Case Study

Service Providers
Intel® Xeon® Scalable processors

AI model based on Indian languages connects broader society to technology

Tech Mahindra's Makers Lab created an open-source language model on Intel® Xeon® processors called Project Indus that includes Hindi and 37 regional dialects. Project Indus makes AI more accessible to people across India who do not speak English, young and old alike, including those in rural communities.

At a glance

- Tech Mahindra built a large language model (LLM) based on Indian languages, supporting Hindi and 37 regional Hindi-based dialects.
- The solution runs on high-performance, cost-effective 5th Generation Intel® Xeon® Scalable processors and has also been tested on Intel® AI PCs.
- Farmers can get advice to improve the health and yield of their crops, while people in rural communities can access loans and financial advice, and children can research schoolwork in their native dialect.
- Running on 5th Generation Intel Xeon Scalable processors, the Indus LLM processed an average of 33.9984 tokens per second.[3] The total response time was between 0.249 seconds for 22 tokens and 4.27 seconds for 167 tokens.[3] The average end-to-end latency was 3.07 seconds.[3]

Executive summary

People across India speak more than 19,000 dialects. However, because many large language models (LLMs) are based on English, it is hard for non-English speakers to access the benefits of AI technology. Tech Mahindra's innovation center, Makers Lab, set out to create an LLM for India based on Indian languages. On-the-ground teams collected dialect samples from villages, and citizens submitted 150,000 additional samples online. The samples then helped to train an AI model with 1.2 billion parameters called Project Indus. The model was deliberately limited in size so small businesses and individuals could readily deploy it.

The Indus LLM has many applications across India. For example, Tech Mahindra group companies such as Mahindra Finance are looking to potentially use it to support low-income families with financial services in their dialects, while students benefit from chat-based AI features. The solution runs on high-performance, cost-effective 5th Generation Intel® Xeon® Scalable processors and has also been tested on Intel® AI PCs.

Challenge

There's huge potential for innovation across India, but many people can't access the technology they need because it doesn't support their dialect.

"There's a language barrier," says Nikhil Malhotra, Chief Innovation Officer and Global Head of AI and Emerging Tech at Tech Mahindra. "Once you go to the smaller cities in India, only 20% of people speak English."

Across India, 27 officially recognized languages have more than 1,600 dialects. Unofficially, there are more than 19,000 dialects. "Every 50 kilometers in India, you see the dialect change," says Malhotra. "Global tech companies do a good job with the basic languages but don't cover the dialects. A large sector of the population is missed out when the dialects are missed out."

Nikhil Malhotra, Chief Innovation Officer and Global Head of AI and Emerging Tech at Tech Mahindra.

Makers Lab is Tech Mahindra's innovation center, established in 2014. Its missions are to drive research and development initiatives that benefit customers and India's economy by creating more sustainable artificial intelligence solutions.

Makers Lab took on the challenge of creating Project Indus, a large language model (LLM) that would enable conversational systems to bridge the gaps between urban and rural populations across India.

With so many languages spoken in India, Malhotra and his team had to pick one to start with. "We thought we'd take one of the most widely spoken languages in the world, which is Hindi," says Malhotra. "It is spoken in 42-plus different dialects, and some are endangered. Our first principle was to ensure that we include the dialects in our LLM so that we can preserve them."

"Project Indus was also a beacon to the world to say that a big IT services company can create a large language model from the ground up without basing its model on ChatGPT or Meta's Llama," he adds.

While Tech Mahindra was looking for commercial value from Project Indus, Malhotra and his team also planned to ignite innovation by making it open source. For the model to be used widely, it had to be possible for small enterprises, and even individuals, to run it.

"Is it really necessary for a large language model to require GPUs to solve a problem?" asks Malhotra. "Not many customers, whether enterprises or young innovators, can sustain GPUs. They're expensive. Our philosophy of making India the producer, and not just consumer, of technology can only be done with frugality. We don't have enough money to spend a lot on data centers and GPUs. We needed to ensure GPUs were not a requirement, and that it could run on a typical PC as well as a server."

Solution: Open-source AI for India

One of the biggest challenges was gathering dialect samples from across India to train the AI model. "We sent teams to the Hindi heartland. We had to speak to the village elders, and it was hard to convince them because they were concerned about giving away their language," says Malhotra. "It was difficult to find people who knew the local dialects and our dialect to help with translation."

Makers Lab also set up a website where visitors could submit recordings in their dialect without providing personal information. The non-profit Tech Mahindra Foundation, which connects young people with employment opportunities, also supported collecting dialect samples. "We were helped by a lot of people as this was a cause for India," says Malhotra. "People came forward to provide dialects."

While some had estimated the cost of building an LLM from scratch at millions of dollars, Makers Lab created Project Indus with a total spend to date of $400,000. "This is in part thanks to support from Intel," says Malhotra.

Project Indus today covers Hindi with 37 dialects and can be used with a voice or chatbot interface. It is available to download from the AI community Hugging Face and to use online within India.

Solution spotlight

- Tech Mahindra's Project Indus is a large language model (LLM) that processes language in Hindi, including 37 dialects.
- The LLM can be used with text input, for example, through a chatbot, or can use a voice interface.
- The solution is open source and runs on Intel® AI PCs and servers without GPUs, so small businesses and individuals can easily adopt it.
- Benchmarking on 5th Generation Intel® Xeon® Scalable processors showed the system processes an average of 33.9984 tokens per second.[3]
- Applications include agriculture technology, financial advice, technology access for young people, and language preservation.

