Video title: Microsoft Build opening keynote | KEY01

Video URL: https://www.youtube.com/watch?v=_r9em36n2b0&t=28s

Video language: English - Default

I think that our industry has to have the common vision. It was a time that connected us to incredible things. My name for this vision is information at your fingertips. And three decades later, we find ourselves in a new era, one where access to information becomes access to expertise. From the farm to the lab, from the boardroom to the classroom, this new generation of AI is for everyone, everywhere. Now anyone can save time with a personal assistant. About 50% of time and that's time that I can use to do other innovative things. Anyone can access a personal tutor to learn new skills. I need that. Inside. I think the technology has the potential to completely reimagine the way every single student learns in the world. This is a new way to analyze with a personal coach. We're going to be able to have not only productivity gains, but insights surplus near real time. Generating I can learn from the data to help improve the farmer productivity. AI is unlocking creativity for us all. Descriptions are so detailed in my imagination I can paint the artwork. Adri Yiga Eagan Lesson plan. Create a teacher. Your freedom even. With expertise at your fingertips, you can build what matters. Welcome to the age of AI transformation. Good morning, Good morning. It's fantastic to be back here at Microsoft Build. Welcome to everyone here and joining us on the web. You know, developer conferences are always most exciting, most fun when there's these fundamental changes that you can sense in the air. You know, I've mocked all my adult life by coming to PDC's and Bills, you know, for the last three decades. I still remember, you know, distinctly the first time Win 32 was discussed. I guess it was 91.net Azure, right? These are moments that I've marked my life with, and it just feels like we are yet again at a moment like that. It's just that the scale, the scope is so much deeper, so much broader this time around, right? Every layer of this tech stack is changing, you know, from everything from the power draw and the cooling layer of the data center to the NP user. The edge are being shaped by these new workloads, right? These distributed synchronous data, parallel workloads are reshaping every layer of the tech stack. But you if you think about even going all the way back to the beginning of modern computing, say 70 years ago, there have been two real dreams we've had. 1st is can computers understand us instead of us having to understand computers? And second, in a world where we have these ever increasing information, that of people, places and things, right? So as you digitize more artifacts on people, places and things and you have more information, can computers help us reason, plan, and act more effectively on all that information? Those are the two dreams that we've had for the last 70 plus years and here we are. I think that we have real breakthroughs on both fronts. The core underlying force. One of the questions I always ask myself is like, OK, this is great. This is like maybe the golden age of systems. What's really driving it? I always come back to these scaling laws, just like Moore's Law, you know, helped drive the information revolution. These scaling laws of DNN's are really along with the model architecture. We interesting ways to use data, generate data are really driving this intelligence revolution. Ohh, you could say

Moore's Law was probably, you know, more stable in the sense that it was scaling at maybe 15 months, 18 months. We now have these things that are scaling every six months or doubling every, you know, six months. You know. What we have, though, with the effect of these scaling laws, is a new natural user interface that's multimodal, that means supports tax, speech, images, video as input and output. We have memory that retains important contexts, recalls both our personal knowledge and data across our apps and devices. We have new reasoning and planning capabilities that helps us understand very complex context and complete complex tasks, right, while reducing the cognitive load on us. But what stands out for me as I look back at this past year is how you all as developers have taken all of these capabilities and applied them, quite frankly, to change the world around us, Right. I'll always remember this moment in January 2023 when I met a rural Indian farmer who was able to reason over some government farm subsidies that he had heard about on television using GPD 3/5 and his voice. It was remarkable, right? For me. It just brought home the power of all of this because the a frontier model developed in the West Coast to the United States just a few months earlier was used by a developer in India to directly improve the life of a rural Indian farmer. The rate of diffusion is unlike anything I've seen in my professional career and it's just increasing. In fact, earlier this month I was in Southeast Asia, I was in Thailand where I met a developer and I was having a great round table. And he was talking to me about how he's using 5, three and GPD 4 and he was using 53 to just optimize all of the things that he was doing with RAG. I mean, this is crazy. I mean, this is unbelievable. It had just launched a few weeks earlier and I was there in Thailand in Bangkok, listening to a developer talk about this technology as a real expert on it. So it's just great to see the democratization force, but we love to talk about, but to witness it, it's just been something and this is quite frankly the impact of why we are in this industry and it's what gives us, I would say that deep meaning in our work. So I want to start though with a very big thank you to everyone of you who is really going about bringing about this impact to the world. Thank you all so very much. You know, when I think about what progress we have made even since last time we were here at Build, we built really three platforms. The 1st is Microsoft Copilot, which is your everyday companion. It puts knowledge and expertise at your fingertips, helps you act on it. And we built the Copilot stack so that you can build your AI applications and solutions and experiences. And just yesterday we introduced a new category of Copilot Plus PCs, the fastest AI first PC's ever built. All three of these things are exciting platforms, but I want to start with Copilot Plus PCs. You know, we are exposing AI as a first class namespace for Windows. This week we are introducing the Windows Copilot runtime to make Windows the best platform for you to be able to build your AI applications. Yeah, you know what Win 32 was to graphical user interface. We believe the Windows Copilot runtime will free for AI. It starts with our Windows Copilot library, a collection of these ready to use local APIs that help you integrate into your new experiences all of the AI capabilities that we shared yesterday. Now this includes no code integrations for Studio effects, things like creative filters, teleprompter, voice focus and much more. But of course if you want to access these models itself right, you can directly call them through APIs. We have 40 plus models available out-of-the-box, including five Silica on US, member of our small language family model which we can specifically which we specifically designed to run locally on your NPC's on Copilot Plus PC's, bringing that lightning fast local inference to the device, you know. The other thing is the Copilot library also makes it easy for you to incorporate RAG inside of your applications on the with on device data, right. It gives you the right tools to build a vector store within your app. It enables you to do that semantic search that you saw with Recall. But now you can, in your own application, construct these prompts using local data for RAG applications. No, I'm so thrilled to announce as well today that we will be natively supporting PITARCH and new Web NN framework through Windows Direct ML. You know native Pie Torch support means thousands of OSS models will just work out-of-the-box on Windows, making it easy for you to get started. In fact, with web and non web developers finally have a web native machine learning framework that gives them direct access to both GPU's and NPC's. In fact, last night I was playing with it, turning it on an edge and seeing that the web and sample road running. It's just so cool to see it, you know? Now use even the NPU. Both Pie Tart, Pie Torch and Web Non are available in developer preview today. Let's take a look and these are just one of the many announcements today. We're introducing more than 50 plus new products and partnerships to create up, you know new opportunity for you. We've always been a platform company and our goal is to build the most complete end to end stack from infrastructure to data to tooling to the application extensibility so that you can apply the power of this technology to build your own applications. And so today I want to highlight our top news for this event across every layer of this copilot stack. So let's dive right in with infrastructure. You know we have the most complete scalable AI infrastructure that meets your needs in this AI era. We're building Azure as the world's computer. We have the most comprehensive global infrastructure with more than 60 plus data center regions more than any other cloud provider. Over the past year, we've expanded our data center regions and AI capacity from Japan to Mexico, from Spain to Wisconsin. We're making our best in class AI infrastructure available everywhere and we're doing this with the focus on delivering our cloud services sustainability. In fact, we're on track to meet our goal to have our data centers powered by 100% renewable energy by next year. Yeah, you know we're we're optimizing power and efficiency across every layer of the stack from the data center to the network. Our latest latest data center designs are purpose built for these AI workloads so that we can effectively and responsibly use every MW of power to drive down the cost of AI and the power draw. And we are incorporating advanced data center cooling techniques to fit the thermal profile of the workloads and that match it to the environment and the location where it operates at the silicon layer. We are dynamically able to map workloads to the best accelerated AI hardware so that we have the best performance. And our custom IO hardware and server designs allow us to provide dramatically faster networking, remote storage and local storage throughput. You know, this end to end approach is really helping us get to the unprecedented scale. In fact, last November we announced the most powerful AI simple computer in the cloud for training using just actually a very small fraction of our cloud infrastructure. And over the past six months, we've added thirty times that supercomputing power to Azure. Yeah, it's crazy to see the scale. And of course, we're not just scaling, training our fleets, we're scaling our influence fleet around the world, quadrupling the number of countries where Azure AI services are available today. And it's great to see that at the heart, at the heart of our AI infrastructure, are the world's most advanced AI accelerators, right? We offer the most complete selection of AI accelerators, including from NVIDIA and AMD, as well as our own Azure Maya, all dynamically optimized for the workloads. That means whether you're using Microsoft Copilot or building your own Copilot apps, we ensure that you get the best accelerator performance at the best cost. For example, you know, you see this in what has happened with GPT 4, right? It's 12X cheaper and 6X faster since it launched. And that's, you know, the type of progress. You can, you know, continue to see how, you know, to continue to see the progress as we evolve the system architecture. It all starts though with this very deep, deep partnership with NVIDIA which spans the entirety of the Copilot stack across both all of their hardware innovation as well as their system software innovation. Together we offer Azure confidential computing on GPU's to to be really help you protect sensitive data around the AI models end to end. We're bringing in fact the latest H2 hundreds to Azure later this year and we will be among the first cloud providers to offer Nvidia's Blackwell GPUs, B1 hundreds as well as GB 200 configurations. And we are continuing to work with them to train and optimize both large language models like GPT 4 O as well as small language models like the Five, Three family. Now beyond the hardware, we are bringing invidious key enterprise platform offerings to our cloud like the Omniverse cloud and DGX cloud to Azure with deep integration with even the broader Microsoft cloud. For example, Invidia recently announced that they're DGX cloud integrates natively with Microsoft Fabric. That means you can train those models using DGX Cloud with the full access to fabric data and omnivores. API's will be available first on Azure for developers to build their industrial AI solutions. We're also working with NVIDIA Nim industry specific developer services and making them fantastic on Azure. So a lot of exciting work with NVIDIA. Now coming to AMD, I am really excited to share that we are the first cloud to deliver general availability of VMS based on a MD MI300XI accelerator. It's a big milestone for both AMD and Microsoft. We've been working at it for a while, and it's great to see that today as we speak, it offers the best price performance on GPT 4 inference, and we'll continue to move forward with Azumaya. In fact, our first clusters are alive and soon if you're using Copilot or one of the Azure Open AI services, some of your prompts will be served using Maya hardware. Now beyond AI, our end to end Systems optimization also makes cloud native apps and the development of cloud native apps better, right? Six months ago is when we announced our first general purpose ARM based compute processor, Microsoft Cobalt. And today I am really excited to announce the public preview of Cobalt based VMS. You know Cobalt is being used for video processing and permissions management and Microsoft 365 helping power billions of conversations on services like Microsoft Teams already. And we're delivering that same arms based performance and efficiencies to many customers in fact including Elastic Mongo, Seamen, Snowflake and Teradata. In our most recent benchmark, data and tasks are Cobalt 100 VMS delivered up to 40%. Better performance than any other generally available ARM based VM. So we are very, very excited about Cobalt getting into the market. Now let's move up the stack to the Foundation models. Look, with the Azure AI, we offer the broadest selection of Frontier and open source models, including LLNS and Slams. So you can choose the model that makes the most sense for your unique needs and your application needs. In fact, more than 50,000 organizations use Azure AI today. Yeah. It's great momentum and it all