

IEEE Spectrum

NEWS AI

Small AI Models Gain Traction Around the World > In places with unreliable networks and no data-center infrastructure, smaller is better

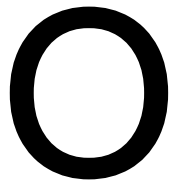
BY DAVID BERREBY

10 HOURS AGO



José Alberto Ferreira, a researcher at the Patient Simulator Lab at the

2020 REPORTER FOR THE YEAR, A RESEARCHER AT THE TATIANA GEMELLI LAB AT THE University of Itajubá, in Brazil, tests a TinyML (Tiny Machine Learning) model that generates electrocardiograms. MARCELO ROVAI



NE MORNING IN 2019, ADEBAYO ALONGE WAS

in a Cape Town hotel room, preparing to demonstrate his startup's AI answer to a serious problem in African health care: counterfeit medication, which kills thousands of people across the continent every year.

The RxScanner is a handheld spectrometer that scans a pill with infrared light, then sends the item's molecular profile to an AI model equipped with a pharmaceutical database. In seconds, the AI identifies the medication from its molecular profile—or reports that it's phony.

Pharmacies were using the system in more than a dozen countries, including Ghana, Kenya, Myanmar, and Alonge's native Nigeria. But that morning in South Africa, it didn't work. "I was shocked," Alonge says.

The spectrometer connected to the AI model—but the data center was 14,000 kilometers away and bandwidth was limited. "Our server was in the United States, and just to get the result of a single scan was taking me over 5 minutes."

So Alonge immediately asked his engineers to shrink the AI model down to a smaller, low-power, unconnected version that could run entirely on his Android phone. They produced it 2 hours later, and that saved the demo.

More importantly, the work birthed a new version of his device, which can authenticate a pill in places without broadband, computers, or even reliable electricity. It also turned Alonge into an advocate for this kind of “small AI.”

Small AI for Global Health Care Access

Small AI is a far cry from wealthy nations’ colossal large language models (LLMs), hyperscale data centers, multibillion-dollar investments, and debates about AI consciousness. But for millions of people around the world, the only AI that matters, and often the only kind available, is small. (According to a World Bank Report issued in November, only 0.7 percent of internet users in the world’s poorest countries have used ChatGPT, compared to a quarter of all internet users in the most developed nations.)

“Most people are discussing AI from the LLM/generative side. But that needs a lot of computing power, electricity, massive data, and skilled people to manage it.” Aiav Banga, president

of the World Bank, said last January at the World Economic Forum, in Davos. “Outside the developed world, other than maybe India and China, very few countries have that combination.”

By contrast, small AI can deliver useful, even life-saving services to people in areas that have none of those things, Banga said. In India, where the government’s AI plans call for more development of small AI, many such systems are working for farmers.

For example, a drone-based system developed by Bala Murugan and colleagues at the Vellore Institute of Technology, in India, takes photos of cashew plants and quickly identifies those with splotches that indicate disease. All the processing takes place on the drone itself, so there’s no need for a computer on-site, nor for a connection to a central server.

Using small language models trained for a specific problem, and sometimes running on cheap, low-power devices, other small-AI implementations have been developed to identify ant infestations in a Uruguayan vineyard, detect the presence of malaria-carrying mosquitoes in a number of nations, and run electrocardiograms from an Arduino device in parts of Brazil

that lack access to more complex equipment.

“This is the most important area in AI nowadays,” says Marcelo José Rovai, a professor at the Institute of Engineering and Information Systems at the Federal University of Itajubá, in Brazil, who was involved in all three projects. “It’s growing very fast.”

Low-Power, Small-AI Models on Devices



Small AI models can run on a variety of low-power devices, including [from left to right] an Arduino Nano 33 BLE Sense, a Sseeed Wio Terminal, and an Arduino Portenta. MOEZ ALTAYEB

For Alonge, Rovai, and other advocates, small AI is not just “a promising trend,” as that November World Bank report calls it. It may be, in the long term, the form of AI that will touch

the most lives and remain sustainable after some of the giant models become too costly for most users.

