**First Name (must be your legal first name)**

Swapneel

**Last Name (must be your legal last name)**

Mehta

**Email**

swapneel@onefact.org

**Phone Number**

5513287074

**Your Location (City)**

Cambridge

**Your Location (Country)**

United States

**Registered Name of Organization**

One Fact Foundation. 501 (c)(3) nonprofit in the USA

**Location of Organization’s Principal Office (City)**

New York

**Location of Organization’s Principal Office (Country)**

USA

**List of additional team members and partner orgs**

Raghav Jain
Kevin Aslett

**Is the individual applicant/ organization applicant a (i) government entity, (ii) political entity, or (iii) government official? If yes, please provide details**

No

**Proposal**

**Proposal Title**

Operationalizing Claims Verification through WhatsApp Conversational Agents

**Proposal Track**

Open

1. Describe the challenge being addressed and why this should be prioritized
   Note: 500 word limit

**Final: We present a community-focused claims verification system to deliver accurate information to local users and source suspicious claims, providing a novel extension to the concept of conventional "tiplines" with Llama 2. There are over 40 national elections coming up in 2024, presenting a complex challenge for social platforms to monitor the integrity of the online information ecosystem in each country, across languages and data modalities. This problem is exacerbated by the widespread availability of generative artificial intelligence to threat actors influencing electoral outcomes [1]. There are many local organizations including fact-checkers combating these harms through active monitoring. Even platforms frequently employ 'misinformation labels' based on verified fact-checks since there is evidence of their impact on user belief in falsehoods [2]. However, as research has shown, the distribution of accurate information often lags the spread and influence of fake news making it challenging to "operationalize" fact-checks**

**effectively. We address this imbalance by empowering NGOs with LLM-powered tiplines to conduct claims verification in real-time, for a large volume of existing debunked claims sent to them by users. Our system uses Llama 2 to perform retrieval augmented generation (RAG) over local databases of their local knowledge as documents structured in a question-answer format for accurate retrieval. Information relevant to user-submitted claims present in the database will be retrieved by the LLM to respond to users with context. For unknown information, these tiplines collect suspicious claims from groups of local users to prioritize organizational efforts on countering the most prevalent misinformation narratives. Furthermore, these documents are designed to be open-access to enable effective local stewardship and data governance akin to a "local Wikipedia". This decentralized model for tiplines run by locally verified orgs. is what we hope will serve as a global infrastructure to support users against threat actors that manipulate online discourse. Tiplines have been shown to be effective ways in creating national communication highways even in large, complex democracies [3] and WhatsApp has previously released tiplines as well. We believe it is essential to democratize such systems at least to IFCN member orgs. and other verified entities invested in a healthier internet and election integrity. We have already built a different WhatsApp chatbot called "Sakhi" (https://sakhi.simppl.org) for menstrual health and hygiene based on the same system architecture 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 WaterAid's capacity building efforts, WHO and UNICEF documents on menstrual health management for women and adolescent girls, by experts in. We are deploying Sakhi with local communities in Bangladesh. We also ensure our architecture is language-agnostic and have secured letters of intent and set up partnerships to deploy such claims verification systems with Fundamedios, Chequeado Bolivia, Chambal Media, Heal Station Foundation, DigiSwasthya, and potentially Jagran New Media to operationalize their local knowledge and collaborate on tiplines for civic and public health issues to augment their efforts.**

**[1] https://www.ajpor.org/article/12985-does-fake-news-matter-to-election-outcomes-the-case-study-of-taiwan-s-2018-local-elections**
**[2] https://www.pnas.org/doi/full/10.1073/pnas.2104235118**
**[3] https://misinforeview.hks.harvard.edu/article/research-note-tiplines-to-uncover-misinformation-on-encrypted-platforms-a-case-study-of-the-2019-indian-general-election-on-whatsapp/**

We present a community-focused claims verification system to deliver accurate information to local users and source suspicious claims, providing a novel extension to the concept of conventional "tiplines" with Llama 2. There are over 40 national elections coming up in 2024, presenting a complex challenge for social platforms to monitor the integrity of the online information ecosystem in each country, across languages and data modalities. This problem is exacerbated by the widespread availability of generative artificial intelligence to threat actors