starts though with our most strategic and most important partnership with Open AI. Just last week Open AI announced GPT 4 O their latest multimodal model which was trained on Azure. It's an absolute breakthrough. It has text, audio, image and video as input and output. It can respond and just have a a human like conversation that's fast and fluid. It can even be interrupted mid sentence. Our GPD Forro is also the top performing model on benchmarks across a variety of modalities in it's always going to get you to the best answer. It has state-of-the-art performance at understanding the information you actually provide in your prompt, because that's really what it matters. What matters? What opening? I demoed last week. I mean, that was just pretty magical stuff. And yesterday we showed how Copilot will leverage GPT 4 O in the video. I'm just gonna play notice that you can share your screen or session as a prompt in Copilot and have Copilot help you in whatever it is that you're doing. Let's roll the video. Hey, Copilot, how's it going? Hey, I'm doing great and it looks like you're about to dive into some Minecraft ready to build, explore, and maybe dodge a few moms. Yeah, I'm trying to figure. Out how to make a sword. My son plays this game all the time and I have no idea what I'm doing. No worries, I'm here to help you impress your son. I see you've got a crafting table in front of you, which is a great start. To craft a sword, you'll need some materials. Can you open your inventory? Just press E. On your keyboard? Yeah. Do I have what I need? Let's see. You've got some sticks perfect for the. Sword's handle. But it looks like you're missing the material for the blade. You can use wood. Stone, iron, gold, or diamond. Let's start by gathering some wood or stone. Shall we head out and find some resources? Yeah, that sounds great. Oh my God. What's that? Oh no, it's the zombie run. You need to get away from it fast. Either Sprint away or quickly build a tower of blocks straight up to get out of its reach if you can find a shelter or. Dig into the side of a. Hill, do it now. Just make sure it can't reach. You. All right. I think I might be safe now. That was a close one. Great job finding shelter. It's really cool to see that. You know the magic here is that Copilot is your personal assistant that's helping you with any task, helping you stay, quite frankly in the flow in the moment, whether it's a game or a document or a you know, line of code. And as Opening Eye innovates, our promise is that we will bring all that innovation to Azure too. In fact, the same day that Open A announced GPD 4 O, we made the model available for testing on Azure Open AI service, and today it I'm excited to say that it's generally available on Azure AI. You know what this means though is that now we can have these groundbreaking apps that all of you can build using this capability. One of the coolest things is that now any app, any website can essentially be turned into a full multimodal, full duplex conversational canvas. Let's take a look. All right, let me ask the agent here. Hey Seth, how's it going? What can I help you with today? I am fantastic. I'm sorry I'm in such a hurry. I'm going to go on an overnight camping hike or camping. All this. It's going to be dark, cold. Should I? Should I use these shoes? It's the mountains. It's probably gonna be cold. Hey, no worries, let me take a look. Ah, those are the. Truck Star Hiking Sandals. While they're fantastic for summer hikes and keeping your feet cool, they might not be the best choice. For college, I'm sorry to interrupt, you pick the best thing that I should use and can you just add it to my card, just one product. No worries. Let's go with the trek. Ready, hulking boots. They're durable, provide great ankle support, and most importantly, they'll keep your feet warm in the cold mountains. I'll add them to your cart now. Ready for checkout? That amazing. It's OK. Maybe Internet Campamento? They not a Seth. It's true that they to campamento Imentet Caliente Conesus Potas Buena suerte divierte Los Montanas while Seth was you know Seth was in a hurry but there you go, GPD 4 or kept up with him, It's incredible. I really want to thank the open AI team for their partnership and and really they're responsible approach to innovation helping our industry move forward. Sam will be here in fact joining Kevin in a little bit to talk a lot more about what's coming right because that's the exciting stuff. How do you all sample what comes next. We also brought bringing lots and lots of other models as well, from Cohere and Databricks and Desi, Madam Mistral, Snowflake, all to Azure AI. We want to support the broadest set of models from every country, every language. I'm excited to announce. In fact, we're bringing models from Cohere, G42, NTT Data, Next Slide, as well as many more as models of services because that's the way you can easily get to managed AI models. And we all love open source too. In fact, two years ago at Build, we were the first to partner with Hugging Face, making it simple for you to access the leading open source library with state-of-the-art language models. We are Azure AI and today I'm really excited to announce that we're expanding our partnership bringing more models from Hugging face with text generation influence with text embedding inference directly into Azure AI Studio. And and we're not stopping there, we are adding not just large language models, but we're also leading the small language revolution. So small language model revolution, you know our five three family of SLM's are the most capable and most cost effective. They outperform models of the same size with the next size up even across a variety of language reasoning, coding as well as math benchmarks. If you think about it by performance to parameter count ratio, it's truly best in class. And today we are adding new models to the five three family to add even more flexibility across that quality cost curve. We're introducing 5 three vision of 4.2 billion parameter multimodal model with language and vision capabilities. It can be used to reason over real world images or generate insights and answer questions about images, as you can see right here. Yeah, and we're also making a 7 billion parameter 5 three small and a 14 billion parameter 5 three medium models available. With Phi you can build apps that span the web, your Android, iOS, Windows, and the Edge. They can take advantage of local hardware when available and fall back on the cloud when not, simplifying really all of what we as developers have to do to support multiple platforms using one AI model. Now it's just awesome to see how many developers are already using Phi 3 to, you know, do incredible things. From Amity Solutions, the Thai company that I mentioned earlier, the ITC, which has been built a copilot for Indian farmers to ask questions about their crops. Epic in Healthcare, which is now using Phi to summarize complex patient histories more quickly and efficiently, and one of the very, very cool use cases in education. Today I'm very thrilled to announce a new partnership with Khan Academy. We'll be working together to use 5 Three to make math tutoring more accessible, and I'm also excited to share that they'll be making Conomy Ago, their AI assistant, free to all US teachers. Let's roll the video you're already. You forgot already felt like I was in a place in my teaching career where I felt like I was kind of losing my sparkle and I would just feel really defeated when I looked out on the classroom and I would see students that just didn't look engaged. Teachers have an incredibly hard job, and what we think we can do is leverage technology to take some of the stuff off of their plate to really, actually humanize the classroom. By some miracle we became Conmigo pilots full. With new advances in generative AI, we launched Conmigo. The point is to be that personalized tutor for every student and to be a teaching assistant for every teacher. Started to build these more robust lessons and I started to see my students engage. We're working with Microsoft on these Phi models that are specifically tuned for math tutoring. If we can make a small language model like Phi work really well in that use case, then we would like to kind of shift the traffic to Fiji. In those particular scenarios, using their small language model, the cost is a lot lower. We're really excited that Conmigo, and especially in the partnership with Microsoft, being able to give these teacher tools for free to US teachers, is going to make a dramatic impact in US education. I think we're going to make them the innovators, the questioners, isn't that really just why you wake up every morning, right? Because that's our for our future, our next generation. And to me, that's everything. You know, I'm super excited to see the impact of this All will have and what Khan Academy will do. And Sal's going to in fact join Kevin soon to share more. And I'm really thankful to teach for teachers like Melissa and everything that they do. Thank you very much. You know, of course it's about more than just models, it's about the tools you need to build these experiences. With Azure AI Studio, we provide an end to end tooling solution to develop and safeguard the Copilot apps you build. We also provide tooling and guidance to evaluate your AI models and applications for performance and quality, which is around one of the most important tasks as you can imagine with all these models. And I'm excited to announce that Azure AI Studio now is generally available. It's an end to end development environment to build, train and fine tune AI models and do so responsibly. It includes rebuilt in support what is perhaps the most important feature which is in this age of AI, which is AI safety. Our Azure AI Studio includes the state-of-the-art safety tooling you know to everything from detecting hallucinations in model outputs, risk and safety monitoring. It helps understand which inputs and outputs are triggering content filters, prompts, Shields by the way to detect and block these prompt injection attacks. And so today we are adding new capabilities including custom categories so that you can create these unique filters for prompts and completions with rapid deployment options, which I think is super important as you deploy these models into the real world. If in an emerging threat is you know appears beyond Azure AI Studio, we recognize that there are advanced applications where you need much more customization of these models were very specific use cases and today I'm really excited to announce it Azure AI custom models will come giving you the ability to train a custom model that's unique to your domain, to your data that's perhaps proprietary. That's same builders and data scientists who have been working with Open AI brought all the Phi advances to you will work you know with all of you to be able to build out these custom models. The output will be domain specific. It'll be multitask and multimodal, best in class as defined by benchmarks, including perhaps even specific language proficiency that may be required. Now let's just roll up the stack to data. Ultimately, in order to train, fine tune, ground your models, you need your data to be in its best shape. And to do so we are building out the full data estate right from operational stores to analytics in Azure. We've also added AI capabilities to all of our operational stores, whether it's Cosmos DB or SQL or PostgreSQL. At the core though is this of the Intelligent Data platform is Microsoft Fabric. We now have over 11,000 customers including leaders in every industry or using Fabric. It's fantastic to see the progress with Fabric. You get everything you need in a single integrated SAS platform. It's deeply integrated at the most fundamental level with compute and storage being unified. You're experience is unified, governance is unified and more importantly the business model is unified. And what's also great about Fabric is that it works with data anywhere, right? Not just on Azure, but it can be on a WS or on GCP or even on your on premise data center. And today we are taking the next step. We're introducing real time intelligence and Fabric. Customers today have more and more of this real time data coming from your IT systems, your telemetry systems. In fact, cloud applications themselves are generating lots of data. But with Fabric anyone can unlock actionable insights across all of your data estate. Let's take a look. Introducing real time intelligence in Microsoft Fabric, an end to end solution empowering you to get instant, actionable insights on streaming data. At its heart lies a central place. Discover, manage and consume event data across your entire organization with a rich, governed experience. Get started quickly by bringing in data from Microsoft sources and across clouds with a variety of out-of-the-box connectors. Route the relevant data to the right destination and fabric using a simple drag and drop experience. Explore insights on petabytes of streaming data with just a few clicks. Elevate your analysis by harnessing the intelligence of Copilot in Microsoft fabric using simple natural language. Make efficient business decisions in the moment with real time actionable insights and respond to changing landscapes proactively. Allow users to monitor the data they care about, detect changing patterns and set alerts or actions that drive business. Value all your data, all your teams, all in one place. This is Microsoft Fabric. And we're making it even easier to design, build and interoperate with Fabric with your own applications, right? Right. And in fact, we're building out a new app platform with Fabric workload development kit, so that people like Ezra, for example, having, you know, who have integrated their spatial analytics with Fabric so that customers can generate insights from their own location data using Azura's rich tools and libraries. Right on Fabric, Right. This is just exciting to see. That's the first time you know where the analytics stack is really a first class app platform as well. And beyond Fabric we're integrating the power of AI across the entirety of the data stack. There's no question that RAG is core to any AI powered application, especially in the enterprise today. And Azure AI Search makes it possible to run RAG at any scale delivering very highly accurate responses using the state-of-the-art retrieval systems. In fact, ChatGPT supports for, you know, GPS. Their assistance API are all powered by Azure AI Search today, and we've built in one Lake integration. Azure AI Search will automatically index your unstructured data too, and it's also integrated into Azure AI Studio to support bringing your own embedding model for example. And so it's pretty incredible to see Azure Search grow over the last year into that very core developer service. Now let's go up to developer tools. Nearly 50 years after our founding as a developer tools company, here we are once again redefining software development, right GitHub Copilot was the first, I would say, hit product of this generative AI age, and it's the most widely adopted AI developer tools. 