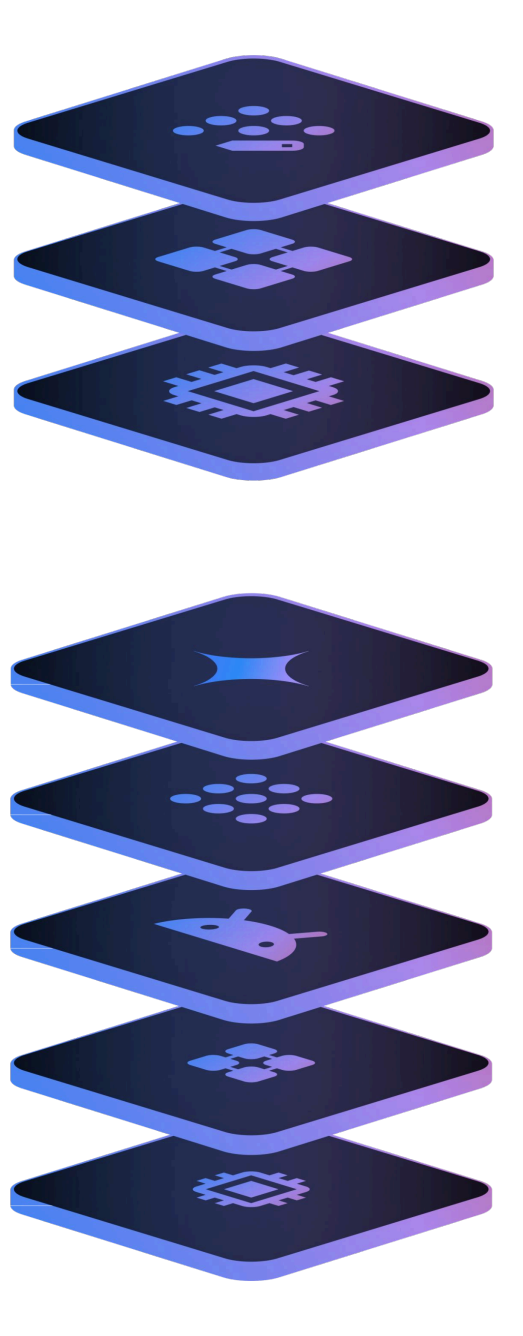
Android

The OS designed for AI

March 2025

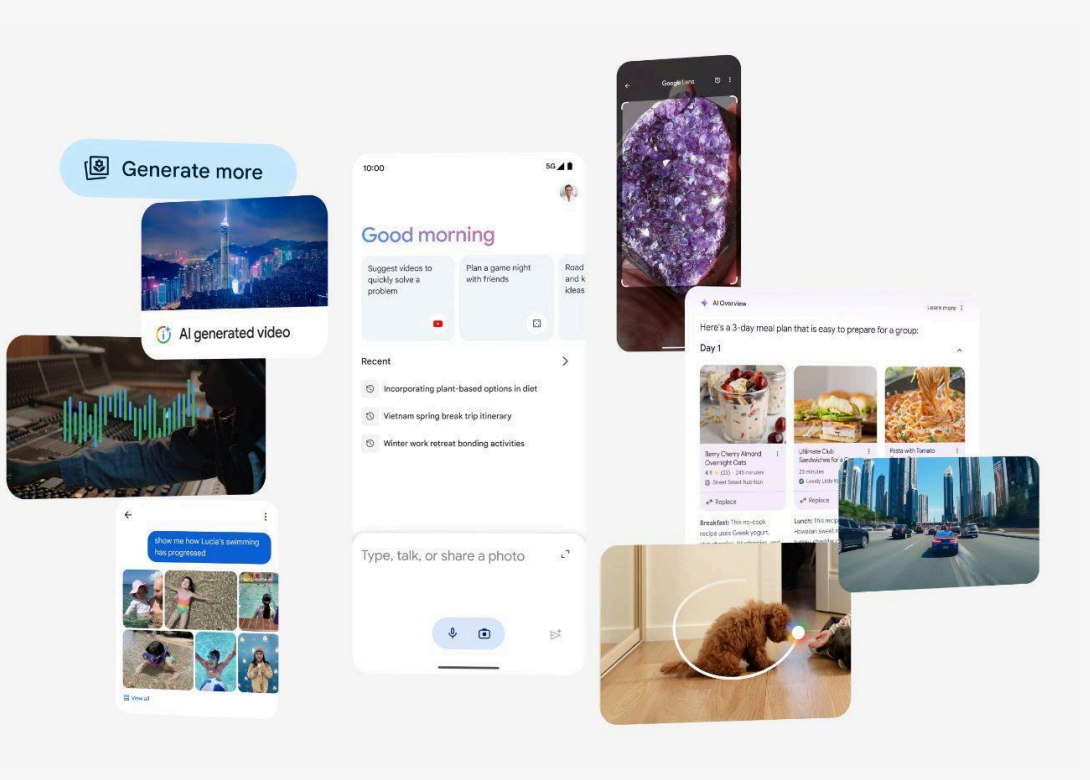


Android, built with AI at the core

The biggest question of 2024 was "what can AI do to help users?" In 2025, the question is shifting to: "*how* and *where* can it deliver the most help?"

That's what Android has been focused on — ensuring the most popular operating system in the world delivers the benefits of AI in the most relevant, helpful, and secure way. Android is putting Google's best Gemini models directly into people's hands, with a holistic evolution of the entirety of Google's unique tech stack.

We know that the true value of AI doesn't emerge from a single app. Or a chip. Or a model. It comes to life when every element in a tech stack is fully integrated — from data centers in the cloud, to entire app ecosystems, to developers being able to design with AI tools, to the OS, all the way to the silicon in your phone. Only this deep level of integration can enable AI to be truly helpful, while keeping personal information secure and protected. And Google can deliver it.



This all comes together in Android, the OS that can uniquely integrate and activate AI across the platform for billions of users and devices. **Android infuses Gemini directly into its core, to make the entire OS layer a powerful, personal, and secure AI experience for each user**.

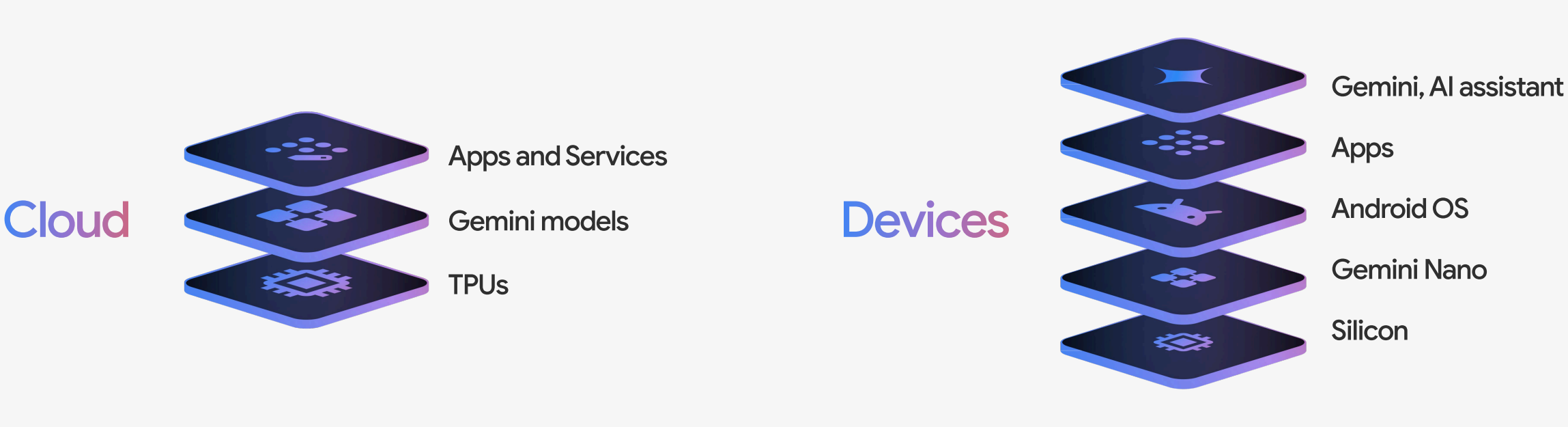
And now, the years we've spent building the world's most advanced cloud-based tech stack are complemented by our mobile compute stack that fits right in your hand.