who can cause widespread online harm, especially influencing electoral outcomes [1]. There are many local organizations including fact-checkers combating these harms through awareness efforts, elections monitors, dashboards. In fact, platforms frequently employ ‘misinformation labels’ based on verified fact-checks since there is evidence of their impact on user belief in falsehoods [2]. However, as research has shown, the distribution of accurate information often lags the spread and influence of fake news making it challenging to “operationalize” fact-checks and deploy them effectively. Our project addresses this imbalance by empowering NGOs with LLM-powered tiplines to conduct claims verification in real-time, for a large volume of existing debunked claims sent to them by users. We do this by employing Llama 2 to perform retrieval augmented generation (RAG) over local databases of their internal knowledge encoded into documents. This process assumes that there is information relevant to user-submitted claims already present in the database, structured in a question-answer format for accurate retrieval. If the information is not present in the database, these tiplines serve as a dual channel to solicit suspicious claims from large groups of local users to prioritize their efforts on countering the most prevalent misinformation narratives. Furthermore, the databases are designed to be open-access to enable effective local stewardship and data governance akin to a “local Wikipedia”. This decentralized model for tiplines run by locally verified orgs. is what we hope will serve as a global infrastructure to support users against threat actors that manipulate online discourse. Tiplines have been shown to be effective ways in creating national communication highways even in large, complex democracies like the Indian election in 2019 [3]. In fact, WhatsApp has deployed its own tipline called ‘Checkpoint’ previously. Advancing this concept with our novel application, we automate responses to conversational queries leading to database entries for emergent misinformation and the detrimental effects of disinformation despite platform safeguards, we believe it is essential to democratize the deployment of tiplines at least to IFCN member orgs. and other verified entities invested in a healthier internet and election integrity. We have already built a different WhatsApp chatbot called “Sakhi” (https://sakhi.simppl.org) for menstrual health and hygiene based on the same system architecture 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 WaterAid’s capacity building efforts, WHO and UNICEF documents on menstrual health management for women and adolescent girls, by experts in. We are deploying Sakhi with local communities in Bangladesh supported by a local chapter of WaterAid, the UK-based charity. We also ensure our architecture is language-agnostic and have secured letters of intent and set up partnerships to deploy such claims verification systems with Fundamedios, Chequeado Bolivia, Chambal Media, Heal Station Foundation, DigiSwasthya, and potentially Jagran New Media to operationalize their local knowledge and collaborate on tiplines for civic and public health issues to augment their efforts.

[1] https://www.ajpor.org/article/12985-does-fake-news-matter-to-election-outcomes-the-case-study-of-taiwan-s-2018-local-elections
[2] https://www.pnas.org/doi/full/10.1073/pnas.2104235118
[3] https://misinforeview.hks.harvard.edu/article/research-note-tiplines-to-uncover-misinformation-on-encrypted-platforms-a-case-study-of-the-2019-indian-general-election-on-whatsapp/

2. Describe proposed solution, including how you will leverage Llama 2 to implement that solution (what to mention)

We will create a claims verification pipeline for fact-checkers, nonprofits, and civil society orgs. that operate in the areas of civic and public health to operationalize their local knowledge and fact checks through a WhatsApp-based Llama 2 chatbot deployed to perform retrieval augmented generation (RAG). Our system will significantly amplify their existing digital literacy and awareness efforts at a cheaper cost, and dramatically increase their audience by employing the popular messaging platform, WhatsApp. We have developed an RAG pipeline over a pre-collected and indexed set of documents in a vector database, that will be retrieved if a claim matches existing fact checks, or entered into a frequency-based priority queue if not. Compared to other open source large language models (LLMs), Llama has proven to be a more accurate tool for fact checking. While not directly employing an LLM to verify a claim without additional information, we aim to leverage this behavior in our pipeline. We will work with local partners to create a chatbot which is integrated in Whatsapp to allow users a "low-cost high-reward" mechanism to verify claims they come across based on local knowledge bases, including but not limited to fact-checking the content they come across in times of civic volatility and elections. Any user in the local area can submit their claim for verification to the Whatsapp bot. If there is already a database entry related to the claim, the RAG pipeline will pull it up and converse about the evidence for the claim to be labeled as true or false. If not present in the database, it will respond to the user accordingly and enter the provided claim into a priority queue based on the frequency of users raising this claim in the tipline. Eventually, we aim to allow multimodal claims starting with images and screenshots, which we will cluster based on a combination of text and object detection, locality sensitive hashing, matching against external databases, and other ways to serve as a prioritization signal for the fact-checkers. The claims verification process itself will remain human-in-the-loop, relying on multiple signals including the frequency of incoming posts, which allows local orgs. to engage in anonymized information sharing as well as prioritization of incoming tips. We also include optimizations and caching of popular queries to improve the efficiency of our response generation. The information we compile in the vector database of documents will be in collaboration with the nonprofits we work with, and include their knowledge of local issues, verified news articles from national and international press, documentation released by relevant authoritative entities like UN agencies (WHO, UNICEF in the case of public health education, like Sakhi), publicly available fact-checks from global IFCN members, and adapted documentation from civil society orgs. such as the Integrity Institute (we co-authored their elections integrity guide https://integrityinstitute.org/elections-integrity-program and transparency guide https://integrityinstitute.org/news/institute-news/integrity-institute-releases-overview-of-online-social-platform-transparency) and others that support capacity-building efforts in global elections. This public-benefit technology is an important first step to address complex challenges around mitigating mis and disinformation, but we have the technical experience, partnerships across the globe, and an organizational mission to develop this technology to support local development efforts for elections integrity and public health.