“I think the future of AI is not like one giant model, at a center. I think it’s millions of small, precise models deployed at the edge, each one solving like a specific problem, a specific context,” Alonge says. This is partly because much of humanity—including people in parts of rich countries as well as the developing world—lives without access to cutting-edge frontier models. But, he says, it’s also because those models are not sustainable.

“If someone is not subsidizing it, most people will not be able to afford those models. So those of us who are said to be small-AI developers are the ones who will have to build for the majority of the world,” Alonge says.

There is no strict definition of “small AI,” but people often use the term for language models with at most a few billion parameters. (Compare that to cutting-edge models, which can include more than a trillion.) That’s small enough to run directly on a phone or a Raspberry Pi. That’s what allows these applications to run on devices without a connection to a data center and use only a few watts of power, often supplied by a battery or a solar panel.

Despite their small footprint, these models aren't fundamentally different technology from that of gigantic AI models, Rovai says. Many instances of small language models were created the same way the phone-based version of Alonge's pharmaceuticals scanner was—by “pruning” large models, or removing the parameters that weren't involved in the task. The result is a system that's less capable generally but still very good at the specific job it was pruned for, Rovai says.





A lighter version of RxAll's RxScanner spectrometer sends its results to an AI model run locally on a phone to check that a drug's molecular signature is genuine. RXALL

Other small models are created by “distillation.” They are trained to mimic a large model, until their performance approaches that of their “teacher,” Rovai says. In other cases, a larger model's precision is reduced, for example, so that a model run on 32-bit architecture can run on 8-bit designs. In situations where the machine learning application is being used to classify data or predict patterns (like an ant infestation), it's trained from the beginning on a small device, not derived from a larger model at all.

Running all these small, specialized systems is becoming easier, Rovai says, for two reasons.

The first reason is that hardware is getting better and more capable while using less power, he says. This means more and more phones can run small AI—especially those equipped with neural processing units, which are specialized chips that handle AI tasks like facial recognition and changing the brightness, shadows, or contrast in a photo.

In 2025, slightly more than a third of all smartphones shipped worldwide were capable of running generative AI and that

models are capable of running generatively, and that figure will reach 45 percent by the end of this year, according to the technology research firm Counterpoint. By the end of next year, slightly more than half of all smartphones will be able to run a small AI model.

The second reason Rovai cites is the shrinking footprint of language models. Both Google DeepMind's Gemma 4 (released in April) and Alibaba's Qwen 3.5 are “fantastic” for small AI, Rovai says. Both models are “open weight,” meaning users can adjust the connections between parameters to suit their needs. This makes it easy, for example, “to take a lot of data from, say, the milk industry and retrain the model specifically on that,” Rovai says.

Rovai illustrated these reasons on a Zoom call, using one of his most recent experiments. Holding up a device, he says, “This is the new Arduino UNO Q—a US \$50 device with a Qualcomm chipset. I’m running a language model here, which collects data from sensors and analyzes that data to detect tiny pools of water where mosquitoes might be breeding. It takes 3 watts to run it.”

Support for Small-AI Development

Convinced that millions of people are already benefiting from

these kinds of applications, the World Bank now actively promotes small AI with grants, mentorship programs, financing, technical advice, and models of government policies that are friendly for small-AI development. For example, in Rwanda, the World Bank is backing a government program to help low-income households get devices that can run AI.

All that said, no one claims that large language models are going away entirely. To create a generative AI that can run on a phone or other small device requires the architectural insights, data processing, and results of a larger model, Rovai says. “We need the big models to create these smaller models.”

And for all that small AI can benefit people without access to big AI, the technology can’t solve the larger problems of development and digital inequality, Alonge says. Implementing small AI won’t allow nations to escape the challenge of creating an ecosystem to support AI: reliable power, a supply chain that works, and an educational system that develops the talents needed to create AI tools.

Though his drug-scanning system can run for days on a phone with no connection, “you still want to be able to enable

periodic syncing for updates with new signatures for the medications and analytics,” Alonge says. “And even when you are using batteries, reliable power is important. That phone battery is not going to last forever.”

In many parts of the world, the future of small AI isn’t assured, he says. “It works, and many places will eventually need to use it. The question is whether or not the political actors are wise enough to invest in infrastructure to support it long term.”