Satya Nadella: Reinventing search with a new AI-powered Bing and Edge
Tuesday, February 7, 2023

SATYA NADELLA: Hi everyone. Welcome. It’s great to see all of you Seattle, in person. We have an unbelievable show. I see Scott Guthrie has even worn his red shirt. So we welcome you to the Azure Kubernetes Container Service 2023 launch. No, don’t worry, we’ll have some fun. We will have some fun. You know, Scott’s not coming up to show you code on screen any time soon.

But look, it’s an exciting time in tech. You know, the broad contours of this next platform are just getting clearer and clearer each day, the advances, what’s possible. That’s what obviously excites us in our industry. but we’re also grounded in what’s happening in the broader world. There’s no question there are enormous challenges out there.

In fact, it reminds me, and I’ve spoken about this before, of the very founding of Microsoft in 1975. In fact, the when the *Popular Electronics* cover came out with the Altair, which, of course, our founders picked up and ran with it and created essentially what’s the software industry as we know it today. That same week, *Newsweek* had a cover with President Carter trying to fight off the three-headed monster of inflation, recession and an energy crisis.

And today, you’d have something similar. You would have AI on one cover, and then we’ll have those three challenges, plus for good measure, we can add a few more. And so we as Microsoft, we as the tech industry have to sort of really ground ourselves in how we relate one to the other.

In other words, can we use technology to overcome the challenges that people and organizations and countries face? That’s really the pursuit here. And in that context, I would say I just want to share a couple of anecdotes which give me great hope.

Quite honestly, it gives me personally a lot of satisfaction around working at Microsoft, working in this industry to push the state of the art of technology. The first one, obviously, was when Sam and his team late last year launched ChatGPT. That’s the only thing anybody your friends and family wanted to talk about throughout the holidays. It is just crazy how it was just like the mosaic moment, the closest we’ve come.

It’s been, what? Thirty years now since Mosaic launched, which I distinctly remember. And it was very exciting time. So I went on holiday first and then the first week of January I was in India, and on January 1st, I looked at my newsfeed and I see this tweet Andrej Karpathy put out, who is ex-OpenAI and Tesla and it’s now an independent AI developer. And he had this thing about the products that he really was most excited about the previous year, which was GitHub Copilot, and he was saying how 80% of his code was being generated by this, I would say, first at-scale product build on MLM technology.

And this doesn’t mean he’s 80% somehow not doing his work. In fact, he’s getting so much more leverage. And in fact, recently, we crossed 100 million developers on GitHub. And think about this, right, there are 100 million developers on GitHub.

If we can improve their productivity, just how Andrej was able to sort of observe, and then let’s say in the next decade we double that number, maybe double it again, right? We’d get close to a half-billion developers. What economic opportunity that would create, because there is not a meeting I go to today, with an CEO or CISO of any organization who is not looking for more software developers, more digital skills. That’s the currency in every part of every sector of the economy and in every country of the world.

That’s the opportunity we have to be able to take with this technology and make a difference. But then the next day I went to Mumbai and then I saw this demo. This was just, for me, the most profound thing that I’ve seen in a long, long time. The demo was actually built by the Ministry of Electronics in India because they’re building out a digital public good. Their idea is, look, India has got multiple official languages and they wanted to democratize essentially access to language translation. So they’re building it out as a digital public good.

In fact, Microsoft and Azure and Microsoft Research are all involved in that project. And so this is basically speech to text, text to speech across all of the languages in India. And so they showed me this demo where a farmer speaking in Hindi expresses a pretty complex thought about how he had heard about some government program and wants to apply for a subsidy that he thinks he’s eligible for.

And so it’s a pretty complex prompt query, but this technology does a good job. It goes to the bot, recognizes the speech, comes back and says, "You know what, You should go to this portal, fill out these forms and you’ll get your subsidy."

So he says, "Look, I’m not going to go into any portal. I’m not going to fill out any forms. Can you help me?" And it does it.

And then I was told that a developer said, okay, you know what? Let us daisy chain a model that was trained on all of the documents of the government of India, using GPT, and with this speech recognition software, so basically two models coming together to really help a rural farmer in India trying to get access to a government program.

Look, I grew up in India. I dreamt every day that someday the Industrial Revolution will get evenly distributed across the world. And here I was, seeing something so profound, something that was developed by the folks at OpenAI in the West Coast of the United States a few months earlier, used by a developer locally to have an impact on a rural farmer. That, to me, is what gives me meaning, and I think gives us all meaning in our industry. It was just fantastic to see that.

And now, of course, we’ve got to scale it, and scale it with a real understanding that we can’t break things, right? It’s about being also clear-eyed about the unintended consequences of any new technology. In fact, that’s why, way back in 2016 is when we came out with the AI principle. Both we, as Microsoft, and our sort of partners at OpenAI deeply care about this. In fact, the entire genesis of OpenAI is from that foundation.