3. Describe the impact of your solution and its capacity for scale over time.

There are many important problems that are only possible to solve with local context and a key challenge that local fact-checking, civil society, and public health organizations face is the lack of viable mechanisms to scale their work to reach a larger population as well as the technology required to do so. Our solution not only helps mitigate the spread of civic and public health misinformation by amplifying the work of established and verified integrity-focused orgs. but also provides a means to scale their information-gathering efforts through a consumer-facing familiar chat interface. We are employing cloud services such as GCP and AWS for creating scalable solutions globally like https://parrot.report that we designed to identify Russian threat actors on Twitter for the Sunday Times, and https://sakhi.simppl.org that we built as a digital literacy chatbot disseminating menstrual health information for WaterAid in Bangladesh, to serve hundreds of their users in a pilot project. Our solution leverages community-based research design involving local organizations as arbiters of the knowledge bases that the bots are permitted to chat over. We have already had an incredible impact, reaching thousands of visitors for https://parrot.report and raising small grants and fellowships to the tune of USD 140,000 from Google, Wikimedia, Mozilla, Amazon, the Anti-Defamation League, the NYC Media Lab, Craig Newmark, the Knight Foundation, Deutsche Welle, and the Tech Global Institute. We have grown our team from 4 members to 12 members in the past two quarters and operate based on agile project management methodology with weekly sprints, team-based contributions, and a portion of our time dedicated to research in the fields of ML, NLP, and online trust and safety.

4. Describe how your project will exhibit responsible practices with regard to data security, privacy, replicability etc.

We have put together a comprehensive data management plan based on our past work building and deploying consumer-facing technologies in order to offer a safe and secure system with the ability to provide direct feedback and eliminate harms at the source. This includes a structured release of the tool to broader audiences after limited pilots in the local community in order to test guardrail efficacy and response accuracy. We work with nonprofits to amplify their ongoing work in order to approach a technological intervention as a partnership. Additionally, we have confirmed that former OpenAI Trust and Safety lead Dave Willner will devote time to advise our project on potential harms, abuses, and red-teaming of the proposed system.

https://docs.google.com/document/d/1nVN97g5mYLxXj7iyVOYr6xOsMwWbTp4GIfJIDrzd1ns/edit?usp=sharing

5. Describe why your team is best equipped to implement your proposed solution (highlight relevant expertise) (what to write here)

Mozilla Responsible Computing Challenge India

We are a research collective called SimPPL, fiscally sponsored by a US-based 501(c)(3) nonprofit, the One Fact Foundation, developing open-access trust and safety tools. Our expertise spans prior systems like https://parrot.report designed to analyze 80 million tweets and 14 million accounts with optimized pairwise computation to identify coordinated networks of Russian threat actors with high precision, using statistical machine learning and artificial intelligence. We also built https://sakhi.simppl.org as a production-ready WhatsApp chatbot, scaling model inference pipelines and exploring fine-tuning on multilingual datasets through QLoRA in order to improve RAG performance for queries in Bengali and Beng-lish (code-switched queries) for popular user queries. Our core team comprises members who are full-time data scientists working with cloud infrastructure, ML engineers, production engineers, political scientists, professors, and social science researchers. We have previously successfully built multiple student organizations that are training hundreds of students each year called DJ Unicode (https://djunicode.in) and the NYU AI School (https://nyu-mll.github.io/nyu-ai-school-2023/). Our team comprises 20 members, with the team leads as follows:

Swapneel Mehta, Founder of SimPPL and Postdoc at BU and MIT

I'm a postdoctoral associate jointly at Boston University and MIT, where 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 CSMAP. 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 started SimPPL as a volunteer-driven student organization in 2021, and have since built technology products 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 raising USD 150,000 in grants and fellowships alongside my Ph.D. from 2019-23.

Raghav Jain, Research Lead

Raghav is an incoming research associate at the National Centre for Text Mining at the University of Manchester, where he works under Dr. Sophia Ananiadou exploring the scientific document understanding capabilities of large language models (LLMs). He is also currently collaborating with the Oxford TVG Lab (Philip Torr) and Meta researchers (Ser Nam Lim) on identifying and mitigating the potential for influence operations on social media using LLMs. His past research experience includes working with the joint NLP Lab of IIT Patna and IIT Bombay under the guidance of Dr. Sriparna Saha and Dr. Pushpak Bhattacharyya. There he worked on a variety of NLP problems and collaborated with researchers from Microsoft, the University of Tokyo, and Amazon Alexa. In broader perspective, his research concentrates on online safety, AI integrity, and user analytics. This encompasses projects such as developing advanced AI-powered content moderation techniques, testing the boundaries of LLMs, and implementing AI across sectors including education, law, and healthcare. His work has led to numerous publications in top-tier venues including EMNLP, ACL, CIKM, ECML, ACM MM, ECAI, ECIR, and IJCNN.

