have aggregated from both past academic datasets and real-time fact checkers listed on the Wikipedia List of Fact Checkers, under 'India'. If matches are not found, we will adopt an approach similar to News Detective to permit verified public comments, and crowdsource potential fact-checks including those from fact-checkers. For generating answers to user-submitted queries, we use retrieval-augmented generation (RAG) based multilingual LLMs. We create a database of fact-checked documents and vectorize them to form a vector store, which is a database of document embeddings. When we receive a user query, we identify the most similar documents to the query from the vector store using a similarity search metric. These selected documents are then fed to the LLM to ensure that the response is grounded in fact-checked documents, thereby minimizing hallucinations.

**Impact and Relevance to SIO:** Our research is the first to assess the effectiveness of AI-augmented tiplines in mitigating misinformation, understand the differential impacts of misinformation across regional settings, and identify topical variances in misleading narratives submitted by users to the tipline. We will run a field experiment to test the efficacy of tiplines at 5 colleges across 250 students, and in collaboration with a rural nonprofit in Maharashtra operating in 100 villages, recruiting family members to submit information they find suspicious to the tipline, for a period of up to 120 days. We capitalize on our lived experience, relationships, and college network of over 5000 students in the region, including SVKM, to enable us to deploy this tool at scale. By leveraging our existing Sakhi bot architecture, we demonstrate the replicability and potential impact of the proposed solution within reasonable cost margins. We have submitted an IRB for running the Sakhi menstrual health and hygiene chatbot field experiment. Our team includes Data scientists, T&S professionals and NLP researchers with multiple publications at EMNLP, ACL, NAACL, CIKM, ICWSM, ICML, NeurIPS, and those that led AI4Bharat instruction tuning regional language models (Raghav Jain, Jay Gala, Himanshu Beniwal, Caitlyn Vergara, Dhara Mungra).