05212024 Build Kevin Scott

Microsoft Build 2024
Kevin Scott
Redmond, Washington
Tuesday, May 21, 2024

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.

I want to give a call out to our friends that share our strength, but it's not just the really wonderful work that we just showed in the video that folks are doing. As Satya mentioned earlier we have over 50,000 customers using Azure AI tools and the Azure AI platform.

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 a very long while.

I just wanted to highlight another couple of things that I thought have been really noteworthy collaborations that my team has had with some of you all in the audience. 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.

We have the CTO of Etsy here in the audience with us today. Hello.

You all should go check out Gift Mode. Like. It is 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 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. 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 an absolutely amazing tool. If you can imagine for yourself some of the most tedious things that you do as an engineer or software developer, Devin is a tool designed to help you with those tasks.

I can't even tell you all the number of times as an engineering leader, or as an engineer, that I've had to write code or lead teams writing code for doing things like re-platforming an application. Re-platforming is one of the laws of physics as to how we build software systems. It's a thing that must be done, but it's rare that an engineer really enjoys doing that.

And so 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.

I really want to talk with you all today about just a couple of simple things. What's driving all of this progress? What is all of this happening right now?

And so part of it is that we're riding an extraordinary platform wave, where 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 Moore's Law was driving an incredible increase in the power and lowering of the cost of personal computing, which led to it becoming ubiquitous, and something that we now all get to take for granted. A 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.

We are 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, and just the incredible scaling of the capability of AI systems as you apply more compute and more data to training them.

But before we get to that expansion of the frontier and the increase in those capabilities, a super-important part of the emergence of a new, powerful platform is sort of completing the stack. 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.

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 tasks.

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, and a Copilot in Bing and a Copilot in Edge, and the Copilot in Windows.

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, and to have them built efficiently where they're 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 that 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. The Copilot just understands all of your contexts and delivers all of the capability of the model in the context of your data and your task to you when you need it.

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 in filling out that toolkit for you all so that it is easier for you to build software, but we are riding a fundamental wave in the development of this AI platform. If you just sort of look at compute over time, like how much GPU cycles or accelerator cycles that we're using to train the very biggest models in the world since about 2012, that rate of increase in compute when applied to training has been increasing exponentially.

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, and so we're sort of doing two things at once at Microsoft. We are optimizing the current frontier and 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 frontier forward and really, really driving forward on efficiency is what our partners at OpenAI launched last week in the form of GPT-4o.

As Satya mentioned earlier, GPT-4o is like a really stunning achievement, a multimodal model that understands a bunch of different input types, from video to text to speech, and it can respond in a bunch of rich ways, from text to speech and eventually video. It can respond to you or to the interactions for users in their applications in real time. And in the case of the ChatGPT 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-4o, in both the model itself, as well as the supporting infrastructure around it, to ensure that it's safe by design. I wanted to also just remind folks that this efficiency point is real, and so while we're off building bigger supercomputers to get the next big models out and to deliver more and more capability to you, we're also grinding away on making the current generation of models much, much more efficient.

Between the launch of GPT-4, which was not quite a year-and-a-half ago now, it's 12 times cheaper to make a call to GPT-4o than the original chat or the GPT-4 model, and it's also six times faster in terms of time to first token response. It's just really extraordinary how much progress we're making because of the full set of optimizations that we're doing from the silicon that we're building, networks, data center optimization, as well as 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.

The great thing is that, again, there's no point of diminishing marginal return here. 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. It's a really important thing to internalize, and something that we challenge ourselves on at Microsoft all the time is to aim for things that are really, truly ambitious because all of this optimization work is going to accrue to make things really ubiquitous in terms of how you can go deploy them.

I just want to show a little quick demo video here to show you.

(Video segment begins.)

JENNIFER: 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?

GPT: 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 expect it to do?

JENNIFER: Let me just show this to you.

GPT: It looks like you're working on a Python function to summarize main points from a keynote.

JENNIFER: 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?

GPT: Change keynotelist.extend to keynotelist.append, last keynote. This will add Kevin as a single element to your list, just like Satya and Rajesh.

JENNIFER: Ooh, that sounds good. Let me try that.

GPT: Sounds good. Give it a go and let me know how it works.

JENNIFER: OK. I replaced extend with append, does that look right?

GPT: Yes, that looks better.

JENNIFER: All right, I made that change. Let me run it and try.

GPT: Sounds good. Let's see how it goes.

JENNIFER: It worked. Satya, Rajesh and Kevin all came out.

GPT: Awesome. I'm glad it worked out. Sometimes the smallest things can trip us up, especially when we're sleep deprived.

JENNIFER: Thank you.

(Video segment ends.)