Welcome to the future.

A tech stack for the Gemini Era

Our Gemini strategy isn't just a software play. It's a full-stack approach to AI that scales across the ecosystem.

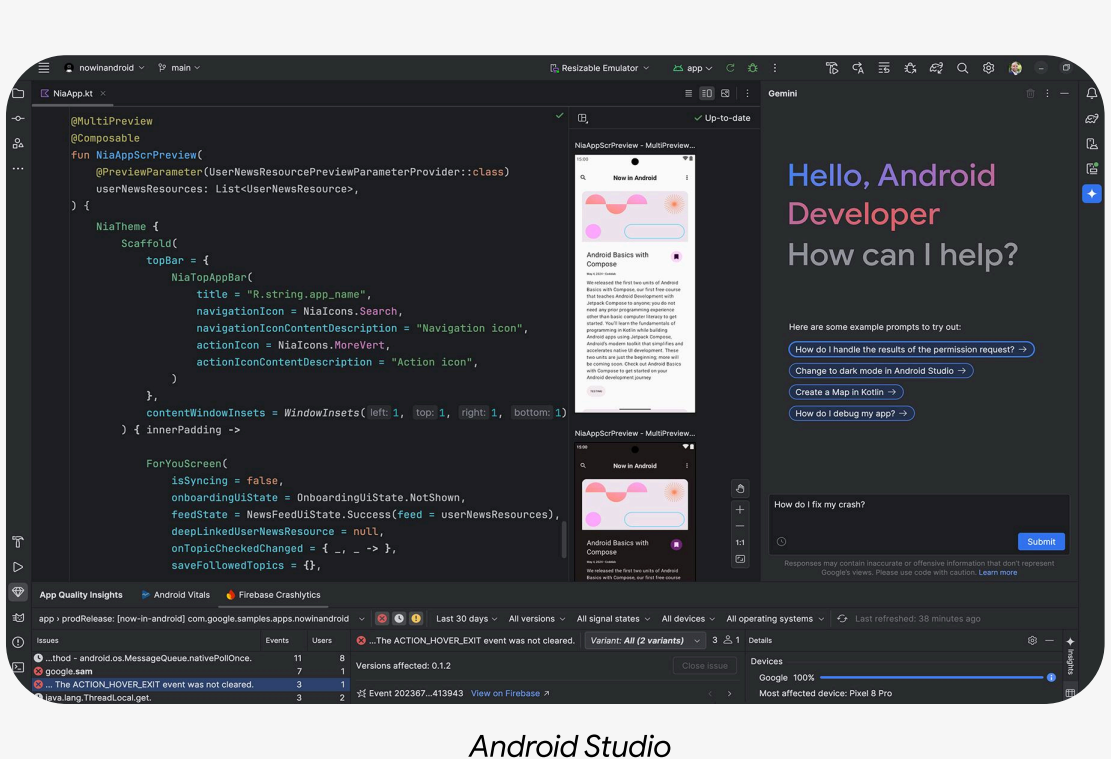
And it's built on 25 years of investment in world-class technical infrastructure that's paying dividends in the Gemini Era. Our long-term strategy to integrate AI across the entire tech stack gives us the unique ability to innovate at every layer.



We've invested deeply over the past decade in state-of-the-art data center infrastructure, including our own sixth-generation Tensor Processing Unit (TPU) silicon that's specifically designed for AI at scale. We have our family of Gemini models, including Gemini 2.0, our newest, most capable model for the agentic era, which is used by leading Android OEMs and partners to create breakthrough AI experiences. And we have Gemini Nano, our powerful on-device AI model that delivers fast, secure AI capabilities like helping you write messages with Smart Reply, and getting a live narration of what's on your screen with Talkback. With Gemini Nano, sensitive data never leaves your device, and AI-powered features even work when you're offline or in airplane mode.

