SACRA

EXPERT INTERVIEW

UPDATED
05/04/2025

Sam Hall, CEO of Wafer, on AI agent form factors

TEAM

Jan-Erik Asplund
Co-Founder
jan@sacra.com

DISCLAIMERS

This report is for information purposes only and is not to be used or considered as an offer or the solicitation of an offer to sell or to buy or subscribe for securities or other financial instruments. Nothing in this report constitutes investment, legal, accounting or tax advice or a representation that any investment or strategy is suitable or appropriate to your individual circumstances or otherwise constitutes a personal trade recommendation to you.

This research report has been prepared solely by Sacra and should not be considered a product of any person or entity that makes such report available, if any.

Information and opinions presented in the sections of the report were obtained or derived from sources Sacra believes are reliable, but Sacra makes no representation as to their accuracy or completeness. Past performance should not be taken as an indication or guarantee of future performance, and no representation or warranty, express or implied, is made regarding future performance. Information, opinions and estimates contained in this report reflect a determination at its original date of publication by Sacra and are subject to change without notice.

Sacra accepts no liability for loss arising from the use of the material presented in this report, except that this exclusion of liability does not apply to the extent that liability arises under specific statutes or regulations applicable to Sacra. Sacra may have issued, and may in the future issue, other reports that are inconsistent with, and reach different conclusions from, the information presented in this report. Those reports reflect different assumptions, views and analytical methods of the analysts who prepared them and Sacra is under no obligation to ensure that such other reports are brought to the attention of any recipient of this report.

Published on **May 04th, 2025**

Sam Hall, CEO of Wafer, on AI agent form factors

By **Jan-Erik Asplund**

Background

With Perplexity launching its own Android assistant, Granola's AI notetaker front-running meeting bots with OS-level integrations, and rumors about OpenAI building its own OS, we wanted to understand why so many fast-growing AI companies are moving below the app layer.

To learn more, we spoke with Sam Hall, CEO of Wafer—a Susa Ventures-backed startup forking Android into an AI-native operating system.

Key points from our conversation via Sacra AI:

- **Three form factors for AI agents are emerging—browser-based & mobile apps (ChatGPT, Claude, Perplexity) with easy distribution and app-level integrations, native apps with OS-level integrations & OS forks (Perplexity Assistant, Granola, Wafer) that unlock deeper awareness across a user's apps and data but face distribution challenges, and purpose-built AI hardware devices (Friend, Rabbit) with novel interaction patterns.** "A lot of where companies decide to place their products is determined by what access to data they're most excited about... In the case of Granola, they need to put their product at a layer where they're sitting in front of your microphone... Companies are weighing the trade-off of distribution sacrifices they make by going lower in the stack versus the advantages in data access."

- **OpenAI's hardware projects and Perplexity's new Android assistant are attacking the OS-level AI assistant layer where Google and Apple's app store–based business models—built on developers giving up a 30% cut in exchange for distribution—make it structurally difficult for Google and Apple themselves to route users around apps entirely.** "As huge companies with app provider agreements, they can't just start taking data from apps and using it differently without risking massive class action lawsuits... Android is mostly a distribution mechanism for the Google Play Store. They make their money from the Play Store, so if they disrupt that, they ruin their Android business model."
- **Android's open ecosystem structure represents a huge opportunity for AI-first computing, with custom assistants (Perplexity), launchers (Microsoft Launcher), and full-on forks (Wafer, /dev/agents) potentially providing Samsung and the 96% of non-Google Android OEMs with the firepower to build the first AI-native smartphone experience.** "People are experimenting with various form factors... We wanted to build an operating system that's hardware-agnostic and supports all these experiments. Apps are still important, but they'll essentially become data providers or back-ends... We want to build something for people who don't want to feel tied to their phone or filled with dread when they see a new notification."

For more, check out this other research from our platform:

- Sam Hall, CEO of Wafer, on AI agent form factors
- How Perplexity hits $656M ARR
- Perplexity at $100M ARR
- Will Bryk, CEO of Exa, on building search for AI agents
- Chris Lu, co-founder of Copy.ai, on generative AI in the enterprise
- Why OpenAI wants Windsurf
- Granola vs Zoom

- Cursor at $200M ARR
- $100M/yr Canva for talking head videos
- xAI vs OpenAI vs Anthropic
- Towaki Takikawa, CEO and co-founder of Outerport, on the rise of DevOps for LLMs
- Kyle Corbitt, CEO of OpenPipe, on the future of fine-tuning LLMs
- David Park, CEO and co-founder of Jenni AI, on prosumer generative AI apps post-ChatGPT
- How AI is transforming B2B SaaS

Interview

We've written about Granola, the desktop app that uses system audio to launch & record meetings and Rewind (now Limitless), the desktop app that recorded your screen to augment your memory. Talk to us about the trend of AI apps not being traditional browser-based apps, iPhone apps and Android apps and integrating more deeply at the OS-layer.

