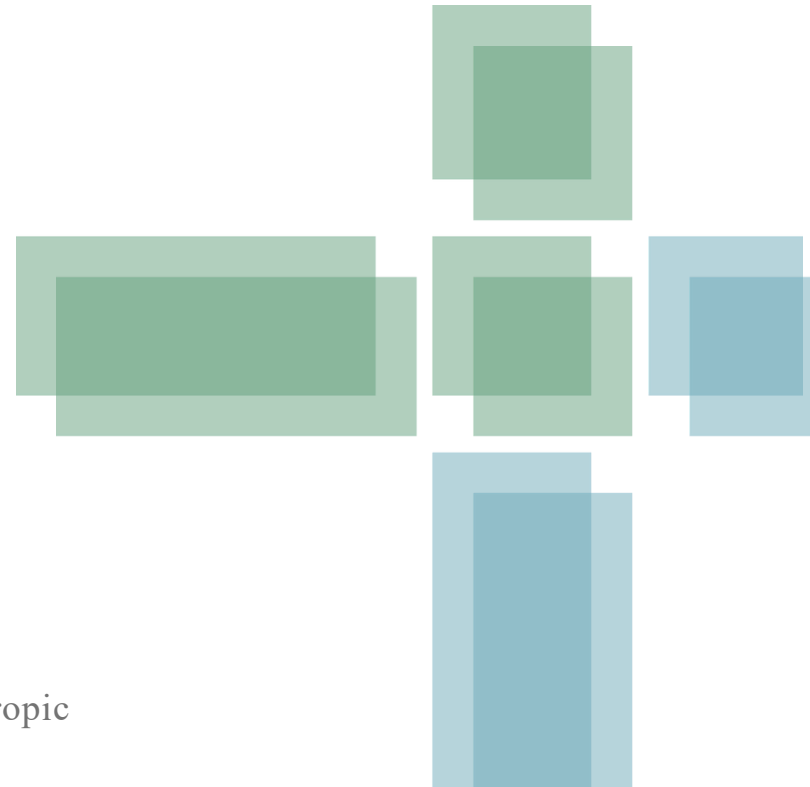
ANTHROP\C

people + ai
An EkStep Initiative

THE GUARDRAILS OF PROGRESS

Making AI Safe for the Real India

FEBRUARY 2026



By People+ai in collaboration with Anthropic

ABOUT PEOPLE+AI

People+ai, an initiative of the EkStep Foundation, that brings together do-ers, dreamers, tinkerers, and innovators to build a people-first AI ecosystem. It connects ideas, talent, and resources to reduce friction in population-scale AI adoption, enabling safe, inclusive, and equitable use cases.

Grounded in the belief that transformation must be people-centric, People+ai ensures that technology follows society's needs, not the other way around.

INTRODUCTION

Imagine a small farmer in Maharashtra. He owns two acres of rain-fed land and navigates the complexities of the mandi in a blend of Marathi and local Hindi dialects. For him, a digital system offering advice on seeds or credit isn't a Silicon Valley demo—it’s a life-altering tool. If that tool hallucinates a pesticide dosage or misquotes a government subsidy, the cost isn't just a "bug"; it’s a ruined harvest or a mountain of debt.

In India, AI safety is a practical necessity. As the nation moves toward integrating AI into its **Digital Public Infrastructure (DPI)**, the challenge is clear: How do we ensure that AI remains predictable, fair, and trustworthy for 1.4 billion people?

This is core to the collaboration between **EkStep Foundation** and **Anthropic**. By merging frontier AI research with the lived reality of Indian scale, they are building a "safety-first" architecture that doesn't just work in a lab, but thrives in the field.

BEYOND THE PILOT: THE CHALLENGE OF POPULATION SCALE

Most global AI safety benchmarks are built on Western assumptions—often tested in English and focused on abstract future risks. But in India, the risks are immediate and practical.