1.8 million Subs across 50,000 organizations are using it and get up. Get up, Copilot. We are empowering every developer on the planet to be able to access programming languages and programming knowledge in their own native language. Think about that. Any person can start programming, whether it's in Hindi or Brazilian Portuguese, and they bring back the joy of coding to their native language. And with Copilot workspace, staying in your floor has never been easier. We are an order of magnitude closer to a world where any person can go from idea to code in an instant are you start with an issue. It creates A spec based on its deep understanding of your code base. It then creates a plan which you can execute to generate the code across the full repo, that is multiple files. At every point in this process, from the issue to spec to plan to code, you are in control. You can edit it and that's really what is fundamentally a new way of building software and we are looking forward to making it much more broadly available in the coming months. And today, we're taking one more big leap forward. You know, we are bridging the broader developer tools and services ecosystem with Copilot for the first time with really thrilled to be announcing GitHub Copilot extensions. Now you can customize get up Copilot with capabilities from third party services, whether it's Darker Sentry and many, many more. And of course we have a new extension for Azure to GitHub, Copilot for Azure. You can instantly deploy to Azure to get information about your Azure resources just using natural language and what Copilot did for coding we are now doing for infra and OPS. To show you all this in action, here is Neil from our GitHub team. Naya, take it away. Thanks Satya. GitHub Copilot gives you suggestions in your favorite editor, like here where I'm writing unit tests. Copilot is great at meeting you where you're at, regardless of the language you're most comfortable with. So let's ask for something simple, like how to write a prime number test in Java, but let's converse in Spanish using my voice. Como verificar still numeral dado Essen numero Primo in Java. Look at that Gracias Copilot. Copilot is great at turning natural language into code and back again, but what about beyond the code? With the new GitHub Copilot extensions, you can now bring the context from your connected systems to you. So now I can ask Azure where my app is deployed. I could ask what my available Azure resources are, or I could diagnose issues with my environment. And this isn't just for Azure. As Satya announced, any developer can now create extensions for GitHub Copilot and that includes any tool in your stock, including your in house tools keeping you in the flow across your entire day. Actually, 75% of the developers day is spent outside of coding, gathering requirements, writing specifications and creating plans. Let's show how GitHub Copilot can help with that live on stage for the first time. So typically my day starts by looking at GitHub issues. Looks like we want to support a rich text input for our product description. Let's open Workspace and get some help with that. Copilot interprets the intent of the issue to see what's required, and it then looks across the entire code base and it proposes what changes should be made. This specification is fully editable and the whole process is iterative, but actually this looks pretty good. Copilot can now help us build a plan on how to implement this change. All right, that's a great start, but we must not forget about our documentation. So let's edit the plan and have Copilot update our README, and then we can even get Copilot's help in starting to implement the code for us. Now this was just a simple example, but in a large enterprise code base, there are 10s of thousands of files and dozens of stakeholders involved. And that means meetings. So many meetings. Workspace helps you focus on what you need to change. And by the way, as a developer I'm always in control. I can see exactly what changes Copilot is proposing, and I can even get a live preview. All right, let's test out the input. All right, this looks great. So I can go back and I can edit my code in VS Code, or I can submit these changes as a pull request to share with my team. GitHub Copilot, Copilot Extensions and Copilot Workspace help you stay focused on solving problems and keeping you and the flow back to you. Satya. Thank you so much, Neil. I'll tell you, GitHub Copilot and everything that that ecosystem is doing is just bringing back a lot of fun and a lot of joy back to coding. And and really the thing about staying in that flow is I think what we all have dreamt for and dreamt about and it's coming back. That brings us to the very top of the stack, Microsoft Copilot. We built Copilot so that you have the ability to tap into world's knowledge as well as the knowledge inside of your organization and act on it. Now Copilot has had a remarkable impact. It's democratizing expertise across organizations. It's having a real cascading effect, right. In fact, it reminds me like of the very beginning of the PC era where work, the work artifact and the workflow, we're all changing. And it's just, you know, really having broad enterprise business process impact. It's lowering. I always say this. It's lowering both the floor and raising the ceiling at the same time for anything any one of us can do. Since No2 business processes are the same. With Copilot Studio, you now can extend Copilot to be able to customize it for your business processes and workflows. Today we are introducing Copilot connectors in Copilot Studio so you can ground Copilot with data from across the graph from power platform, fabric, Dataverse as well as you now have all the third party connectors for SAS applications from Adobe, Atlassian, ServiceNow, Snowflake and many, many more. You know, this makes the process of grounding Copilot in first and third party line of business data just a wizard like experience, enabling you to quickly incorporate your own organizational knowledge and data. We're also extending Copilot beyond a personal assistant to become a team assistant. I'm thrilled today to announce Team Copilot. You know you'll be able to invoke A-Team Copilot wherever you collaborate in Teams, right? It can be in Teams, it can be in Loop, it can be in Planner and many, many, many other places. Someone think about it right It can be your Meeting, Facilitator, or when you're in Teams, creating agendas, tracking time, taking notes for you or a collaborator, writing chats, surfacing the most important information, tracking action items, addressing unresolved issues. And it can even be your project manager, ensuring that every project that you're working on as a team is running smoothly. These capabilities will all come to you all and be available in preview later this year, and we're not stopping there. With Copilot Studio anyone can build copilots that have agent capabilities and work on your behalf and independently and proactively orchestrate task for you. Now simply provide your Copilot a job description or choose from one of our pre made templates and occupied with the necessary knowledge and actions and Copilot will work in the background and act asynchronously our for you right? That's I think one of the key things that's going to really change in the next year where you're going to have copilots plus agents. With this async behavior, you can delegate authority to copilots to automate long running business processes. Copilot can even ask for help when it encounters situations that it does not know much about and it can't handle. And to show you all of this, let's roll the video. Redefine business processes with Copilot Studio. Create copilots that act as agents working independently for you. Simply describe what you want your Copilot to do. Easily configure your Copilot with the details and needs like instructions, triggers, knowledge and actions. Quickly test your copilot before you deploy and seamlessly published across multiple channels. Watch it use memory for context with reason over user input and manage long running tasks. Copilot can learn from feedback to improve. And you're always in control. Put Copilot to work for you. With Copilot Studio, you know all around this stack is perhaps one of the most important things that we at Microsoft are doing, which is wrapping it with robust security. You know, security underlies our approach with Copilot, Copilot plus PC's, Copilot stack, we're committed to our secure future initiative. You can see, you'll see us make rapid progress across each of the six pillars of SFI you know and the core design principles right, which is secure by design, secure by default and secure operations. You'll hear about throughout this conference. In fact, a lot more in Scott's keynote tomorrow, how it underlies everything that we build and everything that we do. So coming to the close, I want to sort of, you know, there are many announcements that you will hear about at Build. But I wanna go back to I think the core of what I think why we chose to be in this industry and why we come to work every day as developers, which is the mission ultimately of empowering every person and every organization. At the end of the day, it's not about innovation that is only useful for a few. It's about really being able to empower that everyone. And it comes down to you all as developers and builders of this new world. For us, it's never, never about celebrating tech for tech sake. It's about celebrating what we can do with technology to create magical experiences that make a real difference in our countries, in our companies, in our communities. Already this new generation of AI is having an incredible impact thanks to all of you, the passion you bring and the hard work you put in. And I want to leave you with this one unbelievable example of how you're all building a more accessible world, which means a lot to me using our platform and tools. Thank you all so very much. Enjoy the rest of Build. Audio description is something that enables me to be able to watch a program or a film and get as much out of it as everybody else who is cited. A white car drives down a road, hands on a steering wheel. I see artists collective good. I think everyone should be able to have access to art. Audio description really helps me get the full experience. A portrait of a group of 17th century civic guardsmen in Amsterdam. The challenge, though, is that there are limited amounts of audio descriptions being incorporated across media and entertainment. Tech and I have the potential to bring the blind and low vision community into the fold. So at WPP, we really care passionately about opening up access to content to people in a way that they want to consume it. The tool that I've made is an application that allows you to upload videos and on the other end with GT4 with vision and as your AI services you get your video back with spoken narrations over the top. Kitchen scene with cats and Hellman's mayonnaise. Lier descriptions Cheaper and faster. Our goal is to be able to offer this product as a service rule of advertisement campaigns. There are so many artworks interact museum there are almost 1,000,000. To. Describe. Ourselves, It would have taken hundreds. Of years. With AI we can do this in a matter of hours. The subject is a male with a reddish beard and mustache, visible brush strokes that add texture. And movement. The first time I heard audio descriptions, it's just brought me delight. It was this opportunity of Oh my gosh, I'm seeing through the power of AI we're. Able to do things only dreamt about until recently. When we strengthen our access to culture, we strengthen the culture itself, connecting our shared humanity. Please welcome Executive Vice President, Experiences Plus Devices, Rajesh Jha. Good morning. I'm Rajesh. You heard Satya talk about the Copilot Stack, the AI architecture of the future. Now I want to talk about bringing that AI stack to Microsoft products. 1st I'm going to focus on how we are expanding Copilot. Then Jeff Teper will talk more about extensibility and then finally Pavan Davuluri will share more on Copilot Plus PCs and the Windows ecosystem. Now diving into how Copilot is evolving. It was only a year ago at Build that we showed you the promise of AI and then we made Copilot from Microsoft 365 generally available in November. I want to spend a moment and talk about the journey that we've been on within Microsoft 365. We brought Copilot to the applications that hundreds of millions of people use every day. Given our users a powerful new way to interact with AI right in the flow of their work. Now we've seen great adoption across a range of customers and industries. Nearly 60% of the Fortune 500 now use Copilot and we've seen accelerated adoption across industries and geographies with companies like Amgen and BP, Cognizant, Moodys Nordisk and NVIDIA, Tech Mahindra and many others for choosing over 10,000 seats each. We've added over 150 Copilot capabilities since the start of just this year. In addition, we continue to integrate Copilot across more of our productivity apps and services from OneNote to Stream, to Forms to OneDrive and more. Now we also launched A standalone Copilot application. Whereas Copilot had been embedded in Office, Outlook, and Teams, the Copilot app now has all of Microsoft 365 embedded in it. And what makes the Copilot app unique and uniquely powerful is a grounding it has. First, our Copilot understands the web more than web grounding, the Copilot is crucially grounded in the users work contexts. Who do they work with? What do they work on? Their meetings, Their conversations, that documents, And that is a Microsoft Graph. That Microsoft Graph represents the user, their team, their permissions, their organization. It represents their context. Now, it's important to note that the Copilot app is no different from any of our other Commercial Services in terms of compliance and data handling. Microsoft has no eyes on Access and your data is not used to train the models. So let's dive in and see how all this comes together in the Copilot app. In the Web tab you can get answers to simple or complex questions grounded in live web data, all with commercial data protection. Now let's struggle over to the Work tab. Here, Copilot is grounded in Microsoft Graph meeting. It has access to your personalized work environment, the people you interact with, important files, teams, meetings, all your communications. By the way, this is not a demo account. This is my actual Microsoft account, using real data to show you how personal this experience really is. Of course you're going to see some of the options. I do want to keep my job now. Learning to ask the right question is key so you can get the most of the copilot to help you. There's a prompt