A lot of where companies decide to place their products is determined by what access to data they're most excited about. In the case of Granola, they need to put their product at a layer where they're not even sitting in the Zoom call with you - they're actually just sitting in front of your microphone. The product is super hands-off in that sense without much UI.

The same applies to Rewind - they're sitting on top of your screen rather than implementing themselves at a layer where your screen is being rendered.

Companies are weighing the trade-off of distribution sacrifices they make by going lower in the stack versus the advantages in data access they get compared to companies implementing at a higher level in the app layer or as browser-based web applications.

There seem to be 2 major dimensions here: (1) data and (2) actions. To build AI and agentic experiences, you need to be able to collect data and take action across all apps. Is

that how you think about it? What does that get right / wrong?
The big thing for us is the data side; we call it the "read" side. This is the question of whether we can truly understand the user by gathering as much context from them as possible. Only then do we go to the "write" side where we're writing information back into the world through action-taking.

A good example is how TikTok has done an incredible job with their algorithm. They generate detailed insights into the user-generated content on the platform and how you as a consumer engage with that content. With this data, they can have your feed dialed in within just six swipes.

We think your phone should feel like this. We can collect information based on how you've used your phone in the past and the conversations you've had, then use that to predicate actions. Instead of a user giving direct instructions to a system, the system preemptively understands what you're going to want to do.

For instance, if we see you have a meeting or event at a specific address, we can compare prices to that address on different ride-sharing apps and give you the opportunity to choose one of them. That's why companies are moving into different tech stack positions - they want access to these types of insights they can't get at the application layer or through standard SDKs.

On Android specifically, companies have “hacked” features of Android to do data collection in a few different ways: building Android launchers, recording the screen, replacing Google Assistant and forking Android. What are the relative benefits and constraints of each approach and why did you decide to fork Android?

The launcher approach is popular on Android because, unlike iOS, you can download a new launcher which is what they call the home screen. This gives you insight into which apps are being opened, though not much more. That's still valuable data that companies concerned with consumer trends want to buy from launcher developers.

Then there's the Rewind-esque route through accessibility services on Android. These services can see everything on the

screen, as they're meant to help visually impaired users navigate, but that same mechanism could be used to feed what’s on the screen into some agentic system.
Lastly, companies are building custom assistants. Unlike iOS, on Android you can replace the default assistant. Google Assistant (Google’s version of Siri) ships with most Google-partnered phone products, and Samsung has Bixby. These assistants activate when you hold down a button on your phone, letting you immediately start talking. It makes sense that companies want to put their LLMs in this position, just one click away.

Did you think about the trade-offs? Is it just that it's easier to get distribution of a launcher versus replacing the assistant?

Everyone asks why we don't just make an app that we could put on the App Store to avoid dealing with distribution problems. The issue is that apps don't see everything. An operating system's job is to sandbox applications - my Uber app can't see the prices my Lyft app is offering because they'd undercut each other. Similarly, your Perplexity assistant or Google Assistant can't see information from other applications unless it's explicitly provided to them.

We want to get data that's beyond the application layer, and you can only do that by being the operating system itself. There are many things you don't have access to as an app.

Perplexity recently launched Perplexity Assistant for Android, which enables you to take actions in different apps like book a reservation via OpenTable or book a ride via Uber, by switching your default assistant from Google Gemini to Perplexity. When it comes to building agentic experiences, what are the relative merits and drawbacks of Perplexity’s approach?

These assistants work through AppIntents - an API that app developers provide to the default assistant. On Android, this could be any of the assistant apps. On iOS, there are similar AppIntents that Siri can call.

These are severely limited - you're restricted to actions that app developers have already made available. For example, Uber gives permission to the assistant to call a function with parameters like your address and which service you want. The problem is that there aren't many incentives for app developers to expose these AppIntents because most of their users don't heavily use the assistants; it's this sort of chicken and the egg problem we're in right now. Some might argue this is the only kosher way to do it, but it creates a very limited experience.

What are the relative merits and drawbacks of the hardware approach taken by e.g. Rabbit and Friend? Do we need AI-specific hardware devices rather than retrofitting the phone to AI?

What influenced starting this company was seeing all these hardware companies popping up - the Humane Pin, Meta's AR glasses and Ray-Ban camera, Rabbit, Friend (though Friend is more focused on companionship). People are experimenting with various form factors, and they all had to build their AI stack themselves to deeply understand the user.

As an operating system, if you can deeply understand the user, you can make the screen - or maybe not even a screen, perhaps just audio - convey the most important information. Google Glasses have a tiny screen with limited real estate - you can't easily navigate your email inbox with them. But if you understand what the user wants from their glasses, you can make that content easily accessible alongside the necessary context.