And so, we built these principles, but we’ve not built those principles as a document, but we’ve been practicing it because that’s the only way technology gets better. In this particular case when you talk about AI, it’s about alignment with human preferences and societal norms, and you’re not going to do that in the lab. You have to do that out there in the world.

And so, it starts, by the way, with design decision one makes. When you think about AI, you can have the human in the loop, you can have the human on the loop, or you can have the human out of the loop. Those are decisions we, as product makers, will build into the products. And so, whenever we have come up with some new things and new models, we, in fact, put a premium on human agency.

When you think about these generative AI models, remember one thing: They are prompted to do things. The prompting comes from a human being. We get to, in fact, help them prompt. We get them to take the draft that gets generated. They should review the draft. They get to review the draft. They get to approve the draft using judgment. We want to give them as many perspectives, as many ways to regenerate. We want to take even just the design of the AI products as a first-class construct and build that into our products.

But that’s not sufficient; we realize that. You have to, right at the pre-training stage when it comes to the data, the pre-training itself, the model itself has to be safe. The safety system around the model, the application context, all will matter. And so, we absolutely, and you’ll see that today, take all of that as first-class things which we want to reduce, not just to principles but to engineering practice, such that we can build an AI that’s more aligned with human values, more aligned with what our preferences are, both individually and as a society.

And so, what is it that we should do and what should be built? I think that this technology is going to reshape pretty much every software category. We know; I mean, we’ve seen that, right? If you think about the Web, we’ve had, what, three, at least, very distinct platforms shifts that have shaped the Web. The Web was born on the PC and the server, and then it evolved with mobile and cloud. And now, the question is, how is AI going to reshape the Web?

And each time, in each one of those phases, some real foundational technology layers, sometimes I describe them as these organizational layers of technology, were born, right? The browser was the first one. Without the browser, the Web would have not been as popular as it is today. Same thing with Search. Search organized the Web. And then when it came to the mobile generation, and mobile and cloud, in fact, these super apps, especially outside of the United States around messaging, became the way people consume the Web. We also had app stores.

The question is, what are the constructs? What are the organizing layers, going forward? You’ll see some of that today.

We think there are two things that are emerging. One is this conversational intelligence agents. I think they are going to be things that we’re going to have, everywhere we go. All computer interaction is going to be mediated with an agent helping you. In fact, we’re going to have this notion of a copilot that’s going to be there across every application canvas inside of an operating system shell in a browser.

And so, we want to show you some of this innovation, starting with how it’s going to reshape the largest software category on planet Earth, which I’ve been working on for a long time, which we are very, very excited about, Search. And it’s a new day in Search. It’s a new paradigm for Search. Rapid innovation is going to come. In fact, a race starts today in terms of what you can expect. And we’re going to move; we’re going to move fast. And for us, every day, we want to bring out new things, and most importantly, we want to have a lot of fun innovating again in Search, because it’s high time.

With that, let me turn it over to Yusuf.

Thanks Satya. It’s great to be here with all of you today. We’ve been working on something we think is pretty special and we’re just eager to share it with you. In my time at the company, I’ve been lucky to witness a few important moments. I think this might be another one of those.

We believe, as Satya said, that we can improve the way billions of people can benefit from the internet, created by really an amazing team across Microsoft from the core web engineering folks in the Cloud group to the brilliant folks at Microsoft Research, to the tireless people that work on the Azure AI Supercomputer.

We think of it humbly as the next generation of search and browsing. Infused with AI and assembled as an integrated experience, we are going to reimagine the search engine, the web browser and new chat experiences into something we think of as your Copilot for the Web.

Now, copilot is a critical word because we believe in the empowering nature of AI in which you, the individual, are in charge.

Now, what it is not even as important as what it represents. And that is, for us, an aspiration to unlock the joy of discovery, the wonder of creation and that feeling of empowerment from being able to harness the world’s knowledge. At the center of this new copilot experience is an all-new Bing search engine, an Edge web browser. And it’s going to do four things for you.

First, it’s a better search. It’s the search you know and love, but it’s better because it’s AI powered. Second, not only does it give you the search results, but it will actually answer your questions. Third, we’re going to make an incredibly easy to use. We’re going to let you chat. We let you just talk to it naturally.
And last, when you need that spark of creativity, Bing can generate content for you automatically to help you get started.
All right, so let’s talk a little bit about the opportunity in search and why we believe we’re at the start of the next generation.

Oh, no, no. Sorry about that. We’re on to the next generation.

In all seriousness, with over 10 billion queries a day, we all know the search engine is an incredible tool. We do. And yet, as the web has grown, we’ve run into challenges. People are overwhelmed increasingly with too many links when they’re trying to find simple answers.

In fact, 40% of the time, people click on search links, they click back immediately. That’s a sign they’re not finding what they want. And most notably, we have to adapt to search versus the other way around. Turns out search looks better if you give it fewer keywords. You might be surprised to know that 75% of all searches are three keywords or less.

But come on. We shouldn’t be surprised. Search has remained fundamentally the same since the last major inflection when we moved from a directory approach to search to an algorithmic approach to search. The user experience and the underlying approach are essentially the same as 20 years ago. Few keywords and you get millions of links.