library to give you suggestions just when you need them. Let me start by getting caught up on the latest my boss. You may have heard of him. Copilot and Routines. Organizational Structure so it knows Satya is my boss. It scans the latest emails and chats and files and in moments I have a detailed breakdown with updates that need my attention. Let's try one more here. I'm asking Copilot to propose a session title based on this very Keynote script, which is a document in SharePoint, and to create a list of topics on generative AI to discuss. In this session, Copilot has analyzed a document from my work environment and propose A suitable headline. But since Copilot is also grounded in web data, it is able to recommend topics source from the web for me to consider. Now switching gears, since its inception, Copilot has been a uniquely personal assistant as you've just seen. But we all work in teams, small and large organizations, intimate and global, and we want to do more to go beyond to empower people when they come together. And like Satya said, today we are announcing Team Copilot, the expansion of Copilot beyond a personal assistant. This will enable Copilot to serve and act on behalf of a team at department and our organization, not just an individual user. Copilot will act as a valuable team member, improving collaboration project management. Let's take a look. Teeth overland. This man copilot from Microsoft 365 from a personal assistant to a team assistant that becomes a valuable member of the team. It gets added to a meeting and group chat like any other participant, making it visible for everyone to interact with. Here, Copilot has been added to a meeting as the meeting facilitator, helping the team focus on having a productive conversation. Copilot shares the agenda in the meeting chat, and anyone can add topics without interrupting the active discussion. It actively tracks time, providing reminders when it's time to move on to the next topic. Copilot will also take notes during the discussion that anyone can edit or add to and flags follow up tasks for everyone to see. Let's switch over to Teams chat here. Copilot has been added as a collaborator to help streamline communications. As the conversation happens, Copilot maintains an up-to-date view of important information for everyone to see. This includes surfacing an unresolved issue for the team. Copilot will also help the team work together to resolve open issues. Here A-Team member asks Copilot for recommendations and grounds the prompt with the relevant document. The copilot instantly responds with suggestions for the team to consider. Once the team has agreed on the solution, Copilot automatically updates the notes pane, removing the open issue and detailing the decision made. With Team Copilot assisting in new ways, the entire team will be more productive and collaborative. The team called Ballot will be a valuable new member of any team and these initial capabilities will be available to our customers in preview later this year. Now in my conversation with the customer is one of the top questions I get is how do I translate productivity gains into transformative business results? And Simply put, the answer is moving to a reimagining of business processes using their own Copilot and agents and extending Microsoft Copilot. Now Jeff is going to walk you through our extensibility platform is going to enable you to do just that. But first, let's take a look at the Copilot architecture that really makes all of this possible. A Copilot, as we spoke, can recall and reason our up-to-date web knowledge. It is grounded in search. All Copilot experiences have this web scale now. In addition to being grounded in the web, it is also grounded in your data through the Microsoft Graph. And when the Copilot is in an application like Office or Teams or Edge, it also understands the application context. So for example, in PowerPoint it is able to draft a slide deck for you from a document. The Microsoft Copilot is architected to compose or inherit capabilities based on the user context. And now as a developer, you can build Copilot extensions at the data layer at the experience layer to further extend and customize the Copilot. And all of this is enterprise grade with tools for IT to manage and personalize for employees. A leading ISP's are already working with us building solutions that can extend the Microsoft Copilot, and I would like to share 2 examples from Service Now and Adobe for Service Now. As you know, ServiceNow helps organizations orchestrate and automate tasks and processes across their enterprise. Here we see Service now as Copilot Extension now assist responding to user prompts inside Copilot from Microsoft 365 with the exact same knowledge, functionality and user experience as it has today in Teams. Service now has included several custom 0 query default prompts to help users get started with the most common tasks without having to know how to crack the right text to start the conversation over to Adobe, who are working on bringing bringing Adobe Experience Cloud Workflows and Insights to Microsoft 365 and Copilot. With Adobe Express Copilot extension users stay in the flow of their work in a Word document and get start a workflow in Adobe Express that allows them to create social content, select and edit images and stage it for publishing. So as we close out this first chapter, I hope you're excited. Microsoft Copilot is already helping people save time, be more productive and creative. Team Copilot expands Copilot in meaningful ways and then there are great developer opportunities for you to extend Microsoft Copilot. Jeff Teper will join us now to share how easily you can build Copilot extensions. But first, I want to close my section with a video showcasing how Lumen is using Copilot for personal productivity to enhance their sales processes and connecting Copilot to their system using Copilot connectors. Let's roll the video. Lumen is a company that's going through a tremendous transformation, an evolution to move from telecommunications to technology. Our goal is to help our 3000 sellers and customer success professionals shift away from transactional selling and move to being customer obsessed. Copilot from Microsoft 365 helps us move faster. It's really hard for a seller to learn everything about all of their customers, and one way they can do this is using Copilot for sales. What takes sellers hours they can now do in minutes. Copilot and Salesforce, for us, that integration has been vital. People are coming to the table more prepared. They have more information. They're armed. They're ready to have customer conversations. Before Copilot, with these third party data connectors, the seller would have to go out to all of the disparate systems. Salesforce Gain site ServiceNow Copilot connectors helped to solve for the fragmentation that customers can feel which drives customer satisfaction. We have seen an increase in outbound calls by 40%. Customer outreach typically takes 4 hours. Now with Microsoft Copilot, with all of these third party data connectors, that is a 15 minute journey. If we can give our sellers back 4 hours a week, it's worth $50 million an annual revenue and that is math that matters. When you unlock people, teams, and culture, you unlock growth. We're at an intersection. It's time for disruption in this industry and we are here for. It. Well, as Rajesh shared, you will able be able to easily and securely use your applications and knowledge to build copilots that help your employees and organization be far more productive and grow your business. And you can now extend the Microsoft Copilot with your own Copilot with handoffs in all the Copilot experiences as well as in Microsoft Teams where you can reach hundreds of millions of users today for both the personal and group assistant scenarios that Rajesh outlined. And we are making building these copilots even easier, from a few clicks in SharePoint to more advanced customization and Copilot Studio, to full control of your models, your data, your applications, your actions, your experience in Visual Studio Code. So let's first look at what this means for end users. Copilot extensions run everywhere. That Copilot is the standalone experience across Microsoft Teams, and as we're showing here in the Microsoft 365 app on the right, you can easily browse your installed Copilot extensions, find new ones, or build your own, which I'll show coming up. And how Microsoft Copilot works is that reasons over the users prompt and maps it to the right extensions or you can explicitly @that extension like we're showing here. You're going to be able to drill into a deeper focus conversation with the extension like we're doing in this marketing example that has suggested prompts for quick actions and to just show the users the capability of your Copilot. This allows the Microsoft Copilot to have real time access to knowledge and applications in your environment. Here we're using a suggested prompt to ask about a key feature of a delivery drone. The Copilot extension you build is gonna come back with a Visual Adaptive card, bringing in all the information to avoid an unnecessary multiturn conversation so the user can just focus on getting their work done. And again, these Copilot extensions also run in teams in one-on-one and group chats and channels and in meetings, so you can reach all these users today. Alright, this is build. So let's get the building with our first custom Copilot extension. You're going to be able to do this with from, again, a few clicks in SharePoint to advanced customization, Copilot Studio to Visual Studio Code. We'll start in SharePoint, which is often the authoritative source of knowledge and content processes with advanced collaboration, workflow, and security, all of which Copilot honors. To make sure users only get access to information that they're they have permission to, I'm in the SharePoint site. I'm gonna go ahead and select a few documents, hit click the copilot and right there. That looks good. I'll go ahead and change the name of this the the delivery drone that looks fine. And just like that, I've created my first custom Copilot that you can use to extend the Microsoft Copilot. Let's go ahead and try this one out. How much does a delivery drone service cost? And we can see it comes back with a flat fee of $5 per order. All that looks pretty good. Secure Grounded. We're going to go ahead and share that with our team and up comes the standard sharing dialogue to honor their security in your organization. We're going to copy this link, go into Teams, paste it in the chat, we'll say try this out. We will go ahead and paste that. And just like that in seconds, we've created a secure, grounded custom copilot and shared it with our team and at Microsoft Teams. So we're very excited about letting anybody create these secure custom copilots. And this support will be available in SharePoint this summer. Sign up today for the preview. We're very excited about that. All right. Next. Yeah, next on the spectrum, we're gonna do some more advanced customization and Copilot Studio starting right from SharePoint, I can launch into Copilot Studio for my more advanced edits. And you can see all the information for the Copilot I just created is carried forward. So that's all there from SharePoint. And let me test this out by asking when the launch event is and it'll return back that it's on June 20th. But what I really want is that Copilot to do work for me, not just answer questions. And so for this I need to go into

Copilot Studio and start by adding additional data sources in the Knowledge tab. Here we can add websites and files and connect to over 1000 Copilot connectors. In this case, our account information is in two tables in Dataverse and so I'll go ahead and select them and we've got the information we need. Next we move to the Actions tab, where I need to see if somebody's already registered for the event, and if not, send them a personalized invitation. Our event registration is managed in an external system, so we've created a custom connector to go get it, and you can see us configure that we can Taylor the action and input and outputs however we need. And then the second step is to automate the sending of that personalized invitation. We've built a custom Power Automate flow, so we'll add that as a Copilot action as well. And there you go in pretty quickly. We've got a complex Copilot extension with content from SharePoint data from Dataverse to disparate actions, and we can go ahead and test this in Copilot Studio to see if Contos actually attending the launch event. It says they're not and the Copilot conversation helps you see why what's happening and map that all through. We can ask a follow-up question again that's routed to Dataverse about who the account manager is, and we can see that's Perry Lang, and then we can ask Copilot to go ahead and send the invitation. Now again what's happened here is the conversation was entirely generated for me identifying and chaining together the key knowledge from data, voice and the appropriate actions we added earlier with generative AI capability. Last, we're ready to publish this Copilot extension back to SharePoint, Microsoft Copilot and Teams. And from Teams you can search for it and the unified marketplace, I can add it to one-on-one or group chats or meetings etcetera. And again, I can see this in the Microsoft Copilot experience, that same Relic Cloud Copilot and continue the experience there. So we're very excited. This is a pattern that people are already doing today. Let me show you what one of our customers, Walters Cluer, who is a leader in information and software and solutions is doing to enhance their tax and accounting professionals workflow with Copilot. They're building a Copilot extension to enable their accountants to complete each step of their workflows, interacting with their back end system just using natural language with no context switching, Copilot takes actions on their behalf, saving time and cutting the process down from minutes to seconds and it's time to communicate back to their client. The extensions helped close the loop, drafts, an e-mail attach the estimates and ensures A seamless end to end productive experience. So we're excited about all of you building these kinds of custom copilots and copilot extensions. Copilot Studio is now generally available to build enterprise grade copilots. The new capability to publish Copilot extensions from your Copilot is in private preview. And as you heard Satya say earlier, Copilot connectors are now in public preview, and they make it even easier to connect your copilots to your business data, your apps and workflows. OK, last, we want to show you how to build a Copilot extension as a professional developer with full control of your models, your data, your actions, your experience in Visual Studio Code. So let me go ahead and do that. So what we've got here is Visual Studio Code with the Teams AI Toolkit installed