Your Team

List any other people working with you on this project, including names, locations, the role they will play in the project, and links to online profiles like github/linkedin.

1. Swapneel Mehta, Ph.D. Boston, Team Lead, https://www.linkedin.com/in/swapneelm/
2. Dhara Mungra, India, Engineering Manager, https://www.linkedin.com/in/dhara-mungra-0aa599126/
3. Raghav Jain, Research Engineer,
4. Uday Sharma, Sophomore, Engineering Intern

Are there any local AI projects that you are excited about (other than what you're working on)?

We are excited about Mozilla's OAT led by Deb Raji and in complement to auditing online algorithms we collect social data. For example, we are advancing a number of efforts to open up access to social media research to undergraduates from underserved communities in India whom we train to become leaders in responsible computing (see our deck here https://docs.google.com/presentation/d/1DsPkmfNudAgU7g_wIQxAnedJ17OEZRn7f5o119F7dJg/edit).

Origin Story

How was your project born? Where did it all begin?

Trust in institutions, media, and even science is at a low point across the world, desperately requiring mechanisms to enable people to identify online facts. However, "getting people to do their own research" only elevates entrenched misbeliefs by relying on biased sources. We aim to provide a balanced set of insights about information propagation that allows users to determine its likely authenticity. In 2022 we successfully built Parrot (https://parrot.report), a system to support the transparent investigation of coordinated networks of accounts on Twitter spreading manipulated information from state-backed Russian media providers by journalists from The Sunday Times. Our agenda was to step beyond reliance on conventional "third-party ratings" alone as a measure of malintent (e.g. the US "saying" Russian state-backed media spreads misinformation) and provide the public with the ability to independently draw evidence-backed conclusions of inauthentic online content. We scaled Parrot to analyze hundreds of millions of tweets, filter tens of millions of accounts and identify at a set of about 400,000 coordinated accounts that were successfully reported through our conversations with the Integrity Leads for X, formerly Twitter, in 2022. We wrote a report about actor analysis, a draft (https://jhagrutlalwani.vercel.app/blog/network_analysis_simppl) of which has been viewed by thousands of visitors, in anticipation of its final publication.

Parrot was based on the simple idea of scaling the detection of coordinated networks and narratives to identify the authenticity of a campaign, and could be applied to multiple other social platforms. It’s success led us to create tools that have collected 100M+ posts across Bluesky and Truth Social, 100M+ comments on YouTube, and nearly half a million instances of flagged content, resulting in 98 million interactions on Facebook. These efforts have informed teams at YouTube and Meta in performing platform functions, moderation of hateful content, and improvements in online safety driving tangible impact from our research. Parrot helped us cultivate the technical, multilingual, and multimodal understanding that informs its successor, a cross-platform network science tool called Arbiter. Arbiter is our attempt to provide public verification mechanisms for online content, making quantitative data on content spread accessible outside of platforms in order to allow for transparent arbitration of issues pertaining to content authenticity.

Project Summary

In a few sentences, describe your project and why it matters.

We will develop a user-friendly platform that allows individuals to obtain digestible reports of the nature of spread of their selected content across multiple platforms, made accessible through natural language querying to support all types of users. Think of it as an assistant that you can request to identify the spread of a link and it will mine multiple social platforms to obtain an assessment of the nature of accounts that promoted that link along with their historical reputation. Allowing laypeople to track how individual pieces of content spread across various social platforms and recognize that these are promoted by inauthentic accounts can provide direct evidence that reduces their belief and sharing of false and low-quality information. Social science and cognition research by our advisor(s) shows that providing this agency combined with a conversational interface is successful even in limiting beliefs in conspiracy theories, and we want to apply this logic to misinformation. The technical knowledge of the users can vary: a researcher can use Arbiter to see how videos containing political advertisements or disinformation on YouTube are shared across other platforms like Meta and Twitter. Tracking how information spreads across multiple platforms is critical given that users are exposed to multiple platform feeds and may be influenced by a variety of recurring content through heterogeneous channels–multiple exposure to misinformation is significantly more influential and Arbiter aims to combat that head on. As a stretch goal, we plan to deploy an AI agent that users can interact with to report any circulating information on social media platforms, check if the information has been fact-checked, and request verified information on various topics relying on publicly accessible databases of fact checkers from media houses (AP) and fact-checkers (IFCN). Arbiter is crucial to scale the efforts of local fact-checking organizations to detect misinformation and fact-check widely viewed social media posts. We have designed a prototype available at:

How your project relates to Local AI*

Describe the parts of your project that are local.

Arbiter is designed to help communities without technical knowledge and resources to collect and analyze millions of interactions on social media and engage in daily conversations in local languages about potentially misleading information they might come across. It serves as an input mechanism to source incoming claims while providing responses for debunked narratives spread online. It offers nuanced conversations and limits itself to a neutral stance allowing users to draw their own conclusions about content veracity.

The usability of Arbiter is twofold - for researchers, local fact-checkers, and civic society organizations to identify misleading narratives and for the general public to converse over daily information. Arbiter will analyze millions of pieces of content just as Parrot did, but it will go further in streamlining the fact-checking process for organizations without technical capacity. We expect it to significantly reduce time-to-action for stakeholders without technical expertise. We create a vector database of collected social media data and employ custom RAG (Retrieval-Augmented Generation) techniques to answer user queries. These features enable users to quickly query data by asking questions or prompting the system. By simply describing their needs through questions, users can obtain the data and results they require without the need to scrape data or perform the analysis themselves. Arbiter will additionally aim to provide access to an AI agent that users can interact with to make reports, request fact-checks and other verified information on a limited but growing set of pre-specified topics.

We aim to make Arbiter multi-modal in addition to being multilingual in order to eliminate linguistic barriers and ensure accessibility for low literacy segments of the global majority.

Objectives*

What questions or problems will your project answer or address?

Trust is crumbling on the social internet and this kind of public benefit infrastructure is critical to advance information integrity. Our vision is for Arbiter to be a user-friendly tool that will reduce the time and cost of analyzing content populating across platforms for a global audience of researchers, journalists, and other civic integrity professionals. It will be able to answer the question “is this content suspicious and what data indicates that?” Fraudulent news costs the global economy $78 billion a year, so the European Commission and the U.S. Department of Defense Advanced Research Projects Agency are investing millions in technical solutions to spot malicious and manipulated videos. As fraudulent content gets even more dangerous with new technology, we aim to position Arbiter as a leader in tracking harmful AI-generated content across social platforms through crowdsourced discovery and governance. There is a clear need for both, platforms and regulators, to prune harmful content to mitigate economic and societal harms at the global scale. Arbiter aims to bridge the knowledge gap across platforms to help professionals stop manipulative content from spreading virally.

Creativity*

Describe how your project is novel in its approach to on-device and local AI solutions. Think through how your project enhances privacy, speed, and/or accessibility, all within the framework of open source.

There are a number of platforms attempting to track cross-platform information including openmeasures.io, formerly SMAT, DataMinr (GenAI product), Alethea (Artemis), Palantir (Gotham), OverwatchData, and others. We used most of these systems or spoke to their users and identified that it was challenging for small-to-medium businesses to access and use these platforms due to a lack of multilingual focus, technical expertise, price point, and general knowledge. Our goal with building a natively multilingual platform that allows natural language interaction is to focus on the global majority much of which is bi and trilingual, and the need to provide cheap and accessible means to verify content because misinformation, taboos, and social stigma create an incredibly challenging ecosystem to drive behavioral change within. Arbiter is unique because it measures the networked spread and influence of key narratives, articles, actors, and popular public figures not only on a single platform but across a variety of mainstream, emergent, and alternative social media platforms and makes it available to consume at the pace of the user, in their preferred language and medium. This design aims to increase the accessibility of the system for end users like local nonprofits and grassroots orgs. invested in social change and able to source localized narratives. It also opens up cross-platform post-sharing data for researchers, local fact-checking organizations, and newsrooms. Instead of relying on traditional, resource-intensive methods to analyze content for harmful or toxic narratives, Arbiter examines the flow and dissemination of information. It starts with a focus on hashtags and media links that spread at an abnormal rate, and grows out to identify images and video reflecting manipulated narratives. This approach reduces computational time and resources, making the tool content-agnostic, tiered, and globally relevant for detection of misinformation and actors involved in its propagation.