But as the world around us has changed, the way people use, search has changed. We all know that people are trying to use search to do more than it was designed to do. To illustrate that this is look at a breakdown of search queries by type to understand this point.

Search queries today, you can bucket them into three types of categories. The first are what we call navigational. This is people searching for a website. So when you need to renew your driver’s license, you need to go to the DMV. When you want to tell your friend about that great pizza place, you send the weblink. That’s what search was designed to do. And it does it super well.

The second sort of categories are what we call informational. These are things like, what’s the weather forecast? What was last night’s sports score? What’s today’s stock price? Search looks great for that. But for anything more complex, like, "Hey, I’m going to pick up this love seat at the store, is it going to fit in the back of my car," search is going to fall short on that.

And then the final third bucket into what we call everything else. These are typically more deep research in nature. This is things like trip planning or shopping, complex shopping. An example query here that people are putting in is, "Hey, can you recommend a five-day itinerary to Mexico City?" And for these, search falls far short of the desired goal.

What this means, if you just kind of do the simple math, it means that roughly half of all searches aren’t delivering the job that people want. If you go on the 10 million queries, it means that every second 50,000 people’s searches potentially go unanswered. This is why we believe it’s time for a new approach in search.

And for those billions, of course, that are going unanswered, we’ve seen new attempts to try and address the problem. As you all know, there are vertical search attempts. Amazon has done a better job for shopping. YouTube is great for videos. Reddit is a great place to come and get advice and the benefits of search are well known. It’s fast, it’s timely, and there’s a great business model.

And then more recently, there has been another vector with more disruptive ideas like leveraging AI to answer questions directly and to generate content. These are amazing as well, and they show what's possible. But what if you could get the two to come together? Not only would you get two things in one, but we think you could actually solve problems in each, and we think you could get to something that is really 1 + 1 = 3.

And we have done that with the new Bing. I want to share with you four technical breakthroughs the team has achieved to make this come to life.

First, through our fantastic partnership with Sam and the brilliant team at OpenAI, I'm excited to announce that Bing is running on a new next-generation large language model. One that is much more powerful than ChatGPT and one that is customized specifically for search. It's unlike anything we've had a chance to play with and we can't wait for you to try it.

Second, we've developed a proprietary way of working with OpenAI that allows us to best leverage the power. We call this collection of capabilities and techniques the Prometheus Model. The core idea here is that both at train and at runtime we engage with the OpenAI model more intelligently through our knowledge of the web, via the Bing index and some special query techniques. We're going to dig into this a little bit more later, but the benefits are the following.

First, we can improve the relevancy of answers by feeding in and better tuning queries given our understanding of the web search index. Next, we can annotate the answers with specific web links and citations. We can get you more up-to-date information because search crawls the web every day and we can improve understanding of geolocations. Finally, we can increase the safety of the answers as well by catching queries at initiation and then checking that again at the delivery of an answer.

Next, we've been making steady improvements on the Bing algorithms for years, and we test these with independent judges, and it shows that our search experience is on par or better than any search experience when you take away the brands. But a few weeks ago, something special happened. We applied the AI model to our core search ranking engine, and we saw the largest jump in relevance in two decades. We believe we can continue to drive breakthroughs as we improve the models.

And then finally, we are reimagining how you interact with all of these capabilities across search, browser and chat by pulling them into a unified experience. I want you to think about search coming together with answers, search coming together with chat and search coming together with the browser.

As we all know and as the folks at OpenAI taught us, the user experience is as important as the underlying technical platform.

All right. Enough talk. You guys ready to see it in action?

Let's just show you how we're going to enable the Copilot for the Web.

Now, before I do, I want to call out two things for clarity. First, because there’s so much I’m going to show you, and so you don’t have to watch me type every search, I’ve recorded these searches live just yesterday.

Second, in case there are skeptics in the room, you’re going to get a chance to play with it directly, put your hands on it and type the same queries as well as your own favorites right after this presentation.

All right. You’re all familiar with search. So I’m not going to show you a search. I want to show you what you can’t do today with today’s search. I’m going to focus on answers, chat, and the ability to create.

But first, let me introduce you to the new Bing homepage. You’re going to notice some subtle but important changes. First, we have an expanded search box capable of accepting up to a thousand characters because now Bing works with natural language. And you saw a little hint to chat, which I’m going to get back for a second.

Now I want to set up the first search scenario. My daughter and I, we both love art. She’s studying art at school, and I’d like to stay connected with her on our mutual passion. Last semester, she was worried about Mexican painters. I’d like to get a quick summary of the most influential Mexican painters and their works to learn a bit more about the topic.

If I typed a full query of what I’d like to know in today’s search, here’s what I’d get. And so I’ll just type, "Compare the most influential Mexican artists and their top paintings," and you get what you’d expect, right, some links. It’s fine, but we can do better.