and we've loaded up a template for building a custom copilot and using it as an extension. The first thing you see in the code is where I configure the model. You can use an off the shelf model here we're using one from Open AI. You can use a refined model or you can build your own completely tailored to what you need. If we go Scroll down the code a little bit farther, this is where RAG integration is and you can see in a few lines of code we've integrated our data, in this case Azura's vector search capabilities. You can use any other data source. The Azure AI Search capability is a great one to use are we keep scrolling down. We see the actions defined and registered, but here is where we go. Search for our product inventory and you can see we've inserted a break point so that we can follow this along. Because I want to show how easy it is to do end to end development here. So we'll bring up teams where we've installed that Copilot extension for the inventory, find information for the Chai Tea product, hit return and we're paused. Why? Because we've hit that breakpoint and so we can go back into Visual Studio, you can see sure enough it's fired and if I hover over the parameters you can see the product name Chai has been passed to it. So this is incredibly powerful right from within Visual Studio to do end to end debugging across the Microsoft apps, Copilot and your Copilot extension. If we keep going down, before we return that to the user, I just wanted to show what the UX looks like. Here's an adaptive card that comes back. It's defined in Jason, but you can also see the user experience for that. That all looks good. Let's go ahead and resume execution from the debugger. You can see that we've now returned back into teams the result and we've gotten attractive adaptive card that gets the user the information they need, does hand off with the Copilot. So right in line the user can complete their work. So pretty exciting. Again, full flexibility within the Teams AI library and Visual Studio Code. We're excited. Just like with Copilot Studio, this is something you can do today targeting hundreds of millions of Teams users, and very soon the ability to turn your custom Copilot into an extension will be available on. This is not new. This is something that hundreds of ISV's are already doing today across all sorts of experiences in Microsoft 365 Teams and now Copilot. Just two examples are leading software organizations Ezri and Thomson Reuters. Ezra is the market leader in Geographic Information systems. They're building a Copilot extension that adds spatial analytical capabilities directly into Teams meetings so that users can ask Copilot from Microsoft 365 a question about map data. And Copilot will seamlessly hand off to a rich interactive experience with visualization and Ezra's custom Copilot with all the associated context. Next up is Thompson Reuters, who's a leading global content and technology company that is transforming the legal profession with AI. Thomson Reuters is extending the Copilot experience in Outlook and Word in Teams for things like risk assessment so that based on the content of the Outlook e-mail, they can update the policy documents in Word and communicate those policy changes to reduce risk right within the Teams meeting. We're super excited to see what you do targeting this huge user base and we're gonna help promote and distribute your application through our unified marketplace, again reaching hundreds of millions of users today in Microsoft 365 and in Teams. And one of the reasons this marketplace is trusted is IT has confidence in the full governance capabilities for their own applications and custom extensions they build as well as the ones they will get from all of you building them. And a vibrant third party ecosystem. So to recap, we have a simple, powerful platform for AI and Microsoft 365 that you can use to be far more productive across the full spectrum. From something everyone can do in SharePoint to advance customization and Copilot Studio to the full power of Visual Studio and Visual Studio Code. But wait, there's one more thing. We are very excited to make Teams a fantastic place for developers to work together with AI to write better code faster. And so we've got a whole set of announcements around that this week at Build as well. First things first, source code inside teams with syntax formatting. And get this, we wondered what would get the applause. And that was my bet. And but wait, you should have held the applause with Microsoft Loop Co editing of that source code right within Teams. But yes, but there's more. Developers have asked us for a while for greater information density in Teams, so you can create and switch to Compact Mood, Compact mode to see much more content on the screen. And you can be much more productive with things like keyboard shortcuts, new slash commands for Teams, and of course developers, always in the flow of resolving issues and chat. And one of the features we're really excited to announce is meet Now, so they're right. Within chat you can bring up a ringless call between members of the team and resolve the issue in seconds. And last, developer teams love to have fun to break from the stress so you can use custom emojis reactions now in Microsoft Teams. And of course this is building on top of a growing set of partnerships with DevOps tools, Jira, Datadog, Pager Duty and of course deeper integration with GitHub and much more integrated with Microsoft Teams. So we are very excited to make Teams a great place for developers to work together to build this next generation AI. The next chapter is how Windows is the best platform for building that next generation AI, and to show that I'm excited, invite pop into the stage Pavin. Thank you, Jeff. Good morning. It is great to be here build. This is one of my favorite times of the year, connecting with fellow product makers about the world's canvas for innovation, Windows. I'm excited about the Copilot extensibility that we're Jesse and Jeff just shared. It really shines on Windows, the platform customers choose from Microsoft 365 and Copilot. Over the last year, we've learned so much about how Copilot can best serve you. We're working hard to make it even more valuable with the vision for Copilot meeting you right in your workflow. Imagine creating a presentation from a document in File Explorer, or helping customers troubleshoot their PCs using Quick actions and natural language right in Settings. We're focused on making Copilot even more contextual and useful across Windows. We took a big step towards that goal yesterday with the announcement of Copilot Plus PC's, the fastest and most intelligent PC's ever built. AI is woven into every layer of these devices, from the silicon to the operating system. With the most powerful PC NPU capable of delivering over 40 trillion operations per second, this new class of PC's is up to 20 times as powerful and 100 times as efficient for running AI workloads compared to traditional PC's from just a few years ago. Built together with our silicon partners AMD, Intel and Qualcomm, and our OEM partners, these PC's will be available June 18th. Starting with Qualcomm Snapdragon X series of chips, Copilot Plus PC's are redefining what you can do on a PC and setting the direction for the next decade of Windows to put this new wave of AI innovation in your hands. We're excited that Qualcomm has announced Snapdragon Dev Kit for Windows. It is designed to be your everyday dev box for AI with the power and flexibility you need As we define this new path for Windows in the era of AI, one thing that will never change is our commitment to openness. We recognize that the real value of Windows comes from the energy and the innovation of the ecosystem. It comes from all of you. So as we enter this new era, let's talk about how we're going to serve over a billion Windows customers together. As such, you said earlier, building a powerful AI platform takes more than a chip or a model. It takes reimagining the entire system from top to bottom. The new Windows Copilot runtime is the system that extends the Copilot stack to Windows. The Windows Copilot runtime is a new, integral part of Windows 11 and has everything you need to build great AI experiences, whether you're just getting started or already have models of your own. It includes the Windows Copilot library, a set of APIs that are powered by on device models that ship with Windows, and includes AI frameworks and tool chains to help you with your own on device models. And it's built on the foundation of powerful client silicon, including the NPU in the Copilot Plus devices. Let's take a look at how the Windows Copilot runtime enables an entirely new class of experiences O experiences like recall, that help users find anything they've seen on their PC. Inbox app experiences in Photos and Paint, which lets you bring your ideas to life using real time image generation. An app Experiences like Cap Cut and Acceptable and Da Vinci resolve some of our first partners using the new NPU and helping us build the Windows Copilot runtime. Looking ahead, the Xbox team has a vision for using the Windows Copilot runtime to empower players and game developers. Let's take a look that's pretty inspiring. This entire class of new experiences now benefit from faster task completion, enhanced privacy, and lower cost by using the Windows Copilot runtime. Next, let's take a look at the Windows Copilot library, the API's and models that support them. Let's take the Recall experience as an example. It relies on on device models deeply integrated into Windows to capture contexts on the screen. That data is transformed into vector embeddings and index in a vector store called the Windows Symantec Index. The Recall User Activity API allows you to extend your app into Recall so users can jump right back to where they were in your app and increase your app engagement in the same way Edge and Microsoft 365 apps like Outlook, PowerPoint, and Teams already have. In fact, soon Recall will draw on contexts from the Microsoft 365 Graph. You build your own semantic index store. You can use the Vector Embeddings API. That makes it possible to use retrieve log manner generational rag within your applications with your data. Imagine you have a Winforms or WPF app that works against a large corpus of sensitive data. With Vector embeddings API you'll be able to create on device vector stores for those records. That's powerful when combined with the RAGE API to enable natural language search in your applications for your users. Of course, that's just one example. The API's in Windows Copilot library cover the spectrum from low code APIs to sophisticated pipelines to fully multimodal models like the recently released Phi 3. The single best SLM in the world, 53 Mini does a better job than models twice its size on key benchmarks. Today we're thrilled to announce Phi Silica. Built from the Phi series of models specifically designed for the NPU and Copilot Plus PC's, it offers lightning fast on device inferencing and state-of-the-art first token responsiveness. Windows is the first platform to have a state-of-the-art SLM custom built for the NPU shipping in box. Now let's take a look at what you can do to bring your own on device models to Windows using frameworks and tool chains. It starts with Direct Mail, the lowest level machine learning framework in Windows. Similar to DirectX for graphics. Whether it's your own open source models or an open source model from Hugging Phase, direct mail helps you scale the breadth of your efforts across the Windows ecosystem by giving you to the metal access to GPU's and NP's. We also know that a lot of you do your development on Pytorch on Windows, and we're thrilled to announce that Windows will natively support Pytorch through direct mail. That's right, pretty exciting. Native Pie Torch support, of course, means that hugging face models will just work on Windows. And not just that, we're collaborating with NVIDIA to bring these workflows to over 100 million RTX AI GPU's in the Windows ecosystem. Now that's incredible. You can download the Pie Torch and Director Mel Developer Preview today. We're also going to extend Director Mel to our web developers, but introducing web and then on Windows Web and then is a web native machine learning framework. Microsoft has been working with Intel and other partners to unlock the access to local ML Excelerator so you can build performant AI experiences in your web. Apps. Behind me you see Clip Champs auto compose feature, achieving faster video composition experiences and cloud savings by leveraging the NPU through RT, Web and web. And then I'm excited to announce that Web and then is available in developer preview today. OK. So that's a glimpse of the Windows Copilot runtime, how it lays the foundation for innovation, giving you the largest catalog of models on the largest ecosystem of devices, making Windows the most open platform for I heard you there. That's fantastic. As Windows transforms for the era of AI, we're continuing to reach the expanse of the platform, including all the AI experiences you create with the Windows Copilot runtime. We're delivering Windows from the cloud with Windows 365, so your apps can reach any device, anywhere, and we're introducing Windows experiences to new form factors beyond the PC. For example, we're deepening our partnership with Meta to make Windows a first class experience on Quest devices, and Windows can take advantage of Quest unique capabilities to extend Windows apps into 3D space. We call these volumetric apps. Let's take a look. Workflows are transforming with Mixed Reality. Microsoft is partnering with Meta to bring Windows 365 and local PC connectivity to Quest and enable developers to easily extend their Windows apps into the 3D space. PTC has been working with this platform, bringing Creo into Mixed Reality in under a day. This extension allows users to enhance spatial understanding without leaving the app that powers their work. Sign up for the developer preview today. As developers, you'll have access. That's good to hear. As developers, you'll have access to volumetric API, and this is just one of many ways to broaden your reach through the Windows ecosystem. For decades, Windows has been a stage for the world's innovation. With Copilot Plus PC's, the Windows Copilot runtime, and Windows 365, we're going to unlock a new era of innovation together. Thank you. Back to you, Rajesh. Thank you, Pavan, and thank you, Jeff. Now, we've covered a lot of ground over the last 40 minutes. From the expansion of Copilot beyond a personal assistant to acting as a valuable team member, to how you as developers can extend Microsoft Copilot with your own copilots and agents in just a few clicks in SharePoint, to more advanced customization and Copilot Studio. You can use VS Code, full control of your models, your data and actions. And of course a phenomenal opportunity for developers with over a billion Windows