KEVIN SCOTT: I mean, it really is extraordinary. I should say, by the way, that Jennifer would never make that actual mistake in writing a Python application, but Kevin might, and so this is all giving me hope that a rusty old CTO might one day be able to program again if the need arose.

I do want to make sure that we're paying attention to just how much has changed over the past year. I mean, what you just saw would have been absolutely inconceivable to think about actually working, and this was not a tortured demo, you know, where we were sort of looking for something to show GPT-4o to you, last night, and Jennifer showed me this. And then she just recorded the demo, and it's just crazy that it works this well.

Another set of things that have been really making a huge amount of progress is what's possible with smaller models. We have been working for a while on this series of models called Phi that are small language models. Satya chatted a little bit about this in his keynote earlier.

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. You can trade size off, which is related to, you know, performance and costs and a whole bunch of other things versus quality. The smaller the model is like the cheaper it is to do inference and the less compute that you need to actually run the model, and so small models are more amenable to running on devices, but it usually means that you have to take a hit on quality.

What we're discovering in particular over the past year is that there's this notion of an efficient frontier. We don't even show the GPT-4o point on this slide. It would be way, way, way off to the right, just in terms of the size. If you want extreme levels of quality and performance, a frontier model is your friend, but in some cases, 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. 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.

I just wanted to show this as an illustration. If you can remember back to ancient history, to the launch of ChatGPT in November of 2022, it launched on top of GPT-3.5, and everybody was just absolutely gobsmacked at what was possible with GPT-3.5, with just this sort of stunning revolutionary thing that happened.

We fast forward a few months to March 2023, and ChatGPT gets an upgrade to GPT-4, which is even more extraordinary in what it’s able to do. You are able to ask extremely complicated question these things and get very rich interesting, compelling completions.

If you forward to today, 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.

This is not to argue that Phi-3 running on this device is just as powerful as GPT-4, because it is not, but the way that you all should be thinking about it is that, 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.

All of this is kind of abstract in a way, and so I wanted to really motivate why this matters with the following example.

Satya mentioned earlier the partnership that Microsoft has formed with Khan Academy, and Khan Academy’s mission is really interesting and important. 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 a Phi-3 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. The challenge with doing something like this is that you have to not just have the model give the student an answer, but you want it to lead them towards discovering the answer themselves.

A tutor is very different from an answer agent, and so it’s just exciting to think about how many tools that organizations like Khan Academy have to solve 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.

We have been chatting about your mission, I think, for quite a while, and 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. I think it was actually a reasonable reaction where like, "OK, we don't understand this. We don't want our students using it. They're going to do things that we would prefer that they didn't do."

On the other you hand, you looked at this and said, "This is amazing." You leaned all the way in. Can you explain a little bit about what drove your first reaction to this new technology?

SAL KHAN: Yeah. Some of you don't 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 tutored one cousin. The word spreads in my family that free tutoring is going on, and before I know it, I'm tutoring 10 or 15 cousins and I start writing tools for them, and software.

I started making videos, which is what a lot of people know about Khan Academy. And if you think about that journey from then until 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?

We were approximating it with software and videos and teacher tools, but to some degree, we were going to asymptote on how far you could get with pre-generative AI tools. And then when we saw it, it was really GPT-4 that opened our mind. Greg and Sam from OpenAI showed it to us in the summer 2022, and we realized there were things that have to be worked out, but it could get that much closer to emulating what a real tutor would do.

It was obvious that 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.

KEVIN SCOTT: Yeah, I think one of the other things that you all have done, and this is a really important thing to internalize about these models and systems, but the model is in a product and the systems aren't silver bullets. You still actually have to understand who your customer is, what problem you're trying to solve and how to go deal with a whole bunch of gnarly things on top of this incredibly interesting and powerful tool so you can do something useful.

Do you want to talk a little bit about what you had to do there?

SAL KHAN: Yeah, and I have to admit, and maybe everyone – a lot of people in this room or in the world right now 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 is going to be able to do everything. But then when you sit down and you really think about how a school system, a teacher or a student is going to use it, and you are going, "What are the guardrails, where is the privacy," and how do you make sure that it really does the tutoring interactions appropriately, it's aligned to standards. You realize that there's a lot to do at the application layer.

Now, I think we're all discovering together this new world of developing applications on top of large language models. It's not deterministic in a traditional way. You have to have evals, you have to constantly test it, but we're realizing that there's just so much to do. It really is a very exciting time.

KEVIN SCOTT: Yeah. I mean, one of the things that I'm especially excited about is this mission that you all have for ubiquity. And the partnership that we're doing with you all is going to enable you to get the hands or get every teacher in the United States hands on Khanmigo and your tools.