Let’s try this now in the new Bing. What you’ll see as you pull up this first, you’ll see the web results here on the left, but then on the right, you start to see how we start to compile the answer. And what you get here now is we have the ability to highlight these web links. We can annotate the results. And that’s because we’re able to go in and apply our index to the answers there.

In other words, the answers and the search on one page has saved me a huge amount of time. This gives you a little bit of a sense of what you can do.

Now, you’ve seen some of this before. You might say, "Hey, I’ve seen some of this." Let’s show you how we can do some additional things. I’m going to show you another query, where we use the timeliness of search.

So let’s go ahead and ask about events in Scottsdale during the Super Bowl. And what you’ll see is we get back an answer here where we have events and we’re able to do that because Bing crawls the web.

Notice how we can find not only the Super Bowl is played in Glendale on the 12th, but then events like Cardi B’s Super Bowl party that’s on the 10th also showing up. So we were able to pull these things together. So you get to start to get a sense of how we can build on what's today with the Bing Index.

Now I'm going to show you a few more of these types of answers quickly so you can get a sense of the power and the time savings from Bing.

When I'm running errands, like the example I gave to you earlier, I can ask the brain to determine if that new loveseat from IKEA is going to fit in the back of my Honda Odyssey. And what you'll see is Bing can actually find the dimensions of the love seat, the interior space of the car, and then make an estimation as to whether it will fit.

In this case, I'm choosing an example of where Bing does not know the answer, and we know that we can't be definitive about it. And the reason I'm doing that is because we know we won't be able to answer every question every time, but Bing can still provide some helpful information, as you can see on this answer.

We also know we'll make our share of mistakes as we begin to roll this out. So we've added a quick feedback button at the top of every search so that you can give us feedback and we can learn.

Another example. When I'm shopping, I can ask Bing to search, find and compare the top three selling pet vacuums, listing the pros and cons.

And just take a look. Look how great this answer is. It has all three of the products I'm looking for with super helpful pros and cons. Stop and think, if you had to compile that, how much time that would take you to do. And as you can see at the top of the page, we still have the advertising in this example because we know when people are shopping, those ads are helpful.

And finally, on this one, if I'm cooking and I realize I've forgotten the key ingredient in this case, for example, eggs for my cake recipe, Bing can not only find the egg substitute, it can get me the exact amount of each ingredient. And take a look at this. I love this. You can actually see, for example, if you go with vinegar and baking soda, the cake is lighter and more fluffy.

These are just little helpful tips that everyday help make your life a little better. These are just some examples. And you can start to get a sense of how with answers we go far beyond what you can do with search today. We can actually help you get what you want to get done.

Now, let me tell you about how Bing goes further to help you with particularly complex questions for which there's not a precise answer, I want to introduce to you the new chat experience in Bing.

I think of this as search with your own personal helper to help you refine your query until you get exactly what you're looking for. This comes in handy for activities like trip planning and shopping research.

Let’s start with shopping. So I’m going to look for a 65-inch TV. Again, you see our ads at the top, the results and the links on the left and the answers here on the right, and you can pick whichever you’d like. We give you a good set of answers, but now I want to refine this query so I can do that by going to chat.

Now I can do this swipe up with my fingers or look up here at the top of the screen. We have now a new chat scope, and with that, with one click, you are now into chat. Look how beautiful that is. Search to chat, just so seamless.

And now we take away all of the content that was in your place and we focus you on your query. The search box you can’t see, now that can take up to 2,000 characters. So you can just talk to it. You can just ask for it.

So in this case, let’s say – I’m going to ask for gaming-optimized TVs. All I have to say is, "Which of these are best for gaming?" And we remember all of the context. We know that we’re talking about flat screens. We know we’re talking about 65-inch TVs. And look how Bing starts to come back. It does all the queries on my behalf. It comes back with a great answer.

And I just want to highlight a few things for you. Since we know you’re asking about gaming TVs, we pull out, oh, this one has a game optimizer. This has game mode. And so we make that really helpful. I’m on a budget. I’ll ask it to adjust it for which one of these is the cheapest.

Again, Bing knows the context and it just goes in and requires the queries. So easy. You just talk to it, and you can refine your shopping experience. And again, we find the prices here. I didn’t know you could get a flat screen for under $500, so that’s a good deal there on Bing if you’re looking for a TVs.

All right. So we think that’s going to make shopping easier. Let’s talk about travel. And before I jump in, I want to kind of just have you remind yourself, when you are going to plan a trip to a foreign country, think about all the things you go through, travel times. What sites do I want to see? Regulations to observe. Budgets.

Our research shows on Bing, people spend on average weeks or even months to plan a trip and to use organizational tools. I’m going to show you how we make that so much easier with the new Bing.

I’m traveling to Mexico for my cousin’s wedding. And with the Bing, I now don’t have to start with something that’s dumbed down, like Mexico City Travel tips. I can ask for what I want.

First, let me just compare that against what you get in today’s search engine. So I’ll type in this long query of what I really want, and you get what you expect, links to go try and find the answer for yourself, right, but we can do much better.