"Safety must be built into AI systems from the beginning, designed like road guardrails, rather than left to individual drivers to figure out."
— EkStep Foundation

When AI meets a farmer over a noisy phone line, it must do more than just answer a question. It must handle "pushback"— users who challenge advice or ask complex follow-up questions. To solve this, EkStep is drawing on its history of building national-scale infrastructure to treat safety as a **shared public protocol.**

THE CONSTITUTIONAL APPROACH TO SAFETY

At the heart of this collaboration is Anthropic's concept of **Constitutional AI**. Instead of just training a model on what not to say, Constitutional AI gives the model a set of guiding principles for Claude's values and behavior—a "constitution"—that it must follow. Anthropic's approach to Claude's constitution is for AI to understand values and knowledge necessary to behave in ways that are safe and beneficial across different circumstances. Most cases in which AI models are considered unsafe or insufficiently beneficial can be attributed to models that respond in overtly or subtly harmful ways, because they have limited knowledge about the world or the context in which they're being deployed, or that lack the understanding to translate good values into good actions.

To adapt this for India, EkStep is exploring contextual safety challenges, including:

- **Groundedness**: Every response must be anchored in verified sources like government gazettes or university research.
- **Stability under Challenge**: The system shouldn't "people-please" by changing its answer just because a user questions it.
- **Temporal Awareness**: AI must distinguish between current policies and those that have expired.
- **Jurisdiction Sensitivity**: Recognizing the difference between Central and State-specific regulations.

AGRICULTURE AND MSMEs: THE STRESS TESTS

The partnership is currently focused on two critical sectors where "safe AI" can have the highest impact:

1. Agriculture (OpenAgriNet)

By deploying **Claude** models within the OpenAgrinet, the team is testing how AI behaves when faced with local accents and regional dialects. The goal is to achieve "Domain Boundedness"—ensuring a system giving crop advice that doesn't drift into legal or political opinions.

2. MSME Empowerment (Data Unlock)

For small business owners, navigating export incentives is a nightmare of red tape. Using the Model Context Protocol (**MCP**), the collaboration is building a prototype that "unlocks" data on government schemes and policies that are scattered around a multitude of open public sources. The ultimate goal is to enable agents to provide accurate, unbiased, and intelligence on export policies and promotion schemes, without the risk of uncertainty from "lock-in" to a single data provider.

Following are the key safety pillars that are adhered to in building prototypes for these two sectors:

Safety Pillar	Indian Context Application
Steerability	Ensuring the AI stays in "advisor" mode and doesn't promote specific brands.
Uncertainty Awareness	Flagging when data is outdated rather than guessing a turnover threshold.
Bias Neutrality	Avoiding ideological positioning in policy-related queries.

Table 1: Safety Pillars and Applications

BUILDING LOCAL CAPACITY: THE CLAUDE CODE REVOLUTION

Safety is not just about the model; it's about the people building with it. EkStep is leveraging **Claude Code** to empower a new generation of developers in the non-profit and public sectors. By using AI-native approaches to refactor existing Digital Public Goods (DPGs), small, nimble teams can now deploy at a scale and speed previously impossible.

Furthermore, EkStep is supporting AI research fellowships across Indian universities. These researchers are tackling uniquely Indian challenges: how voice systems perform in rural settings, and how to map Indian-specific risks, like caste or regional bias, into safety frameworks. One such example is the ongoing steerability project in collaboration with IIIT Hyderabad and EkStep.

The core objective is to develop a steerability framework that tailors language model (LLM) responses to the Indian context, particularly based on users' regional linguistic patterns. The current approach uses Anthropic's persona vectors and Deep Embedded Clustering (DEC) to categorize LLM outputs into one of several distinct "personas" based on a small set of features. Using Claude Code, the researchers are able to swiftly build prototypes to validate these models with “personas”.

A GLOBAL EXCHANGE OF KNOWLEDGE

The partnership between EkStep and Anthropic amounts to something unusual in the tech world: a genuine two-way exchange helping scale deployments of frontier AI to 1.4 billion people.

