Why Apple Is Talking to a 15-Person Startup Instead of Building Bigger AI Servers

TechRounder PDF Edition

Live article:
https://www.techrounder.com/ai/why-apple-is-talking-to-a-15-person-startup-instead-of-building-bigger-ai-servers/

By Vipin PG | Published July 15, 2026 | Updated July 15, 2026 | Format: Explainer | 7 min read

In brief

Apple is evaluating technology from PrismML, a Caltech-founded startup, that compresses huge AI models so they can run entirely on an iPhone instead of a data center. PrismML already shrank a 27-billion-parameter model from 54GB down to under 4GB - small enough to fit on a modern iPhone. Talks are very early, nothing is confirmed, and this technology won't be part of the iOS 27 update launching later this year. But it points to where Apple's AI strategy is likely headed next.

Apple has spent years telling us that Siri will get smarter. This time, the path there might not run through Apple's own labs at all - it might run through a 15-person startup in Pasadena that figured out how to shrink a massive AI model down to the size of a few phone photos.

Apple is reportedly in early talks with PrismML, a startup spun out of Caltech, about technology that could let iPhones run AI models that are currently far too large to fit on any phone. If this actually makes it into a future iPhone, it could mean a Siri that's not just faster, but genuinely more capable - without ever needing an internet connection.

Here's what's actually going on, what it means for your iPhone, and what's still just speculation.

What's Actually Happening

PrismML was founded by Babak Hassibi, a Caltech electrical engineering professor, and is backed by Khosla Ventures along with support from Google and Cerberus Capital. The company recently came out of stealth mode with a $16.25 million seed round - a relatively small amount by Silicon Valley standards, which makes the level of attention it's getting from Apple somewhat unusual.

According to CNBC, Apple has been in discussions with PrismML about using its model-compression technology to run much larger, more capable AI models directly on iPhones. Hassibi confirmed the talks himself, describing them as very early but said that things are "progressing nicely." Apple hasn't commented publicly.

It's worth being clear about what this is and isn't. This is not an announced product, a signed deal, or a confirmed acquisition. It's a technology evaluation - the kind of conversation Apple has with dozens of startups every year. But the specific problem PrismML claims to have solved is one Apple has been visibly struggling with, which is why this particular story has generated so much attention.

The Big Number: From 54GB Down to Under 4GB

To understand why this matters, you need one number: PrismML took an AI model that normally requires about 54GB of memory and compressed it down to under 4GB - while keeping all 27 billion of its internal parameters active and usable.

For context, a 54GB model has no realistic path onto a phone. Even a high-end iPhone typically has only 8 to 12GB of memory total, and the operating system only allows a single app to use a small slice of that - usually around 6GB at most. A 54GB model isn't just tight, it's impossible.

PrismML's compressed version, which it calls Bonsai 27B, was officially released on July 14, 2026, and is free to download under an open license. The company says its most compressed version keeps roughly 90% of the original model's intelligence, and a slightly larger version (around 6-7GB, aimed more at laptops) keeps closer to 95%. On a recent iPhone, PrismML says the compressed model can generate responses at a usable, real-time chat speed.

How Do You Shrink an AI Model That Much?

You don't need to understand neural networks to get the basic idea. Think about packing for a trip with only a carry-on bag instead of a full suitcase.

A regular AI model stores every piece of information with a lot of precision - like packing every outfit on its own hanger, with room to spare. That takes up a lot of space, but it's not wasted space; it's what lets the model be so detailed and accurate.

PrismML's approach is closer to vacuum-sealing everything. Instead of storing each piece of information with fine detail, it forces the model to represent almost everything using extremely simple values - in the most aggressive version, each individual data point can only be one of two states, essentially an "on" or "off" switch. That's a drastic simplification, and normally it would badly damage how the model performs.

PrismML's key claim is that it built the model this way from the start, rather than shrinking a finished model after the fact. That approach, it says, avoids the kind of quality loss that usually comes with heavy compression.

