India's evolving AI ecosystem

By A Damodaran

After the proverbial digital tech stack, artificial intelligence is hailed as the next big thing for India in the global tech arena. Sometime ago, Niti Aayog had estimated that AI would lead to a 15% boost in gross value added to the Indian economy by 2035.

More recently, tech forecasters have predicted that AI would add USD1,000 billion to India's GDP by 2035. India's AI growth story has been fuelled by well-targeted investments by national and global players. Wipro, Infosys, HCL Tech, TCS, Tech Mahindra, and LTI Mindtree have been exploring generative-AI models customised for enterprise-based solutions. IIT Delhi,

NIT Trichy, and IIT Roorkee have centres of excellence working on AI-related R&D projects that focus on health, finance, business, transportation, and societal crisis management, among others.

R&D projects

What makes India's AI scenario noteworthy is the euphoric response of global tech majors to the country's evolving AI ecosystem. India is home to significant R&D investments in AI. The Nvidia-CSIR Centre of Excellence initiative, which goes back to 2018, seeks to develop AI-based applications on the foundation of industry-class supercomputing. Microsoft Azure's OpenAI is the catalyst for AI4Bharat, a government-backed initiative piloted by IIT Madras that focuses on natural language processing (NLP) models (more on this later).

Another collaborative venture is Vaani, an AI-enabled multilingual project that Google has initiated in partnership with Indian Institute of Science and Artpark. This project seeks to generate a large language model (LLM) model for Indian languages based on Google's Bidirectional Encoder Representations from Transformers.

The startup scene

India's AI startups have largely focused on scalable solutions involving machine learning, computer vision, and edge computing. The targeted sectors include finance, actuarial and banking, e-commerce, high-tech farming systems (which make use of satellite and geospatial data to guide crop husbandry practices) and health-sector ventures that utilise AI for rapid screening of diseases.

AI-powered language-translation platforms add to the repertoire. Transliteration and translation services make extensive use of deep-learning tools and NLPs. The Centre for Visual Information Technology of IIIT Hyderabad has developed tools to map Indian languages based on audio samples, using deep-learning techniques. The AI4Bharat programme has spawned open-source projects focused on Indian languages. Being open source, these projects have attracted inputs from talented developers in GitHub.

With generative AI increasingly finding applications in process automation, a few startups have taken up generative-AI projects in manufacturing. In the meantime, the advent of AI chatbots in educational institutions has led to new ventures that focus on tools that detect AI-generated content in test papers and assignments.

AI, language, and the Web

The People's Linguistic Survey India places the number of languages in India at 780. Many "to-be-extinct" languages are without scripts. Consequently, audio sources become

critical for their revival. The social costs of languages becoming extinct can be tremendous for India since it could lead to the disappearance of tacitly held knowledge systems associated with these languages.

Indeed, India's linguistic diversity has spun interesting AI based language projects. By virtue of their ability to translate the aural to the oral and the oral to the text, deep-learning tools and AI-enabled language models such as BERT, GPTs, and XLNET, can play a major role in reversing the extinction of orphaned languages. However, there is a catch.

The logic of economic viability causes LLMs to gravitate towards languages that scale up well on the Web. LLMs such as GPT 4, which are over 1 trillion parameters strong, require humongous training data, most of which are sourced from the Web.

The sourced data is lodged in centralised cloud data centres. Unlike English, India' demographically formidable languages such as Hindi and Tamil cannot make the cut in the LLM scheme of things since they have low Web presence. Consequently, they fall in the category of "low-resource languages". While the Vaani project and Meta Inc's Meta Meta Tool (MMT) have developed training data sets for a fairly large number of languages, it is unviable for these ventures to take up languages with low demographic quotient. Further, cloud-based centralised data-management systems will not work out to be economical for low-

population languages.

It follows that any low-resource language with unfavourable demographics, cannot advance to the large-resource language category, without public investments. More fundamentally, communities speaking such languages need to have access both to the Internet and the Web.

The Bhashini programme of the Ministry of Electronics and Information Technology (MEITY) seeks to open Internet access for non-English speaking Indians. What is especially noteworthy is the strong policy buy-in here, with Prime Minister Narendra Modi pushing for natural language processing (NLP) tools to impart quality education in regional languages and dialects. The larger challenge is to ensure equitable access to the Web for speakers and writers of demographically disadvantaged languages. It is widely conceded that the current version of the Web is controlled by a handful of tech platforms. This could change with Web3, which seeks to uphold the principles of inclusivity and democratisation in content creation and protection. However, in the Web3 scheme of things, there are no gatekeepers to restrain users from posting harmful content. Since LLMs are trained on Web-based data, using unsupervised learning tools, they are vulnerable to the risks of hallucination, misrepresentation, data biases, and discrimination. One

way out is to lay down standards for unsupervised learning algorithms under the proposed Digital India Act.

A more fundamental move would be to ensure that India plays a prominent role in the formulation of content-creating norms for Web3.

An AI ecosystem with Indian characteristics

During his recent visit to India, Sam Altman was fulsome in his praise of India's dynamic AI user base. In the coming years, as the number of Internet users hits 900 million, the country will experience a rapid surge in demand for AI-enabled services. One of the biggest concerns expressed by AI critics is about the emergence of centralised AI apparatuses (mammoth LLMs and super-intelligent computers) in data-rich countries, which could suffocate alternative channels of content creation and utilisation in these countries.

Judging by its present pattern of evolution, India's AI ecosystem is fated to follow a dualistic-system whereby LLMs and large cloud-data apparatuses will coexist with "community-based" generative-AI systems that work on smaller language models and small data systems. For such a dualistic order to function well, it is essential that India develops a balanced AI infrastructure system that seeks to augment both supercomputing and distributed-computing systems in a complementary manner.

The author is senior visiting professor at the Digital Economy, Startup and Innovation (DESI), ICRIER

The infrastructure to develop skills