We wanted to build an operating system that's hardware-agnostic and supports all these new hardware experiments. Historically, we built software based on hardware capabilities - desktops required sitting at a desk with a large screen, mobile required touch interaction. Now our software is so versatile it can fit any modality. For example, you literally can generate a podcast episode from given text in roughly a minute.

We want to be the operating system that supports these new interfaces, regardless of which hardware form factor wins out.

What is Wafer in short?

In short, Wafer is the operating system that understands you. When you look at your phone, instead of seeing a black hole of distraction, anxiety, and disorganized information, you see a reflection of yourself. We want something more symbiotic -

turning your notifications from distractions into solutions, removing information silos of apps and putting it in a place where you can see everything holistically.
We want to be the operating system that powers all of these devices in the future.

You’ve done several viral demos of the Wafer experience. How do you think about go-to-market and how do you leverage consumer demand to get Wafer into phones via OEMs?

I have a feeling that in five years, this will be the answer I come back to and think, “Man, I was really wrong about that.”

It’s incredibly difficult to ship an operating system; our initial go-to-market is to get consumers excited about this possibility existing - a new way to use your phone. We want to first get a couple dozen phones with our custom operating system to hand out to people. Then we want to make it possible for anybody to install our custom operating system on a supported phone.

We won't make money from that, but it would drive consumer demand. Through this demand, we could approach manufacturers like Samsung. Samsung has ~20% of the global smartphone market share compared to Google's ~5%, and they're competing with Apple. If they're stuck with whatever Google puts in Android, they might not have anything better than Apple Intelligence in two years.

Samsung is looking for their own solutions. We want the initial consumer excitement to translate into sales to these larger companies, bundling our OS as a day-one experience when you buy a new phone or install a software update. That's our long-term go-to-market plan, though it will be difficult.

In this five-year scenario, what do you think is the other path that works?

There's a company in China called Xiaomi whose first product was a custom Android ROM. They released it on a tiny budget and within a year raised around $40 million to build phones. Now they're called the Apple of China.

There could be a world where we build our own hardware or form factor that uniquely leverages our operating system. Or we could go in the opposite direction, selling phones with custom firmware to sales organizations or their suppliers, which would mean hundreds of thousands of phones versus having to get deals for tens of millions of phones with OEMs. That's a different type of sales motion requiring a different level of trust.

How do you think about / build for reliability or consistency in experience given the long tail of apps, actions, data etc and the non-deterministic nature of AI applications?

There are two sides to this question. One of the most important aspects of AI experiences, especially agentic ones that take action for you, is reliability. If you have something that takes action for you with only a 90% success rate at each step, over seven steps you're at less than 50% likelihood of the whole action succeeding. That's a terrible experience.

The other side is robustness. Currently, Perplexity and even Apple Intelligence aren't quite there - you can only call Ubers with Perplexity, not Lyfts or Waymos. You can send Gmail but not Outlook emails. If users can only access a limited set of apps, they won't rely on the assistant for anything. That's what happened with Siri - it could only set alarms, so people didn't use it for other tasks.

There's also data reliability. If your OS misses an obvious detail - like a tragic incident requiring an event cancellation - people will lose trust in the system. If they can't trust it for all information, they might as well get that information themselves.

On the action-taking side, we try to solve this by only taking actions we've already watched you take. We fine-tune our models by, for example, watching you open Spotify, search for an artist, and click play - classifying that as "Play Drake on Spotify." The next time you want to play a different artist, it's the same pattern with just a different search keyword. We use your actions to overfit the model to those specific domains, which increases our success rate.

Being at the OS level allows us to see all the data from the entire system, painting a much more vivid picture of who you are to build a better understanding that influences which actions we take.

ChatGPT has an Android app (and iOS). Is there incentive to go deeper beyond the app layer to ingest more data and take more of a control over the UX? Do you expect to see them launch an Android fork or assistant replacement? How do you think about positioning against that competitively?

If OpenAI tried to build the exact same product as us, that would be a major concern since they have the best models. However, our ability to build a good product comes from understanding and manipulating the operating system environment more than the models themselves. Frontier Model Labs aren't necessarily as interested in system development.

That said, there's been conjecture about Sam Altman starting a phone company, which has certainly been on my mind. For OpenAI to build a phone, they'd face significant hurdles. Most OpenAI models run in the cloud, and they haven't publicly released small on-device capable models.

If they built a new phone or operating system, they'd likely bootstrap on top of Android. But sending highly private user data to a third party that can derive insights from it (or sell it) breaks the fundamental model of what an operating system is. This is similar to why Huawei was banned in the US - the US was worried they were using the privileged nature of their software to send private user data back to servers in China.