And just a personal anecdote for me is my daughter is in the 9th Grade. She's taking 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 way, way, way more complicated than a 15 year old, by rights, has to understand, dump them into ChatGPT and then just ask a million questions about it. And her learning acceleration, because she's figured out how to use this tool, is extraordinary. And I just want every kid in the world to have the same experience that my daughter has.

SAL KHAN: Absolutely, and we realized at Khan Academy, there is a subset of students that if you give them the tool, and it sounds you're lucky enough to have a daughter 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… We are using state-of-the-art models that use real compute. It has real cost associated with it, when we launched Khanmigo, 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 these 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 a productivity improvement.

(Applause.)

KEVIN SCOTT: Yeah, big, big deal.

SAL KHAN: Big deal. I actually think teaching will be 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.

KEVIN SCOTT: One last thing before we go. I know that you have just written a book, and having written a book myself, it's a lot, a lot of work. And everybody should pay attention to these fantastic endorsements the book has gotten and go read a copy of it. It's a fascinating work, not just about education and your mission, but I think also, it has a bunch of really interesting lessons about how you can ambitiously use AI to solve hard problems. But why did you choose to write a book now?

SAL KHAN: This is the second book that I wrote. The first book, I wrote back in 2011, and I remember when 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 share it with the world." But there's something about 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 in 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.

KEVIN SCOTT: 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. Thank you so much for being with us today.

SAL KHAN: Thank you.

(Applause.)

KEVIN SCOTT: Another really incredibly impactful area where these new AI tools and platforms are going to have an enormous impact is healthcare. I just wanted to share another personal anecdote with you all.

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 a 74-year-old Southern woman and has been suffering from a thyroid condition for 26 years, which is entirely under control. She's been taking medication for it. And just last fall, she had some change in her system, where 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, trying to figure out what was going on.

And the healthcare system in the part of the world where my mom lives is super overburdened. It's not a place where tons of people are moving for economic opportunity. 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 was trying to help my mom navigate this situation, I was looking at how could AI have helped relieve the suffering that she was experiencing? And if doctors everywhere had access to these tools, a lot of what she went through could have been immediately alleviated. If you just take her symptoms that she was presenting and put it into GPT-4, along with her chart, it would immediately say, go get this test, which is going to help find root cause. And then the results of the tests could be input into the same session, which would then give a set of recommendations to doctors about a course of therapy. And if that had happened in her case, what was six visits to the hospital could have been compressed to one.

And I think about this a lot, because 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, who are interacting with one of these resources that is overburdened.

And so, I just want us all to think about, as we're imagining what the set of possibilities are for what we go do with AI, things what Sal is doing with Khan Academy and some of the amazing potential that we have to reduce suffering in the world, and to help make things high quality healthcare more equitable and accessible. It's just super inspiring to think about.

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.

That's where all the interesting stuff is. If you look at the history of platform revolutions, 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 exponentially progressing, it's the only reasonable place to go aim, because if you're aiming somewhere different, the platform is becoming so much more capable and so much cheaper over time, that everything that you have in your imagination that's too expensive to do right now or too fragile, is going to become cheap and robust before you can even blink your eye.

And so, that is really 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 really focus on those phase transitions.

While you all have been out there grinding away, building really extraordinary things over the past year with all of these AI tools that are coming, we've been hard at work trying to make forward progress on our AI platform. We talked a lot about how we're optimizing the current frontier, 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 OpenAI to push that frontier forward.

And we showed this slide at the beginning. There's this really beautiful relationship right now between 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, without mentioning numbers, which is hard to do, to give you all an idea of the scaling of these systems.

In 2020, we built our first AI supercomputer for OpenAI. It's the supercomputing environment that trained GPT-3. And so, we're going to just choose "marine wildlife" as our scale marker. You can think of that system as about as big as a shark.

The next system that we built, scale wise, is about as big as an orca. And that is the system that we delivered in 2022 that trains GPT-4.

The system that we have just deployed is, scale wise, about as big as a whale relative to the shark- sized supercomputer and this orca-sized supercomputer. And it turns out you can build a whole hell of a lot of AI with a whale-sized supercomputer. (Laughter.)

And so, one of the things that I just want everybody to really, really be thinking clearly about, and this is going to be our segue to talking with Sam, is the next sample is coming. 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.

(Applause.)

Good to see you.

SAM ALTMAN: You, too.

KEVIN SCOTT: You are one of the busiest people on the planet.

SAM ALTMAN: Wild week.

KEVIN SCOTT: Yeah, it's a wild week, a wild year, man. But so, I really appreciate you taking time out to chat with us today. I guess what I really wanted to start our conversation about, and I asked you this question last week, is there's just been an extraordinary amount of change over the past year and a half, year. What has been the thing that has surprised you most, particularly relevant to an audience of developers?