customers. So I'm going to close with highlighting another customer, Amgen, A pioneering biotechnology company that harnesses the power of biology and technology to fight the world stuff as diseases. We are going to see how they harness Microsoft Copilot and their mission. Kevin Scott, our CTO, along with some special guests but round out the day one of keynotes. But before the video, let me just finish by simply saying thank you. Thank you for spending your time with us here at Bill. It means a great deal to all of us. And thank you for the trust you place working together, building the future with us every single day. Let's roll the video and thank you. We are at a once in a generation moment where we have the opportunity to harness the power of tech and biotech to revolutionize drug discovery and development, manufacturing and commercialization. We started with 300 licenses and now 20,000 of our employees have access to Copilot. In manufacturing, you're moving at a very fast pace. You have a lot of data coming at you. Copilot is 1 tool that allows us to synthesize that information and ensures that we can focus on what's most important the production of that drug. We've taken advantage of extensibility options using graph connectors. Identifying opportunities to use. Copilot for daily efficiency gains and then multiplying those by 100 or 1000 times in the life of a molecule from idea to achieving marketing authorization starts to give you a sense for what's possible. We have capabilities in our hands with these new tools in the early days of this new platform to absolutely do amazing things where literally the challenge for you all is to go do some legendary shit that someone will be in awe of you for one day. When people are looking for food resources, it can be a humbling experience. Childhood hunger and food insecurity is a problem that we can solve, and we are using AI to connect to those in need. We know through research that people prefer sometimes to talk to an AI agent to talk to a chat bot because they're not talking to a human. Technology can really help people get what they need with multiple resources in one place and it's making it mobile friendly, which is really important. Most people access this information via smartphone. What we're doing at No Kid Hungry is ensuring that we end childhood. Hunger. Round 2019 Brian started having some issues with his ankle. I was able to get him an appointment with a doctor and he was like, I hate to tell you, but this is ALS. How would I live in a world where my kids wouldn't be able to hear me say I love you? The biggest opportunities when it comes to emerging technologies is the gap that it can fill. AI has the ability to increase communication, increase independence. It's the next. WAVE for accessibility. What Microsoft is doing, their voice banking technology really gives someone back that intonation and a little bit more of their personality. Today we're going to play like newer version of your voice, so you just choose any phrase and then I'll play it on here as well. People, I am so happy to be with all of you. Item type it out. They're beautiful people. I'm so happy to be with all of. You. I think this sounds just like me and I am overwhelmed with joy that my family gets to experience this part of me. In many ways, I feel like I am back. Please welcome Chief Technology Officer and Executive Vice President of AI, Kevin Scott. Thank you all so much for being here with us today. So I've been fascinated my entire life by tools and the power that they give us as individuals and teams to really create extraordinary things. In my own personal making. My wife sometimes wishes that the ratio of tools to extraordinary things was a little bit different than they are. But as a person who has dedicated their entire career to building systems and infrastructure, tools and frameworks for other developers to use to go make the things that they need to make for whatever reason that they need to make them, it's just extraordinarily gratifying to see the really, truly consequential things that you all are choosing to do with these new AI tools that we had. A role in bringing into existence. So I just want to thank all of you so much for all of the great shit that you have made over the past year. So you know I want to give a call out to our friends at share our strength, you know, but it's not just the really wonderful work that we just showed in the video that folks are doing. Like as Satya mentioned earlier, we have over 50,000 customers using Azure AI tools and the Azure AI platform. And like this range of customers spans everything imaginable from all different industry categories, all different scales of business, from small startups to Fortune 100 companies and all different stages of exploration of how to use AI to do transformative things. From doing explorations like trying to find product market fit to scaling things where you found the product market fit to folks who are just trying to figure out how to optimize and enhance the things that they've been doing for very long while. So I just wanted to highlight another couple of things that I thought had been really noteworthy collaborations that my team has had with some of you all in the audience. So we've been doing some really cool work with Etsy using generative AI and the Azure AI platform to build features like their new gift mode. So we have the CTO of Etsy here in the audience with us today. Hello, and you all should go check out Gift Mode. Like it is like a really innovative way to help you figure out how to buy things for people who are difficult to buy for, which I'm guessing like almost all of us in the room are. It's like super stressful trying to figure out what the right gift is for the right person that you care about in your life. And this is just one of many interesting things that Etsy is doing with generative AI to really enhance the experience of the

Etsy product and to bring delight to their customers. Another thing that I'm super excited about is the work that Cognition has been doing. So Scott and the Cognition team are also here with us today. We have just recently announced A partnership between Microsoft and Cognition. Their product Devin is like an absolutely amazing tool. So like if you can imagine for yourself like some of the most tedious things that you that you do as an engineer or software developer, Devin is a tool designed to help you with those tasks. Like, I can't even tell you all the number of times is an engineering leader I've either or is an engineer that I've had to write code or lead teams writing code for doing things like re platforming and application. And like re platforming is like one of the laws of physics of like how we build software systems. Like it's a thing that must be done. But like, you know, it's rare that an engineer really enjoys doing that. And so like the incredible work that Devin is doing on top of these incredibly powerful tools is just really extraordinary. And we're super excited to be partnering with them and to bring all of the power of what they're doing to Azure and having their systems and infrastructure run on Azure. So you know, I I really wanna talk with you all today about just a couple of simple things. What's driving all of this progress? Like, why is all of this happening right now? And so part of it is like we're riding an extraordinary platform wave, Like something is fundamentally changing in the universe of technology. Much in the same way that it changed when we were going through the PC revolution, where Moores Law was driving an incredible increase in the power and lowering of the cost of personal computing, which led to it becoming ubiquitous. And you know something that we now all get to take for granted? Similar thing happened with the Internet revolution where networking technology connected all of this compute together and allowed us to do things that previously were unimaginable. And we're going through one of those major technological changes right now, being partly driven by a set of things that we'll talk about in a few minutes. Just the incredible scaling of the capability of AI systems as you apply more compute and more data to training them. But like before we get to that expansion of the frontier of the increase in those capabilities, like our super important part of the emergence of a new powerful platform is sort of completing the stack. So it's actually hard work even when you have a piece of technology that is improving at an exponential rate to figure out how to do all of the things that have to be done in order to deploy it in real applications so that you can go out and deliver value to real customers who care about what it is that you're doing. And we've done a huge amount of work over the past year on the Copilot stack. It is both optimizing a bunch of systems so things are getting cheaper and more capable. And it's also building that whole cloud of capabilities and systems, services and tools around the core AI platforms, the big models that you all need and the choices that you all want so that you can build the things that matter to you under the constraints that you're operating under. So. One of the reasons that we have been able to do this is no other company has deployed more generative AI applications over the past year that Microsoft has. And so you have probably heard us over the past year talking about all of these different copilots like this new software pattern that we originated with GitHub Copilot where you pair powerful generative AI with this user interface paradigm where you're using the AI to help assist users with task. And so you can apply this to everything. And I know many of you in the audience are building your own copilot. So Microsoft itself is building copilots for service for sales, a copilot in Bing, Copilot in Edge, Copilot in Windows. And the reason that we've been able to do all of this work is because we have the Copilot stack that we built for ourselves to help us have real agility in getting these products built quickly, to have them built efficiently where their price and cost optimized and to build them in a way where they're safe and secure. One of the things that you have heard from Rajesh and that you'll be hearing a lot more of it build is like part of what the Copilot stack is allowing us to do is to unify the experience across all of these copilots into one logical Microsoft Copilot where you don't have to really pay attention to which Microsoft product or service you're in. Like. The Copilot just understands all of your context and delivers all of the capability of the model in the context of your data and your task two you when you need it. So. The other thing that is really driving progress is not just this sort of completion of the Copilot stack, this sort of progress that we're making and filling out that toolkit for you all so that it is easier for you to build software. But we are riding like a fundamental wave in in the development of this AI platform where if you just sort of look at compute overtime, like how much? How much? Yeah, GPU cycles or accelerator cycles that we're using to train the very biggest models in the world since about 2012. Like that rate of increase in compute when applied to training has been increasing exponentially and we are nowhere near the point of diminishing marginal returns on how powerful we can make AI models as we increase the scale of compute. So we're sort of doing two things at once at Microsoft. So we are optimizing the current frontier and like building that toolkit to help you all leverage it, while at the same time investing at a pretty incredible rate in pushing the frontier forward. And so one of the Super interesting things that has just happened as we're sort of pushing the the frontier forward and like really, really driving forward on efficiency is what our partners Open AI launched last week in the form of GBT 4 O. As Satya mentioned earlier, GP Four O is like a really stunning achievement multimodal model that understands a bunch of different input types from video to text to speech. They can respond in a bunch of rich ways from text to speech and eventually video. It can respond to you, to the interactions for users in their applications in real time, and in the case of the chat IT demos that folks have seen, you can even interrupt the model so that you can have really fluid interactions with these systems. And then an enormous amount of work has gone into GPT 4 O both the model itself as well as well as the supporting infrastructure around it to ensure that it's safe by design. So I wanted to also just remind folks like this efficiency point is, is real. So while we're off, you know, building bigger supercomputers to get the next big models out and to deliver more and more capability to you like, we're also grinding away on making the current generation of models much, much more efficient. So between the launch of of GPD 4, which is not quite a year and a half ago, now it's 12 times cheaper to make, got a call to GPD 40 than the original chat or the GPD 4 model and it's also six times faster in terms of like time to 1st token response. So it's just really extraordinary like how much progress we're making because of like the full set of optimizations that we're doing from the silicon that we're building, networks, data center optimization, as well as like an incredible amount of software work that we're doing on top of all of this hardware and infrastructure to really tune the performance of these systems. And the great thing is, is again, there's no point of diminishing marginal return here. And one of the messages that I want to land with you all today is that you can count on things getting more robust and cheaper at a pretty aggressive clip over time. So it's like a really important thing to internalize like something, you know, like we challenge ourselves on at Microsoft all the time is like aim for things that are like really, truly ambitious. Because like all of this optimization work is going to accrue to make things like really ubiquitous in terms of like how you can go deploy them. I just want to show a little quick demo video here. So let's roll the video. Show you. Hey Microsoft Build conference is tomorrow and I have not been sleeping and I am making stupid mistakes. Can you help me take a look at this code? I'd be happy to help. Could you tell me a bit more about the issue you're facing with the code? For example, are you getting an error message or is the code not doing what you expected to do? Let me just show this to you. It looks like you're working on a Python. Function to summarize main points from a keynote. What I'm trying to do is go through a list of keynoters and summarize all of their work, and my problem is that I wanted to just add Kevin's name to the end of the keynote list and it added every character of his name instead of just his whole name. What did I do wrong here? Change keynote list dot extend to keynote list dot append last