These Gemini models also optimize our incredible collection of Google apps and services that billions of people turn to for help. Like helping you find information more quickly with AI overviews in Search, and helping you create presentations, synthesize data, and get more done with Gemini in Docs, Sheets, and Slides.

Lastly, custom versions of Gemini models are deeply integrated into Android Studio for coding, dramatically boosting coding and developing workflows for app developers and partners such as PayPal. With a full suite of Gemini-powered developer tools and Gemini SDKs, first-party and third-party app developers can use AI to be more productive, and also integrate AI into their own apps and services.



Android Studio

Every layer of the stack is informed by the latest research from Google DeepMind. And every layer is tuned specifically for our Gemini models, and built responsibly by design using Google's Privacy Policy, AI Principles, and industry-leading data security protocols.

It all means we can check three critical boxes that no other mobile ecosystem can.

We can deliver breakthrough AI experiences with Android's **open-ecosystem approach** to billions of devices from dozens of device OEMs.

We can deliver a rich, personalized assistive experience that connects **real-time information to helpful app actions on your device**. And bring the smarts of AI across Google apps and third-party apps.

As Gemini continues to evolve at a rapid pace, we're able to **quickly prototype, test, and deploy new features and model upgrades** across the ecosystem. Improvements at any layer of the stack make the entire stack better, so Android devices continue to get more powerful and more helpful over time.

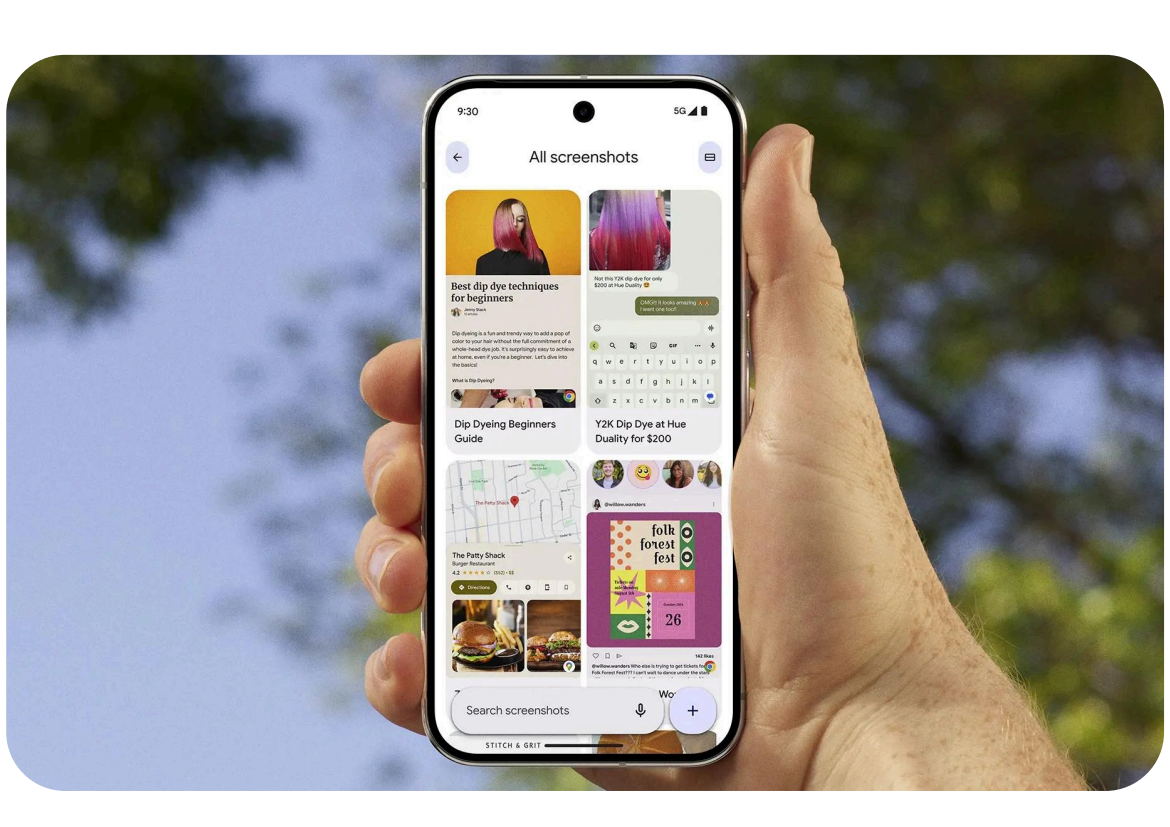
Rapid change can only happen when the stack is fully integrated. It leads to fewer dependencies and development bottlenecks, quicker iterations, and downstream benefits to every OEM and app developer in the Android ecosystem.