Technical Implementation*

How will you technically implement this solution? Describe the technology stack you plan to use and the datasets you will leverage.

Data will be collected from social media platforms including YouTube, Bluesky, Instagram, Telegram, Sharechat, and Truth Social annually and without violating platform terms of service. A vector database will be created using ElasticSearch to store queried social media interaction data, facilitating faster RAG (Retrieval-Augmented Generation) data querying and retrieval. The project uses Django as its backend server and accesses data through endpoints for visualizations. This data is extracted from Parquet files stored on AWS S3 buckets using DuckDB queries or Apache Spark if we want to scale fast. The project frontend is built with Next.js and React, with visualizations created using Charts.js and D3.js libraries. Arbiter is deployed on the Vercel platform, from the creators of Next.js. We have built systems at scale using this tech stack and can attest to its flexibility to accommodate the proposed use cases.

Deliverables*

Expected results of the project and how you will measure success.

For the general public users of Arbiter, we will measure success by the number of reports we enable users to share related to the verification of content through arbiter–both short form and long form reports about content spread across online platforms. We further aim to publish long-form reports (like we have in the past) in collaboration with civil society partners focused on election integrity, combating online manipulations, and limiting the online abuse and harassment of politicians. These reports inform both the public and platforms about the multifaceted risks from coordinated campaigns and information operations. We plan to test Arbiter with our external collaborators, such as the Wikimedia Foundation, Mozilla, The Sunday Times, Deutsche Welle (DW), Tech Global Institute, and IFCN-verified fact-checkers, to help them identify the concerted spread of misleading content and facilitate their fact-checking workflows for such information to provide a human comparison of the system's performance. Our expectation is that this will reduce costs and the time it takes to analyze content populating across platforms. Separately, we intend to use Arbiter for newsrooms to educate journalists about how coordinated networks work to spread fake news and help them understand what type of content or articles are being circulated and potentially identify bad actors and fake news. The number of partners and published reports with partners will indicate overall success for this project on the civil society front.

Additional Project Links

Add any other relevant links that can give us a sense of your project. (Deck, demo, github, homepage, etc.)

https://arbiter-frontend.vercel.app/dashboard
https://parrot.report/

Other Ideas

If you had any other ideas you considered applying with, please describe them here. One may be something we've been waiting for.

Civic dialogue tool?

Required Funds*

If accepted into the accelerator, Mozilla will provide funding to support your project. How much funding does your project require?

a. $100,000

Use of Funds*

Describe how the funds will be used for your project and describe how you estimated these costs.

We will use 30,000 USD from the grant towards computational and infrastructure costs, including cloud infrastructure and technical equipment required to develop the Arbiter system. 50,000 USD will be used to support our staff (including salary, fringe benefits, tuition, or other direct compensation) to feature development to support professionals across journalism and civic society spaces. 10,000 USD will cover registration fees and travel expenses for conferences, seminars, and networking events that we attend to present our work while an additional 10,000 USD will cover miscellaneous expenses and indirect costs. These cost estimations are informed by actual numbers from the previous grants we have received via Google, Mozilla, Amazon, Wikimedia Foundation, the Center for Tech and Society, and the Goethe Institute.

Prior Funding*

0

Have you formed any legal entity?*

SimPPL

How did you hear about Mozilla Builders?*

LinkedIN