keynote. This will add Kevin as a single element to your list, just like Satya and Rajesh. Ooh, that sounds good. Let me try that sounds good. Give it a go and let me know how it works. OK, I replaced extend with a pinch. Does that look right? Yes, that looks better. All right, I made that change. Let me run it and try. Sounds good. Let's see how it goes. It worked. Satya, Rajesh, and Kevin all came out awesome. I'm glad it worked out. Sometimes the smallest things can trip us up, especially when we're sleep deprived. Thank you. So yeah, I believe it. It really is extraordinary. So like, I should say by the way, that Jennifer would never make that actual mistake in writing a Python application, but Kevin might. So, like, this is all giving me hope that a rusty old CTO might one day be able to program again if if the need arose. But but like I, I do want to make sure that we're paying attention to like just how much has changed over the past year. I mean, like what what you just saw would have been absolutely inconceivable to think about actually working. And like this was, you know, not a tortured demo. Like, you know, we were sort of looking for something to show off GBD 4oD you all last night. And Jennifer showed me this and then she just recorded this demo. Like this is just crazy that it works this well. So another set of things that have been really making a huge amount of progress is like what's possible with smaller models. So we have been working for a while on this series of models called Phi that are small language models. So Satya chatted a little bit about this in his keynote earlier. And the way that we think about Phi is just imagine an efficient frontier. And so usually when you're building these models, you're trading a couple of things off. So you can trade size off, which is related to, you know, performance and costs and a whole bunch of other things versus quality. So the smaller the model is like the cheaper it is to do inference and like the less compute that you need to actually run the model. So small models are more amenable to running on devices, but it usually means that you have to take a hit on quality. And like what we're discovering in particular over the past year is that there's this notion of an efficient frontier. So we we don't even show the GPD 4 O point on this slide. It would be like way, way, way off to the right just in terms of the size. So like if you want extreme levels of quality and performance, like a frontier model is your friend. But in some cases, like you may want to choose one of these other models somewhere else on this efficient frontier where the trade off that you're making between cost to serve or latency or locality is acceptable given the quality that you can get. And the very interesting thing that's been happening over the past year is the quality that you're able to achieve in these small models is getting pretty high. So I just wanted to show this as an illustration. So like, you know, remember back ancient history the to the launch of ChatGPT in November of 2022. So ChatGPT launched on top of GBD 35. And so like everybody was just absolutely gobsmacked at what was possible with GPT 35I mean, just as sort of a stunning, revolutionary thing that happened. We Fast forward a few months to March 2023, and she BT ChatGPT gets an upgrade to GPT 4, which is even more extraordinary what it's able to do. Like you're able to ask extremely complicated questions of these things and get very rich, interesting, compelling completions. So Fast forward to today and like you can sort of see that a version of Phi 3 optimized to run on a mobile phone can respond to a prompt, just like ChatGPT could just a year or so ago with responses that are sort of equivalent and like this is not arguing that, you know, Phi 3 that's running on this device is just as powerful as GBD 4, It is not. But the way that you all should be thinking about it is. In many cases, these models can be appropriate to use for building your applications when you have a particular set of constraints that you're trying to optimize towards. It's like all of this is kind of, you know, abstract in a way. And so I I wanted to really motivate why this matters with the following example. So Satya mentioned earlier the partnership that Microsoft has formed with Khan Academy. And Khan Academy's mission is really interesting and important, Like they are trying to ensure that every learner on the planet, no matter where they are, has access to high quality, individualized instruction. And so one of the things that we are exploring together with Khan Academy is the possibility of achieving that goal of ubiquity of these personalized learning agents by using something like Phi 3 where you can imagine training A53 model that's very good at something like math instruction. So this is an actual interaction with Phi 3 medium that has been fine-tuned to work particularly well for math tutoring. And the challenge with doing something like this is that you have to not just have the model give the student an answer, but like you want it at lead them towards discovering the answer themselves. So like a tutor is very different from an answer agent. And so like, it's just exciting to think about how many tools that organizations like Khan Academy have to solve, like these really, really important missions that they have in the world. And so with that, I'd love to bring Sal Khan from Khan Academy onto the stage. Hey, Sal, thank you so much for being here with us today. So, you know, we have been chatting about your mission, I think for quite a while. And you know, one of the interesting things that happened when ChatGPT burst onto the scene a few years ago is that there was this reaction from a bunch of educators. And like, I think it was actually reasonable reaction where like, OK, we don't understand this. Like we don't want our students using it. They're going to, like, do things that we would prefer that they didn't do it. You, on the other hand, looked at this and said this is amazing and leaned all the way in. Can you explain a little bit about what drove your first reaction to this new technology? Yeah, you know some of you all know how Khan Academy got started. If you go almost 20 years ago, it started with me tutoring a cousin. I was a hedge fund analyst at the time. I tutor one cousin. Words present. My family for tutoring is going on before I know what I'm tutoring 1015 cousins and I start writing tools for them software. I started making videos. That's a lot of people are know about Khan Academy and if you think about that journey from then. Till now, even right before we started really working on generative AI, everything we've been doing is how could you scale that type of personalization that I was originally doing with my cousin Nadia and we were approximating it with software and videos and teacher tools. But what we saw, but to some degree there was going to be we're going to asymptote on how far you could get with pre generative AI tools. And then when we saw it was really GPT 4 that opened our mind. You know, Greg and Sam from Open AI showed it to us end of summer 2022 and we realized there's things that have to be worked out, but it could get that much closer to emulating what a what a real tutor would do. And it was obviously it could also be used as a cheating tool and you have to worry about safety and privacy, especially with under 18 users. But I told the team, let's turn those into features. Let's put the guardrails on it, because this could get us that much closer to our mission, which is free world class education. Yeah. I think one of the other things that you all have done and like this is a really important thing to internalize about these models and systems is like the model isn't a product. And like the systems aren't silver bullets. Like, you still actually have to understand like who your customer is, like what problem you're trying to solve, like how to go deal with a whole bunch of gnarly things on top of the, you know, this incredibly interesting and powerful tool so you can do something useful. Like you want to talk a little bit about what you had to do there? And I have to admit, and maybe everyone, a lot of people in this room where in the world around are experiencing this every now and then you see some of these demos and you're like, does my application even have relevance anymore? This thing's going to be able to do everything. But then when you sit down and you really think about how a school system a teacher as a student's going to use it and you're going to what are the guardrails what are the privacy. How do you make sure that it really does tutoring interactions appropriately it's aligned to standards. You realize that there's a lot to do at the at the application layer now I think we're all discovering together this new world of developing applications on terms of on top of large language models. It's not deterministic in the traditional way. You have to have a evals you have to constantly test it but but we're we're realizing that there's just so much to do. It really is a bit of a a. It's a very exciting time. Yeah. I mean, one of the things that I'm especially excited about is like this mission that you all have for ubiquity and like the partnership that we're doing with you all is going to enable you to get the hand or like, get every teacher in the United States hands-on Conmigo and your tools. And, you know, like just a personal anecdote for me is like, my daughter is in the ninth grade. She's taken biochemistry and just in love with science in general. And she on her own, without any prompting from dad, figured out how to use the free version of ChatGPT to take a bunch of biochem papers that were like way, way, way more complicated than a 15 year old by rights has to understand, dump them into Chad GBT, and then just ask a million questions about it. And like her learning acceleration because she's figured out how to use this tool is extraordinary. And like, I just want every kid in the world to, like, have the same experience that my daughter has. Absolutely. And you know what? We realize that Khan Academy there is a subset of students that if you give them the tool and it sounds like you're lucky enough to have a daughter like that, they will run with it. But what you really need in most cases is you need caring adults, primarily teachers in the room, motivating students, driving that usage. And So what we're really excited about this partnership and this is a big deal. I want to make sure you know we are using state-of-the-art models that use real compute. It has real cost associated with it. When we launched Conmigo and which is still out there and it's a tutor for students, it's a teaching assistant for teachers. But what we're launching today as part of this partnership is the state-of-the-art teacher tools. We're going to be able to give free to every teacher in the United States so that they can get it productivity improvements. It big, big, big, big deal. I, I, I, I actually think teaching will be the the first mainstream profession to really benefit from generative AI, lesson planning, progress reports, grading papers, etcetera, etcetera. And I think if we can win teachers, hearts and minds, then it gives us that much a better chance of also being able to reach students. So one last thing before we go. I know that you have just written a book and like having written a book myself, like it's like a lot, a lot of work. So why? And like everybody should like, you know, pay attention to these of fantastic endorsements the book has gotten and like, go read a copy of it. It's like a fascinating work, not just about education and your mission, but like, I think also it has a bunch of really interesting lessons about how you can ambitiously use AI to solve hard problems. So, but but why did you choose to write a book now? You know, this is the second book that I wrote. The first book I wrote back in 2011. And I remember when the the publisher had reached out to me then I was like, why would I write a book? I could just put it all on YouTube and, you know, share it with the world. But there's something about writing, writing a book that lets you frame the problem. And I felt that 10 years ago, when Or 12 years ago, when Khan Academy first came on the scene. And I think this moment, we all feel even a little bit more overwhelmed. I mean, you mentioned these inflection points, this exponential growth. We're all feeling like things are changing every week. And I wanted to take the time to, for myself, understand where we are and where we're going. And hopefully a reasonably timeless way that's not going to be dependent on whatever the frontier model of the day is. And I think especially anyone who cares about education, cares about work, cares about what skills their kids should learn to keep up, how we can all be more productive, what the future of admissions and recruiting is going to look like. Hopefully this will be useful for them too. Yeah. Well, I am incredibly grateful for the partnership that we have and more importantly, for the work that you're doing in the world. So thank you so much. For being with us today. Thank you. So. Another really incredibly impactful area where these new AI tools and platforms are going to have an enormous impact is healthcare. So I just wanted to share another personal anecdote with you all. So like, I grew up in rural central Virginia and my mom and brother and most of my family still live in the place where I grew up. My mom is like a 74 year old Southern woman and it has been suffering from a thyroid condition for 26 years, which, you know, entirely under control. Like she's been taking medication for it. And just last fall she had like some sort of change in her, like in her system where like all of a sudden her medication wasn't working as well as it used to. Which resulted in her spending a bunch of time in the hospital like trying to figure out what was going on. And the health care system in the part of the world where my mom lives is like super overburdened like it it's you know not a place where like tons of people are sort of moving for economic opportunity. Like it's one of those places where people tend to move away from to seek economic opportunity which has impacts on everything that is happening in that part of the world. And I sort of as I was trying to help my mom navigate this situations or do I was looking at how could AI have helped relieve the suffering that she was experiencing. And you know if doctors everywhere had access to these tools, like a lot of what she went through could have been immediately alleviated. Like if you just take her symptoms that she was presenting and put it into GBD 4 along with her chart, it would immediately say like go get this test which is like going to help find root cause. And then the results of the tests, like could be input into the same session, which would then sort of give a set of recommendations to doctors about course of therapy. And