There's another critical layer of the stack, and that's at the silicon level. We have deep experience from developing our own TPUs and our Google Tensor system on a chip (SOC) for Google Pixel devices. We then take our learnings to the ecosystem, collaborating closely with partners to optimize their own solutions to deliver advanced, on-device Gemini experiences and a seamless integration with cloud experiences, ensuring hardware devices are more performant.

Simply put, every time you pick up your Android device, the apps and services you use on your phone bring AI capabilities that are optimized for that device.

CASE STUDY

Google Pixel 9: Utilizing Gemini Nano for custom app experiences



Pixel Screenshots app

Google's full-stack approach gives OEMs and developers a robust platform for developing their own custom AI-powered apps.

Like the new Pixel Screenshots app on Pixel 9. Pixel Screenshots lets you take a quick screenshot of anything you might want to remember, and files it away for easy access later. It's a fast and easy way to recall QR codes, shopping sites, book recommendations, recipes, or the perfect gift idea.

The app is built to take advantage of Android's on-device Gemini Nano model, with new optimization schemes that boosted inference speeds by roughly 3X over Pixel 8. And it takes advantage of Gemini Nano's new multimodal capabilities, to understand the text and images in a screenshot, organize and categorize the information, and make it fully searchable for users.

While Pixel Screenshots is exclusive to Pixel 9, the AI models it's built on are available to the entire Android ecosystem of OEMs and developers.

CASE STUDY

OPPO: Leveraging Gemini models to build AI phones

OPPO builds AI into the core of an astounding number of devices and experiences, with around 700 million monthly active users globally. The secret to OPPO's successful rollout of new AI experiences lies in its ability to deploy AI across its full development cycle. Our integrated, full-stack approach makes it possible.

OPPO/OnePlus was the first OEM to integrate hardware power consumption feedback into Google's neural architecture search (NAS), to optimize models based on chip characteristics. That customization reduced the power consumption of object detection algorithms by 27% and the latency by 40%, resulting in a more efficient and hardware-friendly AI model in a shorter development time.

OPPO also taps the Gemini models in Google Cloud to create unique AI capabilities for OPPO/OnePlus phones. Like the AI Recording Summary feature, which automatically summarizes meeting content based on an audio recording. And AI Writer, which can generate content in whatever style you want, help polish your text, and generate social media posts based on photos you share.

OPPO has also deeply integrated Android features like Circle to Search, and the new Gemini app, to deliver the latest AI experiences to OPPO users all over the world.