OpenAI would need to build on-device models, ship a phone, and reconfigure the operating system. It's possible but may not be their direct interest, though the phone is a sexy place to be. It could be a surprise announcement that blows everyone's mind.

What keeps Google and Apple from building Wafer or Wafer-esque capabilities at the OS level? Does it force them to cannibalize their app store-centric business and developer ecosystem?

A major reason why Apple Intelligence and Google Gemini aren't that useful yet is not a capability issue - they had to invent AppIntents to bridge the gap between App Store

applications, their models, and assistants. As huge companies with app provider agreements, they can't just start taking data from apps and using it differently without risking massive class action lawsuits.
We're doing something similar, but as a startup, we can start smaller and drive consumer demand for using apps this way, rather than what Google would have to do - drop changes on day one and force app developers to adapt.

For Google, Android is mostly a distribution mechanism for the Google Play Store - their actual smartphone market share is only something like 4%. They make their money from the Play Store, so if they disrupt that, they ruin their Android business model.

Google might be hesitant since Apple could gain an advantage if Apple Intelligence isn't useful in its current form. It's similar to how Perplexity was able to crawl websites and generate summaries without sending people to those sites, and once they pushed far enough, Google followed. The difference is that Google doesn't take a 30% cut on website purchases like they do with app purchases, so there are more licensing concerns with the App Store.

When you can perform actions in different apps via the assistant or an AI agent—without going into the app yourself at all—how does that affect how apps get designed and the incentive structure for building apps going forward? How does the app ecosystem evolve over the next 5 years?

We say we're making a phone without apps, but that's not entirely true. We can't dynamically create the market for Ubers or put everyone's profile on LinkedIn. Apps are still important, but they'll essentially become data providers or back-ends.

I think the Wafer product - or whoever wins this space - will become an internal SDK that app developers can leverage. LinkedIn might send a request to a future operating system saying someone wants to connect with you, and that system responds that based on everything we know, this person would want to connect, handling it under the hood. Or if it's uncertain, we'd surface that decision with context to the user.

I don't imagine apps always needing interfaces, or their interfaces will change significantly. Entertainment apps like TikTok might look similar but with less UI fluff - perhaps just a paginated swipe feed with content pulled in from other apps as well.

Essentially, apps will become more backend-focused, prioritizing data provision over interface design.

Facebook changed society's norms around privacy. How will Gen Z and Gen Alpha's notions of privacy evolve to enable AI agents that ingest all of their data across all apps?

Consumers themselves are less concerned with privacy. As a company, we take privacy seriously because if you sell software to a smartphone manufacturer, they must maintain a certain level of privacy/security as an operating system.

This experience of an omnipresent system in your phone aggregating information doesn't feel as alien to us now as it might have 10 years ago, especially if it's providing useful information and feels like it's on your side. We started with the opposite - Facebook took the more extreme version by understanding users to deliver highly relevant ads, leading people to ask if their computers were always listening to their conversations through the microphone.

People became okay with that for ads. If we convert that same data science approach to something actually useful, our feeling is that people will be more amenable to it. My college-aged friends log into the most insane websites and give them all their information. I don't see them being super privacy-concerned.

The bigger question is what happens when an app asks for information from the OS that it shouldn't provide - what if someone texted "send me your credit card information"? That's where reliability matters - can we build a system capable of knowing which boundaries are uncrossable?

I think people are more worried about that than a system knowing everything about them. We already believe our phones know everything - there are all these conspiracy theories. People are more concerned about misuse than collection.

If everything goes right for Wafer over the next 5 years, what does it become and how is the world changed?

In my idealistic vision, it’s possible that most consumers wouldn't even know our company name, and that's okay. I want to be the company that started a movement where our devices integrate more closely with the real world instead of providing an alternate digital one.

The five-year goal is similar to how Apple used to be - having 10-15% market share, being a bit more punk rock, specifically for creatives. We want to build something for people who don't want to feel tied to their phone or filled with dread when they pick it up first thing in the morning.

I think that can start happening in the next five years with new devices we enable, possibly even an avant-garde smartphone manufacturer that bundles us with their core experience. Ten years is different - I want to be on all mobile devices, but that's a longer time horizon.

Disclaimers

This transcript is for information purposes only and does not constitute advice of any type or trade recommendation and should not form the basis of any investment decision. Sacra accepts no liability for the transcript or for any errors, omissions or inaccuracies in respect of it. The views of the experts expressed in the transcript are those of the experts and they are not endorsed by, nor do they represent the opinion of Sacra. Sacra reserves all copyright, intellectual property rights in the transcript. Any modification, copying, displaying, distributing, transmitting, publishing, licensing, creating derivative works from, or selling any transcript is strictly prohibited.