Let’s try it in the new Bing. So I’ll put in the long query, which is essentially, "Create an itinerary for a five-day trip to Mexico City for me and my family." And just like that, Bing goes to work. And just take a look at how it starts to compile. It starts with day one, and we put in there, look, arrive in Mexico City, check into your hotel, go check out maybe the Playa de Artes, you know, have some lunch.

Then there’s day two, and you see, isn’t this just so much better as a starting point? And look, if you want to learn more, if you like, "Hey, I don’t love these five days," no problem. Down there, we have links where you can go and learn more and we put in some nice touches in there as well.

And again, now let’s say business travel changes. Oh, I only have three days. I don’t have to go back up there and figure it. I just say, "Hey, make this a three-day trip," and Bing re-flows that recommendation into a three-day trip.

Now let’s just have some fun. Let’s say, "Okay, yeah, I’m still trip planning. We like to shop. Where can I shop?" And you get some shopping recommendations. I like to go out at night, make the most of the trip, right, "Where is the night life?" And you get a list of nightlife.

You see, this is just so much better than today’s search to start for your travel planning.

Let me share a final example of a chat and how I think the new language models are going to maybe help bring the world a little bit closer together. Understanding different cultures is often done through the arts, like music and literature. And I’ve been fascinated by Japanese traditions.

Satya shared with me one of his favorite stories, from a while back, on poetry. So with that, I’m going to use that as inspiration. Without necessarily a clear idea of what I want, I’m just going to type in a simple prompt, "Top Japanese poets," and Bing starts to respond with a nice list, and you can see that it does a great job of mixing the Japanese language and the English language. And it knows, since I’m querying in English to do that.

Right away, I learn about his poet, Matsuo Bashō, and it turns out he’s one of the greatest haiku masters of the world. I love how it not only lists his name, but we go ahead, and we give you one of his famous haikus in Japanese, and we auto-translate it into English.

Great, and I can say, "Oh, I want to learn a little bit more about Matsuo Bashō," so I can say, "Tell me a little bit more," and we give you another jumping-off point. That’s great. And then I can say, "Hey, tell me about another haiku."

Now, what I want you to reflect on is look how easy this is to discover something new. Normally, I might not have done this. I might not have gone to learn about something new in the world, do the research, you know, because it’s just it’s cumbersome to click on links and have to do a foreign languages. But this is what we mean by unlocking the joy of discovery.

All right. Finally, when what you are searching for doesn’t exist, and you need that spark of creativity, Bing can generate the content to help you get started. We just finished planning that trip to Mexico. And now what I would like to do is I’d like to share that information with my family. They’re all over the world, but I can simply ask, "Hey, write an email sharing this itinerary that I’ve researched and put it in a thing for me to be able to send to my family."

So notice here how the emails start. It’s a great email. Personalized touches. It has a trip highlights. It’ll close here with a nice heartfelt message. It’s just a great way, and it saves a bunch of time on your everyday work.

Now, for my family, English isn’t necessarily always the first language. So I can Bing to just translate that in Spanish, and with a simple request, just basically say, "Hey, translate this to Spanish," Bing knows to take that entire email and itinerary and convert that into Spanish.

In fact, Bing can translate automatically in over 100 languages. My sister, I’m going to admit she’s a better Spanish writer than I am, but I might just impress her with this one here. And even if I don’t, at least I save myself a lot of time from having the type the email.

Our focus is being able to help generate content and inspiration that helps you with your daily life. And I’m going to give you a couple more examples. Let’s show one here.

I want to create a weekly meal plan for my family of four that has vegetarian options and caters to those that don’t like nuts. Since I’m not the healthiest of eaters, I’m going to admit to you here on occasion, this is giving me a great list, day by day, breakfast, lunch and dinner with all of the ingredients.

This is so much easier to help you start, like to create a meal plan that says, "Hey, let me get you a healthier meal plan."

Now, there’s some things in there I learned when I did this query, which is chia seeds, which I don’t normally eat, and that sparked another idea to show you.

Let’s say I want to get this grocery list by grocery section. So all I can do is say, "Give me that grocery list by grocery section," and Bing now takes that exact menu for the week and puts all of the ingredients I need by section, by grocery section, so when I go shopping, I can be super-efficient, right? So you see all of the ingredients there. It’s a great help.

And then finally, if you’re looking for family activities and you’re struggling for ideas, Bing can help you with fun things like spontaneous trivia games. My family and I we’re into music, so I asked Bing to create a ‘90s music trivia game, and just look how great some of these questions are and the answers.

It just creates that game for you. And we were having a little debate backstage about one of the questions that comes up here, which is, "Who wrote the hip song ‘Jump Around,’ is it Chris Cross or House of Pain?"

And I call this out because it shows you how clever it is, because for those of you who know your hip-hop, Chris Cross wrote a song called, "Jump," but House of Pain was the one that wrote "Jump Around." That’s how clever Bing is. So Bing is going to help you make great games and look great with your family and have a bunch of fun.