What a farmer in Gujarat or a small manufacturer in Uttar Pradesh needs from an AI system — reliability in low-connectivity settings, accuracy across dozens of languages and dialects, trustworthiness with critical decisions — are not just local problems to solve. If a system can be made to work at the scale of a billion across that range of conditions, it can be made to work for anyone, anywhere.

Building AI with guardrails ensures a safe future where AI innovation drives extraordinary outcomes at population scale.

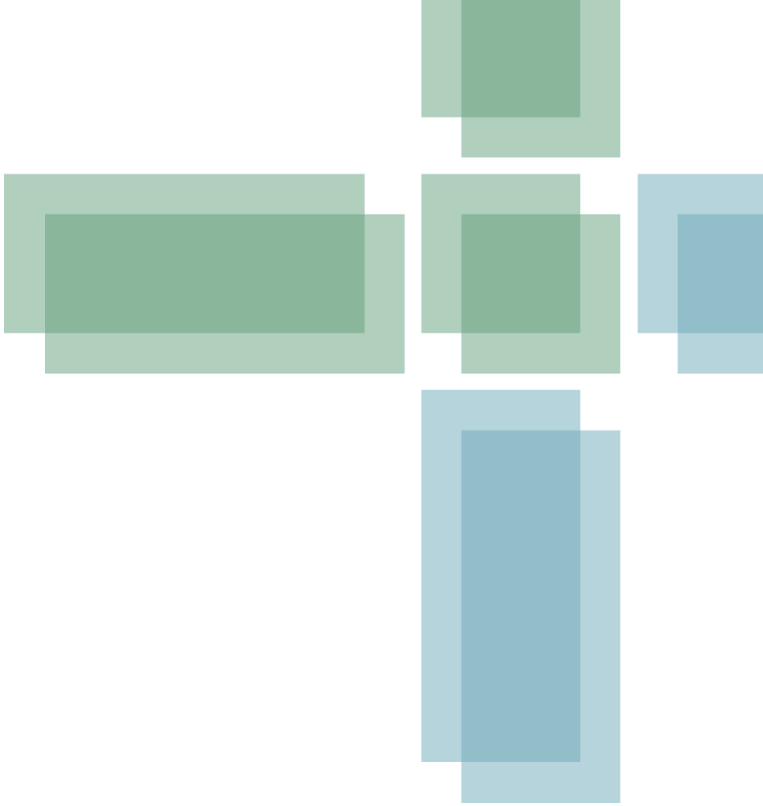
References:
1. https://www.anthropic.com/constitution
2. https://www.anthropic.com/research/persona-vectors

ACKNOWLEDGEMENTS

This article was authored by People+ai, EkStep in collaboration with Anthropic, featuring contributions from the following people:

1. Nitarshan Rajkumar, International Policies, Anthropic
2. Sally Aldous, Communications, Anthropic
3. Jagadish Babu, Chief Operating Officer, EkStep Foundation
4. Shalini Kapoor, Chief Strategist, Data & AI, EkStep Foundation
5. Aparajita Choudhury K, Solutions, Data & AI, People+ai
6. Anuj Gupta, Product Manager, People+ai
7. Lakshmanan Nataraj, Technical Advisor, People+ai
8. Tanvi Lall, Director, Strategy, People+ai
9. David, Menezes, Director, Programs, People+ai
10. Sonia Menezes, Technical Writer, People+ai
11. Kirti Pandey, Mission Director, OAN
12. Ponnurangam Kumaraguru, Professor Computer Science, IIIT Hyderabad
13. Sriharini Margapuri, Research Assistant, IIIT Hyderabad
14. Hari Shankar, MS student, IIIT Hyderabad
15. Vedanta S P, Research Assistant, , IIIT Kottayam
16. Abhijnan Chakraborty, Assistant Professor, Computer Science, IIIT Kottayam

Designed by: Kumar Kanishka, People+ai



ANTHROP\C

people + ai

An EkStep Initiative