India OpenGenAI Data Charter Summary for Stakeholders

Author	Date	Version	Comment
Anuraj Ennai, Kalyan Vaidyanathan, Raj Jain, Ritesh Mathur, Trivikram Prasad	01-Apr-2024	1.0	Initial Version for publishing

We are delighted to present the draft of the OpenGenAI Datasets Charter, a foundational blueprint for fostering the development, refinement, and dissemination of diverse datasets pivotal for the Open Generative Artificial Intelligence (GenAI) ecosystem in India. This initiative is designed to harness the power of AI in reflecting the rich diversity of India's linguistic, cultural, and socio-economic diversity. Below is a concise overview of the charter's core components and action plan:

Introduction

Objective: Develop and facilitate diverse, high-quality datasets for an Open GenAI framework, encapsulating a wide range of Indian languages, India Specific Verticals, and modalities (audio, video, text) to drive AI innovation and inclusivity.

Foundational Tenets

1. **Comprehensive Inclusion:** Prioritize the representation of India’s linguistic and cultural diversity. Build and disseminate India specific data sets for verticals that have the greatest public impact and impact to government services.
2. **Data Integrity and Authentication:** Ensure the authenticity and reliability of datasets through rigorous verification.
3. **Privacy and Regulatory Compliance:** Maintain strict adherence to privacy laws and anonymization of personal data and/or corporate data where data is collected from industry.
4. **Open Access and Collaboration:** Foster an open, collaborative ecosystem for data sharing and innovation.
5. **Quality Assurance:** Commit to continuous dataset quality improvement and community feedback integration.

Action Plan

Outlines systematic approaches for data collection, verification, anonymization, and open distribution, alongside a model for community engagement and contribution.

Action Plan Asks

1. **Government:** Funding allocation, data collection standards, and provision of anonymized data.
2. **Private Sector:** Industry-specific dataset contributions, participation in challenges, and resource provision.
3. **Academia:** Multilingual dataset curation, competition hosting, and ethical data practice guidance.
4. **Non-profits:** Public domain data aggregation and advocacy for ethical, inclusive data practices.
5. **International Agencies:** Best practice sharing, cross-border data sharing facilitation, and funding for data inclusion efforts.

Roadmap

Phase 1: Formation of working groups, standard definition, initial dataset development, and architectural experimentation.

Recommendation on data collection and sanitisation to be useful across verticals.

Phase 2: Open source model releases, dataset expansion, and application validation in key sectors.

This charter aims to catalyse the advancement of an equitable and robust GenAI ecosystem in India, aligning with national interests and global standards. We look forward to your support, collaboration, and feedback to bring this vision to fruition.

Annexure A: Dataset Requirement

The Representative Data Set Requirements for the Open Gen AI Data Charter outline essential data needs across key sectors to support AI-driven solutions in India. Key highlights include:

1. **Healthcare:** Focus on diagnostics, patient data, and clinical trials, emphasizing anonymity and real-time data for pandemic control and health interventions.
2. **Finance:** Includes accounts and transactions data, stressing anonymity and recency for improved financial services and policy-making.
3. **Education:** Covers admissions and outcomes, aiming for curriculum design and targeted educational interventions with a focus on data anonymity and recency.
4. **Retail and E-Commerce:** Involves purchase and returns data, highlighting the need for anonymity and recency to optimize supply chains and demand forecasting.
5. **Manufacturing:** Emphasizes production and supply chain data, with a priority on recency for production planning and supply chain optimization.
6. **Transportation and Logistics:** Focuses on route and traffic data, underlining anonymity, recency, and real-time data for infrastructure and traffic management.
7. **Agriculture:** Involves crop and production data, highlighting anonymity and recency for crop planning and targeted interventions.
8. **Energy:** Covers production and consumption data, with a focus on anonymity, recency, and real-time data for energy planning and demand prediction.
9. **Media and Entertainment:** Includes content and viewership data, stressing anonymity and recency for content planning and pricing strategies.
10. **Government and Public Services:** Focuses on public service data, emphasizing anonymity and recency for infrastructure planning and service efficiency.
11. **Infrastructure Services:** Focuses on basic infrastructure data on water, sewer, roads, energy distribution, land use, and real estate for better planning and maintenance.

Overall, the requirements stress the importance of data anonymity, recency, and the need for real-time information to foster AI-driven planning, analysis, and policy-making across diverse sectors.