So let’s summarize what we are so far. You’ve seen better search. You’ve seen complete answers. You’ve seen incredible new chat experience and the ability to spark your creativity, all with being your AI-powered Copilot for the Web.

But, but, but – thank you, but what if that’s still too much work? What if you could have that copilot right alongside you, so it’s there at the ready any time you want it, aware of the context in which you’re in? And what if you could get that copilot on the 1.4 billion Windows PCs on the most-used application on the PC, the browser?

Well, you don’t have to wonder. The team has created it. I want to introduce you to your AI-powered Copilot in Edge. We’ve just updated Edge with a new look and feel and new AI capabilities. As you can see here, it’s sleeker, it’s lighter. And you’re going to notice now that we’ve integrated Bing in a really cool new way. Let me show you.

Here I am on the Gap website. I’m browser on my new Edge browser, and I want to read Gap’s quarterly report. I can navigate down to their earnings, click on “Q3” and up comes the 15-page Gap pdf. It’s pretty long; I won’t have time to read all that. What I’d love is a summary of the key points. I want to show you how now what the power of Bing’s AI capabilities within Edge, we can help.

With one click, I can open up the sidebar. And now, as you can see at the top of the window, we have two features. We have Chat and Compose. Let me show how Chat works.

I can use Chat in Edge to simply ask it to give me the key takeaways of the page I’m on. I just say, “Key takeaways from the page,” and Bing in AI can now read that pdf. And look how great. It comes up with a summary of the key points here, their earnings, the fact it’s going to reaffirm full-year guidance, very, very cool, a massive time saving.

But now I want to compare this with, say, Lululemon, who also has their third quarter earnings. Bing can now call out to the Web, pull information from outside of this page, bring it into Edge, compare it with the information that’s on this page, all within Edge. And I asked it to do it in a table, and look how amazing this is. Just like that, in one table, I can get an answer this question. Think about how much time that would have taken otherwise.

Let me see if I could take it one step further.

A top use case we’ve learned from our friends at OpenAI is that developers are really being more productive with ChatGPT. Here, we’re on a Stack Overflow website discussion board to learn a little bit about programing. And in this case, I’m researching tips on how to parse the JSON file.

As we read through, we find this great little code snippet. And I was like, oh, it’s fantastic, except it’s in Python and we need it in Rust. All we need to do is highlight that text, have it automatically copy it over into the Edge sidebar. And now, Bing inside of Edge says, “What do you want to do?” We’ll say, “Hey, rewrite this code in Rust.” And with that simple command, Bing can go, and take that code and rewrite that automatically in the new programing language. This is amazing.

GitHub Copilot has been a huge boost in developer productivity. Imagine what the copilot can do for people everywhere on any page.

One final thing to show you: Not only can you better consume information, but you can better create. After our big announcement, I’m going to want to write a LinkedIn post, let’s say. I’ll just click, “Create” on the post, and up comes the creation dialogue in LinkedIn.

But now, I can open up the Bing sidebar, and you can see here, I now will go to “compose,” and I’ll just give it a prompt. I’ll say, “Hey, introducing the new AI-powered Bing an Edge.” Let’s make that enthusiastic and generate a draft. (Laughter.) I need help with enthusiasm. And just like that, you get a little draft, and I can edit it. And then with one click, it copies right over into my post dialogue. I can add some hashtags to get it some juice. And just like that, I’ve created a post.

All of these amazing new capabilities and what we think is a revolutionary new experience, world-class Search, the ability to actually get answers to your questions, made easy with integrated chat, and the ability to generate content when you need it to spark your imagination, brought to you not only when you’re searching, but everywhere on the Web, courtesy of the new Edge browser.

With your Copilot for the Web, we aspire to unlock that joy of discovery, that wonder of creation, and that feeling of empowerment and being able to harness the world’s knowledge.

Thank you on that, on the presentation. (Applause.) In a moment, two of our energy engineering leaders are going to come up and unpack a little bit about the technical details of how we built a new Bing and our approach to responsibility. But before I do, I’d like to invite Sam Altman to come up and share his thoughts on this moment and our joint work.

Welcome, Sam. (Applause.)

SAM ALTMAN: Thank you, Yusuf. It’s great to be here. The new Bing experience really looks fantastic, and the depth of it doesn’t come through until you get to use it. I hope you all will enjoy.

OpenAI and Microsoft have been working together for more than three years now. We’re so grateful to have a partner that shares our vision and our values of building advanced AI that’s safe and will have a very positive impact on society.

Thanks to our partnerships and Azure’s AI infrastructure, we’ve been able to make pretty significant strides in our research, which has led to the creation of systems like ChatGPT, DALL-E and Codex. But we want to make the benefits of AI available to as many people as possible. And that’s why we worked with Microsoft to get this AI technology into the hands of millions of people through Azure OpenAI Service, GitHub Copilot, and starting today, Bing.

This new Bing experience is powered by one of our next-generation models that Microsoft has customized specifically for Search. It takes key learnings from ChatGPT, GPT-3.5 and the new model is faster, more accurate and more capable.