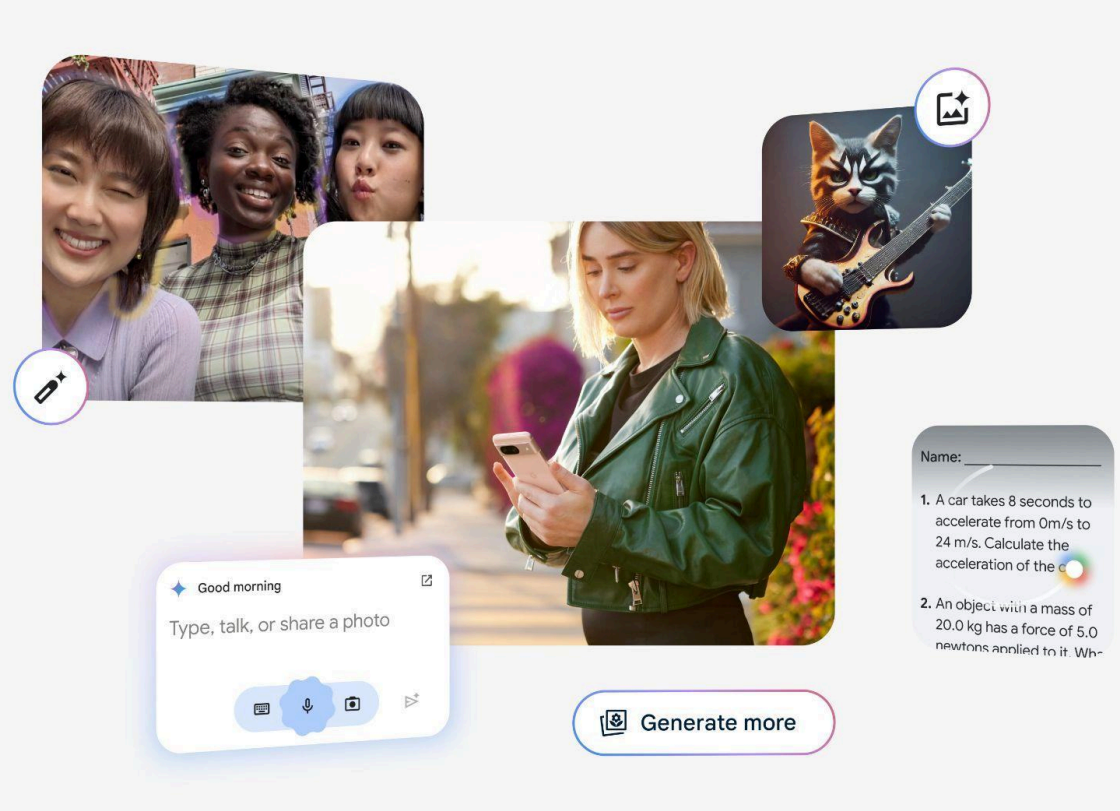
700 million
active users globally

27%
reduction in power consumption of object detection algorithms

40%
reduction in latency

Next-level AI assistance, for everyday help

The quickest, most seamless way to experience mobile AI and the strength of our full-stack integration is with Gemini, your new assistant on Android. It connects and integrates a broad range of capabilities across apps and services into an easy-to-use assistant for everyone, right on their personal devices.



And as Google DeepMind continues to make breakthroughs in AI research, we're able to refine and deliver the latest features and upgrades through Gemini on your Android phone. For instance, advances in multimodality with Gemini 2.0 — like native image and audio output — and native tool use, are leading to new AI agents that bring us closer to our vision of a universal assistant.

With our full-stack integration on Android, we can optimize how we deliver Gemini 2.0 breakthroughs for the mobile form factor. Over the last few months, we've been testing Project Astra; soon we're bringing screen sharing and live video input to the Gemini app on Android devices with Gemini Advanced. The multimodal capabilities let Android users get contextual help with what they see, whether it's on their screens or through their cameras. **The next-gen assistant on Android isn't just able to answer questions, it's also able to get things done for you.**

With contextual understanding, you can ask Gemini to help with something on your screen and fulfill a task. And thanks to Google's full-stack integration, we can connect that with the amazing capabilities of the Android software ecosystem, so you can take multi step actions in popular apps like Calendar, Maps, Spotify, YouTube, Google Home, and many more. Our developer tool, app functions, enables these actions to scale across all your favorite apps and services, which also includes integrations with our partner apps, like those from Samsung, OPPO, OnePlus, and Xiaomi.

USE CASE

Gemini can connect to multiple apps at once

Gemini brings more advanced reasoning capabilities to Android.
For complex tasks, Gemini can break down a single request into multiple steps, and can act as an agent on your behalf, working across your apps to complete your request.