SAM ALTMAN: Yeah. I mean, I'm delighted to be here. And obviously, it's 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.

When we put GPT-3 out in the API, some people thought it was cool, but it was narrow (inaudible). And seeing what people have done with GPT-4 and seeing now what's happening with GPT-4.0, even though it's new and hasn't been out that long, it's quite remarkable. I've never seen a technology get adopted so quickly in such a meaningful way, 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.

KEVIN SCOTT: Yeah, and I think what you just said is one of the most important points to me. There's a version of AI that could have existed, that is a bunch of smart people building things at extraordinary scale and then just building it into a bunch of products, where everybody gets to passively use them. The really brilliant thing that you all have done is taken the exact same set of things and decided to make it available to any developer, who's able to sign up for an API key.

SAM ALTMAN: Yeah, we try to be really thoughtful about what makes a good API for this. There's going to 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're doing, any product, any service, we make that very easy.

KEVIN SCOTT: Yeah. And again, I think the progress has been stunning. I think the setup for introducing you onto the stage here was –

SAM ALTMAN: I saw that big blue whale.

KEVIN SCOTT: (Laughter.) Yeah, you're making good use of the whale-sized computer right now. And so, without getting too specific, which we can't be, obviously, what are the category of things that people should be expecting over the next k months?

SAM ALTMAN: The most important thing, and this sounds 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, got much safer, both because 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 we actually are seeming to increase the general capability of the model across the board, that's going to keep happening. And the jump that we have seen in the utility that a model can deliver, with each of those half step jumps in smartness, it's quite significant each time.

As we think about the next model or 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. With GPT-4.0, we were able to bring the price down by half and double the speeds.

New modalities really matter. Voice mode has been actually a genuine surprise for me in how much I like the new voice mode. And when people start integrating that, I think that'll matter. But it's the overall intelligence that will be coming that I think matters the most.

KEVIN SCOTT: You, for a while now, have been one of the most successful startup investors in the world. And now, you are one of the most successful CEOs of one of the most important companies in the world. And so, you've got a room full of developers here. I think there are 5,000 people in the room and there are about 200,000 people online right now. What's your advice to them as they think about how to spend their precious time, given what's happening in the world? What's your advice?

SAM ALTMAN: Two things. No. 1, this is probably the most exciting time to be 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 big opportunities, the ability to build something new and really kind of 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 this is a special time and take advantage of it. This is not the time to delay what you were planning to do or wait for the next thing. This is a special moment and a few years where a lot of stuff is going to happen and a lot of great new things are going to get going.

The second thing also about platform shifts is when the mobile phone revolution started or really got going, 2008-2009, you would see people say, we're a mobile company. We're having a mobile app. And then only a few years later, no one said they were a mobile company, because it was table stakes.

And amazing new technology, which I'm biased, 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 a new thing to do, but you still have to figure out how you're going to build enduring value in whatever you're doing. And it's easy to lose sight of that in the excitement of the gold rush.

KEVIN SCOTT: Yeah. One last thing before we let you go. You and I, and 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, everything from APIs and developer tools to end products. And I think we have accumulated a really interesting volume of experience, experience that's hard to get if you're not doing deployments at this scale.

And I think you just mentioned something that's really, really interesting. Part of the interesting and surprising progression of capabilities of these models means that they're more useful in helping to make AI systems safer. I don't know whether you had some thoughts you wanted to share there as well.

SAM ALTMAN: When we first developed this technology, we spent a lot of time talking about, all right, 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. 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.

When we started this, we’re like, we’ve got this thing, we’ve got this language model. It looks kind of impressive and kind of not. And even then, how are we going to get it aligned? And what does it mean? 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 figuring out policy to how we do the monitoring, that’s a huge amount of work, but it’s necessary to be able to deploy these and use them.

When you take a medicine, you want to know it’s going to be safe. When you use an AI model, you want to know it’s going to be robust and behave the way you want. And I’ve 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, 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 into the world, which we really want to do.

KEVIN SCOTT: Yeah, it’s definitely table stakes. Thank you so much for being with us here today.

SAM ALTMAN: Thank you.

KEVIN SCOTT: I really appreciate your time. It’s awesome to hear from you.

SAM ALTMAN: Awesome.

(Applause.)

KEVIN SCOTT: All right, I think this is all that’s separating you all from the rest of your Build and probably lunch. My very last thing for you all is the following call to action.

Microsoft and our partners, like OpenAI, are spending an extraordinary amount of energy and investing capital at an unprecedented scale, trying to make sure that we are building a genuinely valuable platform, but all we’re doing is building the platform. And 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. You’re the ones who are making all of these things matter. We could build platforms all day, all night, and if you all didn’t have the great ideas, 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.

(Applause.)

END