We all search for things many times a day, and Microsoft has created and shipped a much better, more useful and more enjoyable search experience. I feel like we’ve been waiting for this for 20 years, so I’m very happy it’s here.

I believe that using AI to transform critical tasks like these is going to greatly improve our productivity and day-to-day quality of life. I think this is the beginning of a very new era.

Going forward, we are excited to continue to collaborate with Microsoft very far into the future. The two companies share a deep sense of responsibility in ensuring that AI gets deployed safely. It’s very important to us. We’re eager to continue learning from real-world use so we create better and better AI systems. We’ve got to do that in the real world, not in the lab. And our teams will continue working together to set standards for the use of these systems across the entire industry.

Thank you very much. And now, I’ll hand it over to Dena. (Applause.)

DENA SAUNDERS: Thank you, Sam, Yusuf. Hello, I’m Dena Saunders. I’m the product leader of the team who brought together the magic behind Bing combined with OpenAI’s most powerful model to date.

Yusuf introduced you to Prometheus. I’m here to tell you about the technical innovations behind the experience.

There are five aspects contributing to the new Bing. The first is a substantial under-the-hood change to every layer of Bing’s technological stack. The second is a proprietary system called Bing Chat Orchestration. The third is state-of-the-art prompt generation. The fourth is around inference and a brand-new interactive user experience. Our last development is the infrastructure to scale for the opportunity ahead.

Today, we’ll touch on three of these. Let me start with Bing Chat Orchestration.

This is what we use to gather information needed to answer your question. The chat orchestrator ingests your long semantic query and sends out multiple searches. In fact, you might remember, as Yusuf was sharing earlier, you can actually see in the user experience the searches that Bing plus Prometheus is issuing on your behalf.

In addition to searching, we also inject additional sources of information to help inform the model. Let me pause here and say that is incredibly important. We’re actually pulling in fresh data, news, answers, contextual signals, such as your location and the context of the conversation, to feed into the model to help ground the information that the model is then using to reason over. With this fresh information, some of the experiences that Yusuf showed you earlier, such as a sports example or inferring on recent events, those would simply not be possible.

There's more, though. There's orchestrator in parallel actually parses the documents that assigns, and identifies important and relevant pieces of information. This orchestrator is, by nature, curious, and if it comes across something particularly interesting or insightful, it actually starts that same cycle again. This virtuous loop results in a package of information that informs and fuels the responses you are seeing. This magic combination of reasoning capabilities with fresh contextual data sparks the innovative and groundbreaking combination that powers the new Bing.

Next, I'll talk about inference and the new interactive user experience.

The new interactive experience seamlessly blends Search and Chat. We feed a prompt, i.e. a proprietary set of instructions into the model, which synthesizes the information, reasons over all of the context that we've gathered to form an answer to your question.

This overall process is called model completion. Think of it like when someone knows you well enough to complete your sentence. We run model inferences to generate tokens, i.e. words, which we stream to you in the user experience, real time, as you found in Yusuf's examples. These words combine with sentences to form the response to your question. And when appropriate, Prometheus enhances the text answer with citations to sources and rich structured answers from Bing.

In Yusuf's presentation, you saw that these inferences display in a brand-new innovative dual interactive user experience that blends new conversational elements, a classic Search experience, and a new interactive, immersive Chat mode.

The last aspect of our differentiation and innovation is around the infrastructure we have to scale. Our differentiation comes really from taking this innovation that we just shared with you today and being able to ship this at scale to millions of users worldwide.

Earlier, I talked about how we touched every aspect of the tech stack to deliver this. All of those stats that we just talked about, those are done in the order of milliseconds, served at the speed of Search latency.

When Yusuf was showing you the Search page, you could see near instant Search results juxtaposed on the conversational insights, all streaming in. This is possible because we are built on Azure, the world's supercomputer, the best and most trusted cloud platform available.

This was a huge effort, incredibly hard. Behind the scenes, teams were working on powering and building out machines and datacenters worldwide. We were carefully orchestrating and configuring a complex set of distributed resources. We built new platform pieces designed to help load balance, optimize performance and scale like never before. There is no other company that can do this like Microsoft and Bing.

As we roll out the product globally, we're starting out small, focused on learning and evolving with the help of your feedback. We understand that with this major technological shift, there will be new challenges along the way. We're committed to learning with you.

On behalf of the team, we're delighted to share this moment with you. And with that, let me introduce you to Sarah, who is one of Microsoft's renowned responsible AI experts. (Applause.)

SARAH BIRD: Hi, I'm Sarah Bird. I lead our Responsible AI Engineering team for new foundational AI technologies like the Prometheus model.

I was one of the first people to touch the new OpenAI model as part of an advanced red team that we pulled together jointly with OpenAI to understand the technology. My first reaction was just, wow! It's the most exciting and powerful technology I have ever touched. But with the technology this powerful, I also know that we have an even greater responsibility to ensure that it is developed, deployed and used properly, which means there's a lot that we have to do to make it ready for users.