For example, you may be on the hook to plan the next group get-together and the pressure is rising. You can ask Gemini about the best suggestions for the group of your size and send a message in just a single command. Gemini can quickly connect the dots and find places on Google Maps that fit all your criteria. Plus, with your permission, Gemini can then draft a text message to your friend.

This can be so helpful throughout your day. If you've come across a PDF with a lot of dates for an upcoming trip, you can even ask Gemini to list the key events and add them to your calendar.

USE CASE

Ask Gemini about PDFs, YouTube videos, and more

We get so much content on our phones, but it can be hard to read and parse through it all. With Gemini, you can now quickly find information or get help over a PDF or a YouTube video you're diving into.

For example, maybe you have a lease agreement up as you're heading to check out that place one last time. You may want to know whether the building accepts pets or what the security deposit policies are.

And when you want to retrieve a specific piece of information from a YouTube tutorial or lecture, Gemini works in the same contextual way. Let's say you're watching a DIY tutorial on YouTube. You can now ask Gemini about a specific part of the video, or go back a few steps, all hands-free.

It couldn't be simpler to access the help of Gemini on Android. Just long-press the power button when you need help, and the assistant will be there for you, with a full understanding of your screen context.

Gemini is delivering this level of assistance across the Android ecosystem at scale. It already supports **45 languages** in more than **200 countries and territories** on billions of devices. Samsung users, as well as people who use hundreds of other phone models from various manufacturers, can access Gemini's power every day.

And talking to Gemini on Android feels more natural than ever. Gemini's ability to navigate the "umms" and "ahhs" of natural speech stems from fine tuning advanced language models, trained on vast audio and text datasets, enabling it to discern context and filter irrelevant sounds. You can also interrupt it naturally, and choose a voice that best fits your style.

People are already turning to Gemini for a wide range of help, like getting answers quickly, learning practical skills, organizing thoughts, unlocking new ideas, and much more. **We're constantly making Gemini a more helpful mobile assistant**, and we're excited to see what more people will do with it.

Coming soon to Gemini Live: Screen and video share capabilities from Project Astra

A true assistant should be able to understand the world around you, and the sights, sounds, and information you come across. That's the idea behind Project Astra, a research prototype that uses advanced multimodal processing to interact with you in real time, understand what you're trying to do, and help with everyday tasks.

The capabilities Project Astra bring go beyond simple voice commands and static input types (images, PDFs).

Later this month, we're bringing some of the groundbreaking capabilities from Project Astra directly to users in the Gemini app. We'll upgrade Gemini Live so you can ask about anything you see in real time, whether it's on your phone or something right in front of you. You'll be able to have a conversation with Gemini while simultaneously live streaming your camera to explore new places, collaborate visually, or get instant help troubleshooting items in your environment.

You can also share your screen with Gemini Live for in-the-moment guidance on app usage, shopping, and skill development, all dynamic to how you move across your phone with every scroll, app switched, and more.

It's a big step forward for the kind of help you can get from Gemini. These capabilities are launching in March 2025, and will be available first on Android devices with Gemini Advanced.

And as Gemini continues to get more deeply integrated, it's also getting more *agentic*, using advanced reasoning, multimodal understanding of the world, and contextual inference to take action for you. This helps unlock a truly helpful assistant that can perform complex tasks, with your permission, in the background without additional input or guidance. You just ask once, and Gemini handles the details.