Makers Lab has benchmarked the Indus LLM on 5th Generation Intel® Xeon® Scalable processors. Intel® Xeon® processors are the most benchmarked host processors for AI accelerators.[1] 5th Generation Intel Xeon processors deliver up to 62% lower TCO on popular AI workloads compared to alternative processors.[2]

Intel® Advanced Matrix Extensions (Intel® AMX) and Intel® Advanced Vector Extensions 512 (Intel® AVX-512) provide acceleration for AI tasks in Intel Xeon Scalable processors. "We wanted to optimize our solution to run well on Intel processors using these accelerators," says Malhotra. In addition, the solution was optimized for non-uniform memory access (NUMA), a memory architecture where processors share some memory.

Makers Lab has worked with Intel to test the Indus LLM on Intel® AI PCs. These devices incorporate a CPU, GPU and neural processing unit (NPU) to handle AI tasks locally and more efficiently. "We're closely knitted with Intel, and the moment we change our model, we send it to the Intel engineering team. They test it within 24 to 48 hours and confirm it's running right," says Malhotra.

Results

Makers Lab wanted to ensure its solution performs well on popular benchmarks for large language models. These include Indic Eval, which assesses LLMs using Indic languages across various tasks. "Project Indus delivered competitive performance for Indic Eval, despite our model having far fewer parameters than alternative models we compared it with," says Malhotra. "We also performed well on the ARC challenge, which measures reasoning, and Hellaswag, a sentence completion benchmark."

From a user perspective, other benchmarks were also important. Tokens are letter sequences that the AI model recognizes. "When you run on a CPU, the important benchmarks for us are how long it takes to get to the first output token, the delay between two tokens, the input prompt length, the output length and the total throughput. The Intel engineering team helped me because I didn't have a way to do an oranges-to-oranges comparison," says Malhotra.

Across India, 27 officially recognized languages have more than 1,600 dialects. Unofficially, there are more than 19,000 dialects.

Running on 5th Gen Intel Xeon Scalable processors, the Indus LLM processed an average of 33.9984 tokens per second.[3] The total response time was between 0.249 seconds for 22 tokens and 4.27 seconds for 167 tokens.[3] The average end-to-end latency was 3.07 seconds.[3]

Makers Lab has a range of use cases planned for Project Indus. While children often learn English at school, their thought processes for learning may be in their native language. It would be helpful for them to be able to research using their first language. "That's one of the biggest use cases for personalized education you can give kids," says Malhotra.

"Farmers can speak to their phone in their dialect and get advice on which pesticides to use, on which crop, in what timeframe, and what the precipitation and soil indexes are," says Malhotra. "Mahindra Group also offers low-income finance across the country, and many queries come in in Hindi and its dialects. We're using Indus now to automate many tasks within Mahindra Finance. Customers can speak in their dialect and resolve their problems much faster."

Future innovations

Malhotra is excited by the potential of the Intel AI PCs to support education. "Some kids walk a long way to school and back each day," he says. "Imagine these kids now have an AI PC in the village community center. They can ask questions and get answers in their dialect, even though there is no representation of the topic they're asking about in English or their dialect. You give those kids a chance to make a better life for themselves. I think that's one of the biggest advantages of AI PCs for us."

He adds: "AI PCs will give almost everyone a decentralized AI system in the future, helping them with any number of jobs they want to do, such as travel bookings, managing expenses, and recipes with videos and pictures."

Tech Mahindra has also built an LLM for Indonesia, where more than 700 languages are spoken today. "Many countries are looking at the sovereignty of AI now," says Malhotra. "Each country has different cultural biases and doesn't want another country's biases projected into their AI conversations."

"For innovation, you need a lot of collaboration," he says. "You also need to be humble. I don't think you can be egotistical with technology. You've got to make hundreds of errors to get a success, so empathy toward yourself and the task at hand is important."

Learn more

- Download Project Indus at Hugging Face
- Project Indus at Tech Mahindra Makers Lab
- Intel® AI PCs
- Intel® Xeon® Scalable processors
- White paper: Benchmarking the Indus language model on Intel® AI hardware

intel ai

[1] Based on MLPerf benchmark testing as of 2024. For details, visit https://mlcommons.org/.

[2] 5th Gen Intel® Xeon® processors deliver up to 3.49x faster performance than the 4th Gen AMD EPYC while running a real-time Natural Language Processing inference (DistilBERT) workload. This performance drives a fleet reduction from 50 to 15 servers which, over 4 years, saves: 1496.5 MWH of energy, 634,428 kg CO2 emissions, and $1,300k of cost. See T206 at intel.com/processorclaims: 5th Gen Intel Xeon Scalable processors. Results may vary.

[3] For full specifications and testing methodology, see the white paper 'Benchmarking the Indus language model on Intel® AI hardware'.

Availability of accelerators varies depending on SKU. Visit the Intel Product Specifications page for additional product details.

Performance varies by use, configuration and other factors. Learn more at intel.com/performanceindex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Details at www.intel.com/PerformanceIndex. Results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

0225/SB/CAT/PDF ♻ Please Recycle 364210-001US