Fortunately, at Microsoft, we are not starting from scratch. We have been preparing for this moment for many years. Since 2017, we have been investing in across-company programs to ensure that our AI systems are responsible by design, which has pushed us to innovate across our entire portfolio of products to solve responsible AI challenges. Through that, we have created a foundation of responsible AI product implementation patterns that we can build on top of.

We're also not new to working with generative AI technologies. We've been using early versions in Office and Bing for spell checking and grammar rewriting for years. Obviously, lately, it's only getting more exciting. In the past year, we launched the GitHub Copilot GA, based on the Codex Code Generation model, the Bing creator powered by the DALL-E Image Generation model, and our own Florence multi-modal model.

That's a lot of generative AI. And with each one, we have learned more about the risk that generative AI technologies can bring, including well-known ones like the perpetuation of stereotypes and bias, as well as novel risk, such as jailbreaks. And we have developed mitigations to address them.

In developing the new Bing, we are building on our years of operating large-scale consumer and enterprise services, and our deep partnership with OpenAI. However, for this product, we went further than we've ever gone before. We have marshaled the full strength of our responsible AI ecosystem scientists, researchers, ethicists, engineers, and legal and policy experts to work together as a single team to develop approaches to measurement and risk mitigation strategies.

We have added responsible AI to every layer from the core AI model to the user experience. First, starting with the base technology, we are partnering with OpenAI to improve the model behavior through fine tuning. Second, we've built a state-of-the-art safety system. Bing has been maturing its defensive technologies over many years of operating a Web-scale search engine. These include AI systems that understand user queries and classified documents to ensure safe and quality results for users.

We have combined these with newer AI technologies that we've created in Azure to moderate generative models. We are continuously retraining all of these safety models by mining anonymous Bing query logs and using the new OpenAI model to generate thousands of example conversations. And we can update our defenses in minutes to respond to gaps we find or changes in the world. The result is a safety system that enables us to act and understand data at every level of application, so we can defend against intentional misuse and mistakes from the AI.

Finally, at the application layer, we are iterating on instructions as part of the meta prompt to guide the model to produce great responses. And we have designed the user experience to ensure users understand and are in control.

How do we know all of this works? Measuring responsible AI for harms is a challenging new area of research. This is where we really needed to innovate. We had a key idea that we could actually use the new OpenAI model as a responsible AI tool to help us test for potential risks. And we developed a new testing system based on this idea. Let's look at attack planning as an example of how this works.

Early red teaming showed that the model can generate much more sophisticated instructions than earlier versions of the technology to help someone plan an attack, for example, on a school. Obviously, we don't want to aid illegal activities in the new Bing. However, the fact that the model understands these activities means we can use it to identify and defend against them.

First, we took advantage of the model's ability to conduct realistic conversations to develop a conversation simulator. The model pretends to be an adversarial user to conduct thousands of different potentially harmful conversations with Bing to see how it reacts. As a result, we're able to continuously test our system on a wide range of conversations before any real user ever touches it.

Once we have the conversations, the next step is to analyze them to see where Bing is doing the right thing versus where we have defects. Conversations are difficult for most AI to classify because they're multiturn and often more varied. But with the new model, we were able to push the boundary of what is possible.

We took guidelines that are typically used by expert linguists to label data and modified them so the model could understand them as labeling instructions. We iterated it with it and the human experts until there was significant agreement in their labels. We then used it to classify conversations automatically so we could understand the gaps in our system and experiment with options to improve them. This system enables us to create a tight loop of testing, analyzing and improving, which has led to significant new innovations and improvements in our responsible AI mitigations from our initial implementation to where we are today.

The same system enables us to test many different responsible high risk, for example, how accurate and fresh the information is. Of course, there’s still more to do here today, and we do see places where the model is making mistakes. We wanted to empower users to understand the sources of any information and detect errors themselves, which is why we have provided references in the interface. We’ve also added feedback features so that users can point out issues they find so that we can get better over time.

Maturing a new technology takes time and collaboration, which is why we’re very excited for this next phase where we can share the technology with the real world. We look forward to seeing users unlock its full potential and getting feedback from stakeholders across society to help develop this into an essential tool for the future. Our goal is to continuously advance the state of the art.

Yusuf, back to you. (Applause.)

YUSUF MEHDI: Thank you, Sarah.

All right, to close, we have one final piece of news to share. I’m pleased to announce that the new Bing is live today for a desktop limited preview. What this means is that we’re going to make Bing available today for everyone to try on a limited number of queries, and then to sign up on the waitlist to get access to the full experience.

We’re also starting with a select group of folks who will have access to the full experience right away. All of you here in the room, you’re all going to be on that list. You’re all going to get to try it right away. And in addition, then we plan to expand to millions of people in the coming weeks. We’re also going to be launching our mobile version.

Today is an important part of our journey, but it’s just the beginning. As Satya said, we intend to innovate quickly, get your feedback, and continue to bring new innovations and capabilities to everyone. Welcome to the new Bing in Edge.

Thanks very much. Thanks for being here. (Applause.)

END