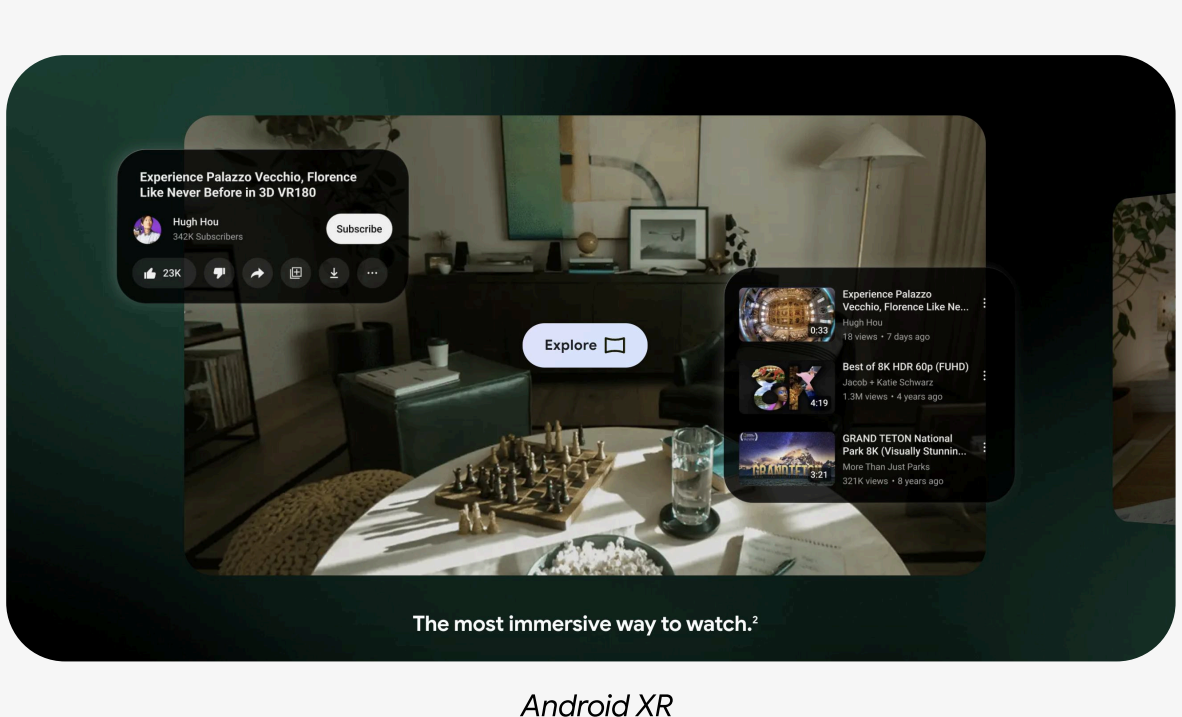
Last month, we introduced Deep Research agentic capabilities to Android users. Instead of spending hours researching a topic on your phone, pulling out your laptop for more complex information, and then painstakingly organizing notes, you can ask Gemini to do the research for you. Gemini gathers the info you need from across the web, and creates a report just for you, easily viewable on mobile. Just a couple of minutes can save you hours.

Gemini's agentic capabilities will help your assistant be even more proactive to help you get things done, so you can focus on the things you care about.

Gemini in a multi-device world

One of the strengths of Android's ecosystem is the platform's extension into other form factors, interaction models, and hardware capabilities. And Gemini's model advances like multimodality are directly pushing hardware innovations forward.

For instance, new extended reality devices like headsets and glasses are designed to take full advantage of Gemini's multimodal capabilities, and our new **Android XR** operating system. With a multimodal AI assistant built right in, your glasses or headset can understand your intent and the world around you, helping you get things done in entirely new ways.



Android XR

On headsets, you can have conversations about what you're seeing, or even control your device with your voice. Gemini can understand your intent, helping you plan and research topics, and guiding you through tasks. With glasses, we put the power of Gemini one tap away, providing helpful information right when you need it — like directions, translations, or message summaries — without reaching for your phone. It's all within your line of sight, or directly in your ear.

And of course, we're taking the same full-stack approach to Android XR that we've developed for Android. So the OS, the application layer, and the hardware are working hand in hand with Gemini models in the cloud and on the device. Developers can lean on the same Android toolkit and standards to create breakthrough XR experiences at a global scale.

The future of mobile computing has never been more exciting.

Android is designed to be an open, unified platform that works across devices. For users, that means more choice of devices and services they love, and in the future, a seamless, shared context for continuous assistance no matter where they are, all optimized for the device they use. For developers, it's a unified platform. For our ecosystem partners, it's a single, integrated solution that will drive rapid innovation at Android scale.

Stay tuned for more updates on what is possible with Android with Gemini in the coming months.