What You Gain, and What You Lose

No compression is completely free. Here's the honest trade-off, based on PrismML's own published numbers:

What Changes: Storage size | Effect: Drops from 54GB to under 4GB

What Changes: Speed | Effect: Several times faster than a full-size model on the same hardware

What Changes: Battery use | Effect: Significantly lower than running a full-size model

What Changes: Reasoning, math, coding | Effect: Stays close to the original model's ability

What Changes: Recalling obscure facts | Effect: Noticeably weaker than the original

That last point is actually a smart trade-off for a phone assistant. Siri doesn't need to memorize every historical date or trivia fact - it can just search the web for that. What it does need is the ability to understand what you're asking, follow multi-step instructions, and reason clearly. That's exactly the part PrismML's approach protects the most.

Would This Actually Make Siri Smarter?

Right now, Apple's current on-device AI model - called AFM 3 Core Advanced - has 20 billion parameters. But it doesn't use all of them at once. It only activates a small portion, roughly 1 to 4 billion parameters, for any given task. This is a common trick used to keep AI models fast and light on phones, but it also limits how much the model can reason about at one time.

PrismML’s compressed model works differently. Even though it’s more heavily compressed, all 27 billion of its parameters stay active during use. In practical terms, that means it can potentially hold more context, follow more complex instructions, and handle more advanced tasks - like coding help or multi-step planning - directly on your phone, without needing to activate an internet connection.

In short: this isn’t just about making Siri respond a little quicker. It’s about giving it noticeably more reasoning power while keeping everything on the device.

Why This Fits Apple’s Privacy Strategy

Apple has built its brand around the idea that your data stays on your device whenever possible. Right now, when Siri needs to handle something more complex, it often has to send that request to Apple’s cloud servers for processing. Apple secures that connection carefully, but it’s still a round-trip over the internet.

A model that runs entirely on the iPhone removes that step completely. That means:

- Faster responses, since there’s no waiting on a server
- It keeps working even with no signal or internet connection
- Sensitive information - health data, messages, financial details - never has to leave your phone

For a company that markets privacy as a core feature, this kind of technology lines up perfectly with where Apple has said it wants to go.

Don’t Expect This in iOS 27

Apple is holding a developer session on July 23, 2026, focused on the new Siri AI features and the "Liquid Glass" design update coming with iOS 27. It’s tempting to connect the two stories, but they’re not related - at least not yet.

These talks with PrismML are still in the early evaluation stage. Turning this kind of compression technology into something that runs reliably across hundreds of millions of iPhones takes serious engineering time - testing battery behavior, heat management, stability under long use, and more. None of that happens overnight.

So this year’s iOS 27 update will still rely on Apple’s existing on-device model. If anything from this partnership ever ships, it’s realistically a story for a future iPhone and a future iOS version, not this one.

Apple Isn’t the Only One Racing to Shrink AI

This isn’t happening in isolation. Google and Samsung have already been down this road for a while. Google’s Gemini Nano runs directly on Android phones through a system called AICore, and Samsung has built its entire "Galaxy AI" push around a similar mix of on-device and cloud processing, with a public goal of reaching 800 million AI-enabled Galaxy devices by the end of 2026.

In that context, Apple’s interest in PrismML looks less like innovation for its own sake and more like an attempt to catch up - and possibly leapfrog - competitors who’ve had a head start on practical on-device AI.

What Happens Next?

A few different things could realistically happen from here, and it’s worth keeping expectations grounded:

- Apple licenses the technology to improve its own AI models without acquiring the company
- Apple acquires PrismML outright - something it has done before with small AI teams
- The talks simply don’t lead anywhere , which happens often with early-stage technology evaluations

There’s also a simpler possibility worth mentioning: PrismML has every incentive to publicize its connection to Apple, since it brings attention to a young startup with a tiny seed round. That doesn’t mean the story isn’t real - multiple outlets have confirmed the talks are happening - but it’s a good reminder to treat "in early talks" as exactly that, and nothing more, until Apple says otherwise.

Meanwhile, PrismML isn’t waiting around. It has already open-sourced its compressed model for anyone to try, and says it plans to apply the same approach to other popular AI models next - with an eventual goal of shrinking even the largest AI systems available today down to a size that can run outside a data center entirely.

Bottom Line

Nothing here is confirmed, and nothing is coming to your iPhone this year. But the direction is clear: the next real leap in phone-based AI may not come from bigger chips or bigger data centers - it may come from figuring out how to make AI small enough to disappear quietly into the device you already carry in your pocket.