Ministry of Electronics & IT

IndiaAI Scales Up Safe AI Efforts with Cutting-Edge Solutions for Deepfake Detection, Bias Mitigation and AI Penetration Testing

Five Game-Changing Proposals Selected to Advance Safe & Trusted AI Ecosystem in India

‘Saakshya’, ‘AI Vishleshak’ and IIT Kharagpur’s Voice Detection System to Fortify India’s AI Forensics and Security

Digital Futures Lab & Karya Tackle Gender Bias in Agricultural AI; Globals ITES and IIIT Dharwad to Develop Tools for Generative AI Security

Posted On: 07 OCT 2025 1:06PM by PIB Delhi

As AI technologies evolve at an unprecedented pace, ensuring their safe, transparent, and trustworthy use has become a national imperative. India is committed to fostering agile and robust mechanisms for developing indigenous governance tools, frameworks, and guidelines that reflect the country’s unique socio-technical context.

To advance this vision, IndiaAI had launched the second round of Expression of Interest (EoI) under the ‘Safe & Trusted AI’ pillar, across a range of critical themes on December 10, 2024.

More than 400 proposals were received from reputed Academic Institutions, Start-ups, Research Organisations & Civil Society. A multi-stakeholder committee was created to provide technical expertise for the evaluation of the proposals, resulting in the selection of 5 projects across various themes. Collectively, these projects translate “Safe & Trusted AI” into practice combining resilience testing, bias audits to support responsible development and deployment of AI.

List of Projects selected under Second EOI in Safe & Trusted AI pillar of IndiaAI Mission

Theme	Project Title	Selected Applicant

Deepfake Detection Tool	Saakshya: Multi-Agent, RAG-Enhanced Framework for Deepfake Detection and Governance	IIT Jodhpur (CI) & IIT Madras
	AI Vishleshak: Improving Audio-Visual Deepfake Detection and Handwritten Signature Forgery Detection with Adversarial Robustness, Explainability & Domain Generalization	IIT Mandi & Directorate of Forensic Services, Himachal Pradesh
	Real-Time Voice Deepfake Detection System	IIT Kharagpur
Bias Mitigation	Evaluating Gender Bias in Agriculture LLMs-Creating Digital Public Goods (DPG) for Benchmarking and Fair Data Work	Digital Futures Lab & Karya
Penetration Testing & Evaluation	Anvil: Penetration Testing & Evaluation Tool for LLM and Generative AI	Globals ITES Pvt Ltd & IIIT Dharwad

The five selected projects under the second EoI demonstrate IndiaAI's commitment to translating the vision of Safe & Trusted AI into concrete solutions. These initiatives will advance real-time deepfake detection, strengthen forensic analysis, address bias in AI models, and build robust evaluation tools for generative AI, ensuring that AI systems deployed in India are reliable, secure, and inclusive. By bringing together leading academic institutions, industry partners, and civil society, the IndiaAI Mission continues to foster innovation, ethical practices, and a resilient AI ecosystem that serves the diverse needs of the nation.

About IndiaAI

IndiaAI, an Independent Business Division under MeitY, is the implementation agency for the IndiaAI Mission. It strives to democratize the benefits of AI across all strata of society, bolster India's leadership in AI, foster technological self-reliance, and ensure the ethical and responsible use of AI.

***

Dharmendra Tewari\Navin Sreejith