like if that had happened in her case, like what was 6 visits to the hospital, like could have been compressed to one. And you know, I I think about this a lot because I, I worry even whether or not my mom would have pulled out of a health spiral that she was in if she hadn't had me for a son intervening on her behalf. And I worry about all of the people in the world who don't have someone to intervene on their behalf, like who are interacting with one of these resources that is overburdened. And so, like, I just want us all to think about as we're imagining what the set of possibilities are for, you know, what we go do with AI, like things like what Soil is doing with Khan Academy. And like some of the amazing potential that we have to reduce suffering in the world. And to like help make, you know, things like high quality healthcare, more equitable and accessible, like just super inspiring to think about. So you know, the last thing that I want to chat about before we get to conversation with Sam Altman is how we at Microsoft have been thinking about building applications on top of this incredible platform that is emerging right now. And so the challenge I think for us, and I think it's the same challenge that all of you face, is that you really want to focus on things that have made the transition from impossible to merely difficult. Like that's where all the interesting stuff is. Like if you look at the history of platform revolutions, like that's where all the interesting companies emerge from. It's where all the innovation happens. It's where all of the value gets unlocked. And in the case of technology platforms that are sort of exponentially progressing, it's like the only reasonable place to go aim because if you're aiming somewhere different, like the platform is becoming so much more capable and so much cheaper over time that everything that you sort of have in your imagination, it's too expensive to do right now, are too fragile, is going to become cheap and robust before you can even blink your eye. And so like that is really, you know, the the thing more than anything else that I would say to all of you to take away from what I'm saying here today is like really focus on those phase transitions. So while you all have been out there, you know, grinding away, building really extraordinary things over the past year with all of these AI tools that are coming. Like, we've been hard at work trying to make forward progress on our AI platform. So yeah, we talked a lot about how we're optimizing the current frontier, like making things cheaper and making them more powerful and complete. But we've also been hard at work building new supercomputing infrastructure and working with our partners at Open AI to push that frontier forward. And like we showed this slide at the beginning, like there's this like really beautiful relationship right now between the sort of exponential progression of compute that we're applying to building the platform to the capability and power of the platform that we get. And I just wanted to you know sort of without without mentioning numbers which is sort of hard to do to give you all an idea of this scaling of these systems. So in 2020 we built our first AI supercomputer for Open AI. It's the Super computing environment that trained GBD 3. And so like we're going to just choose marine wildlife as our scale marker. So you can think of that system about as big as a shark. So the next system that we built scale wise is about as big as an Orca and like that is the system in that we delivered in 2022 that trained GPT for the system that we have just deployed is like scale wise about as big as a whale relative to like you know the shark sized supercomputer and this orca size supercomputer And it turns out like you can build a whole hell of a lot of AI with a whale sized supercomputer. And. And so you know one of the things that I just want everybody to really, really be thinking clearly about and like this is gonna be our segue to talking with Sam is the next sample is coming. So like this whale sized supercomputer is hard at work right now building the next set of capabilities that we're going to put into your hands so that you all can do the next round of amazing things with it. And so with that, I'd like to bring Sam Altman to the stage. Good to see you too. So you are one of the busiest people on the planet. Wild week, it's, yeah, it's a wild week. It's a wild year, man. But so I I really appreciate you taking time out to chat with us today. So I I guess what I really wanted to start our conversation about and like I asked you this question last week is, you know, there there's just been an extraordinary amount of change over the past year and a half year. Like what has been the thing that has surprised you most like, particularly relevant to an audience? Developers, I mean, I'm delighted to be here. Obviously great to see you. But developers have been such a core part of what's been happening this last year and a half.

There's millions of people building on the platform. What people are doing is totally amazing and the speed of adoption and talent and figuring out what to build with all of this over what has really not been very long, like when we put GPD Three out in the API, some people thought. It was cool, but. It was narrow where the very happened and seeing what people have done with GPT for and seeing now what's happening with GPT 4 O, even though it's new and hasn't been out that long, is quite remarkable. I've never seen a technology get adopted so quickly in such a meaningful way. The what people are building, how people are finding out how to do things that we never even thought of possible, which is why it's always great to have an API that's been very cool to see. Yeah. And I, I think, you know, what you just said is like one of the most important points to me. Like there there's a version of AI that could have existed that is, you know, like a bunch of smart people. Like building, you know, things that extraordinary scale and then just building it into a bunch of products where everybody gets to passively use them. Like the the really brilliant thing that you all have done is like taken the exact same set of things and like decided to make it available to like any developer who's able to sign up for an API key. Yeah, we we try to be really thoughtful about what makes a good API. Artists. There's gonna be all kinds of ways people can use this, but the more this can just be a layer that gets built into every product, every service, the better. And we've tried to make it such that if you want to add intelligence to whatever you were doing, any product, any service, we make that very easy. Yeah. And like again I think the progress has been stunning. So you know, I I think you know the the setup for like introducing you on to the stage here was. I saw that Big Blue whale. Yeah, like you know, you're making good use of the whale sized computer right now. And so like I, without like getting too specific, which we can't be obviously, like what are the category of things that people should be expecting over the next, you know, K months? So the the most important thing, and this sounds like the most boring, obvious, trite thing I can say, but I think it's actually much deeper than it sounds. The most important thing is that the models are just going to get smarter generally across the board. There will be a lot of other things too, which we can talk about, but if you think about what happened from GPT 3 to 3.5 to 4, it just got smarter and you could use it for all these things. It got a little more robust, it got much safer, both because of the model got smarter and we put much more work into building the safety tools around it. It got more useful but the underlying capability, this amazing emergent property of like we actually are seeming to increase the general capability of the model across the board that's going to keep happening and the the jump that we have seen in the utility that a model can deliver with each of those half step jumps and smartness, it's quite significant each time. So as we think about the next model and the next one and the incredible things that developers are going to build with that, I think that's the most important thing to keep in mind. Also speed and cost really matter to us. So with GPT 4 O we were able to bring the price down by half and double the the speed. New modalities really matter. Voice mode has been actually a genuine surprise for me and how much I like the new voice mode and when people start integrating that, I think that'll matter. But but it's the overall intelligence that will be coming that I think matters the most. So you for awhile now have been one of the most successful startup investors in the world. And like now you are one of the most successful CEO's of one of the most important companies in the world is that you've got a room full of developers here. You know, like, I think they're 5000 people in the room and they're about 200,000 people online right now. What's your advice to them is like they think about how to spend their precious time given what's happening in the world. Like what? What's your advice? 2 Two things. Number one, this is probably the most exciting time to be doing, building a product, doing a startup, whatever it is that we have seen at least since the mobile boom and probably, I would say, since the Internet and maybe even bigger than that, we don't know yet. But the, the big opportunities, the big the, you know, the ability to sort of build something new and really kind of like change the landscape that comes at the platform shift times. And we haven't had a platform shift in a while and this looks like it's really truly a platform shift. And so my biggest piece of advice is like this is a special time and take advantage of it. This is like not the time to delay what you were planning to do or wait for the next thing. Like, this is a special moment and a few years where a lot of stuff is going to happen and a lot of like great new things are going to get going. I'm the second thing also about platform shifts is when the mobile phone revolution started or really got going like 2008, 2009 you would see people say we're a mobile company, you know we're having a mobile app and then only a few years later no one said they were a mobile company because it was like table stakes and an amazing new technology which I would bias, but we'll put AI in that category. It doesn't get you out of the hard work of building a great product or a great company or a great service. You still have to do it. AI alone is a new enabler, but it does not automatically break the rules of business. And so you can use this as like a new thing to do, but you still have to figure out how you're going to build enduring value and whatever you're doing. And it's easy to lose sight of that and the excitement of the Gold rush. Yeah. So one last thing before we let you go. So you know, you and I and like members of your team and members of the Microsoft team have been doing really an extraordinary volume of work over the past year and a half, two years thinking about safe deployment of an awful lot of AI capability like everything from you, your APIs and developer tools to end products. And you know, I think we, you know, have accumulated a really interesting volume of experience, like experience. It's sort of hard to get if you're not doing deployments at this scale. So I I, you know, and I think you just mentioned something that's like really, really interesting like part of, you know, part of the interesting and surprising progression of capabilities of these models means that they're more useful in like helping to like make AI. System safer. So I I don't know whether you had some thoughts you wanted to share there as well? You know, when we first developed this technology, we spent a lot of time talking about, alright, we've made this thing, it's cool. Are we ever going to be able to get it to an acceptable level of robustness and safety? And now we kind of take that for granted with GPT 4. You know, if you use it, it's far from perfect. We have more work to do, but it is generally considered robust enough and safe enough for a wide variety of uses. And that took an enormous amount of work across both teams and fundamental research. Like when we started this, we're like got this thing, we've got this language model, it looks like kind of impressive and kind of not. And even then how are we gonna like get it aligned and what it what does it mean? You know, what is it going to take to be able to deploy it? The number of different teams we've had to build up to go from research and creation of the model to safety systems to figure out policy to how we do the monitoring, that's a huge amount of work. But it's it's necessary to be able to deploy these and use them. Like, you know, when you take a medicine, you want to know it's going to be safe. When you use an AR model, you want to know it's going to be robust and behave the way you want And have been super proud of the work that teams have done together. And I think it's amazing how fast this much work has happened and that we can all now use this and say, Oh yeah, basically it basically works. As the models get more powerful, there will be many new things we have to figure out as we move towards AGI. The level of complexity, and I think the new research that it'll take will increase. I'm sure we'll do that together, but we view this as a gate on being able to put these things out in the world, which we really want to do. Yeah, it's definitely table stakes. So thank you so much. Thank you for being with us here today. Like, I really appreciate your time. It's awesome to hear from you. Alright, so I will. I I think this is all that's separating you all from the rest of your build and probably lunch. So like my very last thing for you all is the following call to action. So Microsoft and our partners, like Open AI, are spending an extraordinary amount of energy and investing like capital at an unprecedented scale, trying to make sure that we are building a genuinely valuable platform that, like, all we're doing is building the platform. And like, I don't even think that that's the most important part of the AI revolution that's happening right now. It's you who are doing the work. Like, you're the ones who are making all of these things matter. Like, we could build platforms all day, all night. And if you all didn't have the great ideas, like you didn't understand the consequential impact that you wanted to have on the world, it would all be for nothing. And so I'm incredibly grateful for all of the things that you all have done on the platform over the past year and I am incredibly excited to see what you all are going to go do in the year ahead. Thank you all so much.

Generated for FREE with Youtube to Text: https://www.youtubetotext.org

Support our services by donating: https://link.youtubetotext.org/make-a-donation

Thanks for using our services!