Prof. Kevin Aslett, Univ. of Central Florida

I am an Assistant Professor at the University of Central Florida in the School of Politics, Security, and International Affairs and the Cybersecurity and Privacy Research Cluster. I am also a Faculty Affiliate at the Center for Social Media and Politics at New York University and a Fellow at the Political Economy Forum at the University of Washington. My research broadly focuses on the threats digital technology pose to liberal democracy (specifically online misinformation) and policies designed to mitigate these threats. In addition, I have developed new methods in the field of political communication that remove obstacles to extracting information from enormous collections of electronic text and images that users encounter online. You can track my code and find replication files on Github. I have published nine articles in peer-reviewed journals such as *Science Advances*, *Journal of Experimental Political Science*, and *Policy Studies Journal* and in popular outlets such as the Washington Post. I also serve on the Editorial Board of the Journal of Online Trust and Safety run by the Stanford Internet Observatory.

https://simppl.org/team

6. Outline resources needed for implementation (e.g., budget; additional partners; assistance with tech)

We would appreciate assistance with discussions on efficient ways for setting up fine-tuning infrastructure if and when necessary. Our team members have fine tuned Llama 7B with QLoRA adapters on smaller datasets earlier but would appreciate infrastructure advice.

We would also appreciate help walking through the research design, guardrails, and trust and safety consultancy from experts at Meta that could point out potential harms and unforeseen abuses of our system by malicious end users.
Our partners span multiple areas of operation advancing civic integrity and public health, and we are working largely on English chatbots for the purpose of this project. We would appreciate further advice on prioritizing projects in regions where Meta has historically observed outsized impact from community-based research with local partners.

https://docs.google.com/spreadsheets/d/1i9XACNJun4ldbE3lkzp_e8B5CcFTBtU3krFoF_EJXXk/edit?usp=sharing

1. Project Lead
2. Research Engineers
3. Research Assistants
4. Cloud Costs
    a. Web Servers
    b. Storage
    c. NLP Pipelines and Data Infrastructure
    d. Trust and Safety Operations - alerting infrastructure
5. Human Coders for Multilingual Tagging
6. NGO Partnerships for Community-based Research

## Anticipated timeline for implementation

15 months (including 4 months of performance monitoring and testing with partners)

Weeks 1 - 6:
Project Kickoff and Planning
- Define specific goals, roles, and responsibilities
- Set up communication channels within the team

Data Collection and Preprocessing
- Begin collecting fact-checked documents from Iffy News and other sources
- Implement data preprocessing pipelines for efficient storage and retrieval

Initial Model Training
- Instruction-tune the Llama 2 model on the preprocessed fact-check dataset

Week 5-12:
Development of Claims Verification Pipeline
- Integrate the trained Llama 2 model into the fact-checking pipeline
- Develop retrieval-augmented generation mechanisms

Website and Chatbot Development
- Begin development of the website for the claims verification platform
- Initiate the chatbot development for WhatsApp integration

Prototype Testing and Iteration
- Conduct initial testing of the claims verification pipeline and chatbot
- Gather feedback from internal testing and make necessary iterations

Week 13-18:

Full-Scale Model Training
- Fine-tune the Llama 2 model based on initial testing results
- Scale up training for improved performance

Website and Chatbot Refinement
- Refine the website and chatbot based on user feedback
- Implement additional features and improvements

External Beta Testing
- Roll out the claims verification platform for external beta testing
- Collect feedback from users for further refinement

Weeks 20-25:

Integration with WhatsApp and Database
- Integrate the chatbot with WhatsApp for broader accessibility
- Ensure seamless connectivity with the fact-check database

Security and Privacy Audit
- Conduct a thorough security and privacy audit of the entire system
- Implement any necessary enhancements to ensure compliance and user data protection

Tipline Implementation
- Develop the tipline functionality for users to submit new claims
- Implement mechanisms for sourcing prevalent misinformation

User Education and Outreach
- Create educational materials on how to use the platform
- Initiate outreach efforts to promote the platform and encourage user engagement

Weeks 26-30:

Capability Testing for Independent Fact-Checking
- Explore the feasibility of the chatbot conducting its own fact-checks
- Determine the potential for expanding the pipeline's capabilities

Weeks 34-40:

Final Testing and Documentation
- Conduct comprehensive testing of the entire system
- Finalize documentation for the claims verification pipeline and chatbot

Weeks 40 - 60

Deployment with 2 partners in the global north (focused on elections integrity) and 1 partner in the global south