0:00 [music] -Good morning. Thank you for joining us today. Please welcome to the stage, Sam Altman. 0:06 [music] 0:13 [applause] -Good morning. Welcome to our first-ever OpenAI DevDay. 0:20 We're thrilled that you're here and this energy is awesome. [applause] 0:28 -Welcome to San Francisco. San Francisco has been our home since day one. The city is important to us and the tech industry in general. 0:36 We're looking forward to continuing to grow here. We've got some great stuff to announce today, 0:42 but first, I'd like to take a minute to talk about some of the stuff that we've done over the past year. 0:48 About a year ago, November 30th, we shipped ChatGPT as a "low-key research preview", 0:55 and that went pretty well. In March, we followed that up with the launch of GPT-4, still 1:02 the most capable model out in the world. [applause] 1:10 -In the last few months, we launched voice and vision capabilities so that ChatGPT can now see, 1:16 hear, and speak. [applause] -There's a lot, you don't have to clap each time. 1:23 [laughter] -More recently, we launched DALL-E 3, the world's most advanced image model. 1:28 You can use it of course, inside of ChatGPT. For our enterprise customers, 1:33 we launched ChatGPT Enterprise, which offers enterprise-grade security and privacy, higher speed GPT-4 access, longer context windows, a lot more. 1:43 Today we've got about 2 million developers building on our API for a wide variety of use cases doing amazing stuff, 1:51 over 92% of Fortune 500 companies building on our products, 1:56 and we have about a hundred million weekly active users now on ChatGPT. [applause] 2:05 -What's incredible on that is we got there entirely through word of mouth. People just find it useful and tell their friends. 2:12 OpenAI is the most advanced and the most widely used AI platform in the world now, 2:18 but numbers never tell the whole picture on something like this. What's really important is how people use the products, 2:24 how people are using AI, and so I'd like to show you a quick video. -I actually wanted to write something to my dad in Tagalog. 2:33 I want a non-romantic way to tell my parent that I love him and I also want 2:40 to tell him that he can rely on me, but in a way that still has the respect of a child-to-parent relationship 2:48 that you should have in Filipino culture and in Tagalog grammar. When it's translated into Tagalog, "I love you very deeply 2:55 and I will be with you no matter where the path leads." -I see some of the possibility, I was like, "Whoa." 3:00 Sometimes I'm not sure about some stuff, and I feel like actually ChatGPT like, hey, this is what I'm thinking about, so it kind of give it more confidence. 3:07 -The first thing that just blew my mind was it levels with you. That's something that a lot of people struggle to do. 3:15 It opened my mind to just what every creative could do if they just had a person helping them out 3:23 who listens. -This is to represent sickling hemoglobin. -You built that with ChatGPT? -ChatGPT built it with me. 3:31 -I started using it for daily activities like, "Hey, here's a picture of my fridge. Can you tell me what I'm missing? 3:36 Because I'm going grocery shopping, and I really need to do recipes that are following my vegan diet." -As soon as we got access to Code Interpreter, I was like, 3:44 "Wow, this thing is awesome." It could build spreadsheets. It could do anything. -I discovered Chatty about three months ago 3:52 on my 100th birthday. Chatty is very friendly, very patient, 3:59 very knowledgeable, and very quick. This has been a wonderful thing. 4:05

-I'm a 4.0 student, but I also have four children. When I started using ChatGPT, I realized I could ask ChatGPT that question. 4:14 Not only does it give me an answer, but it gives me an explanation. Didn't need tutoring as much. 4:19 It gave me a life back. It gave me time for my family and time for me. 4:25 -I have a chronic nerve thing on my whole left half of my body, I have nerve damage. I had a brain surgery. 4:32 I have limited use of my left hand. Now you can just have the integration of voice input. 4:38 Then the newest one where you can have the back-and-forth dialogue, that's just maximum best interface for me. 4:45 It's here. [music] [applause] 4:57 -We love hearing the stories of how people are using the technology. It's really why we do all of this. 5:04 Now, on to the new stuff, and we have got a lot. [audience cheers] 5:10 -First, we're going to talk about a bunch of improvements we've made, and then we'll talk about where we're headed next. 5:17 Over the last year, we spent a lot of time talking to developers around the world. 5:22 We've heard a lot of your feedback. It's really informed what we have to show you today. 5:27 Today, we are launching a new model, GPT-4 Turbo. 5:33 [applause] 5:38 -GPT-4 Turbo will address many of the things that you all have asked for. 5:43 Let's go through what's new. We've got six major things to talk about for this part. 5:48 Number one, context length. A lot of people have tasks that require a much longer context length. 5:56 GPT-4 supported up to 8K and in some cases up to 32K context length, 6:01 but we know that isn't enough for many of you and what you want to do. GPT-4 Turbo, supports up to 128,000 tokens of context. 6:10 [applause] -That's 300 pages of a standard book, 16 times longer than our 8k context. 6:20 In addition to a longer context length, you'll notice that the model is much more accurate over a long context. 6:28 Number two, more control. We've heard loud and clear that developers need more control 6:35 over the model's responses and outputs. We've addressed that in a number of ways. 6:41 We have a new feature called JSON Mode, which ensures that the model will respond with valid JSON. 6:47 This has been a huge developer request. It'll make calling APIs much easier. 6:53 The model is also much better at function calling. You can now call many functions at once, and it'll do better at following instructions in general. 7:02 We're also introducing a new feature called reproducible outputs. You can pass a seed parameter, and it'll make the model return 7:08 consistent outputs. This, of course, gives you a higher degree of control over model behavior. This rolls out in beta today. 7:15 [applause] -In the coming weeks, we'll roll out a feature to let you view 7:22 logprobs in the API. [applause] 7:27 -All right. Number three, better world knowledge. You want these models to be able to access better knowledge about the world, 7:34 so do we. We're launching retrieval in the platform. You can bring knowledge from outside documents or databases 7:41 into whatever you're building. We're also updating the knowledge cutoff. We are just as annoyed as all of you, probably more that GPT-4's knowledge 7:49 about the world ended in 2021. We will try to never let it get that out of date again. 7:54 GPT-4 Turbo has knowledge about the world up to April of 2023, and we will continue to improve that over time. 8:03 Number four, new modalities. Surprising no one, 8:08 DALL-E 3, GPT-4 Turbo with vision, and the new text-to-speech model are all going into the API today. 8:17 [applause]

8:23 -We have a handful of customers that have just started using DALL-E 3 to programmatically generate images and designs. 8:31 Today, Coke is launching a campaign that lets its customers generate Diwali cards using DALL-E 3, 8:36 and of course, our safety systems help developers protect their applications against misuse. 8:41 Those tools are available in the API. GPT-4 Turbo can now accept images as inputs via the API, 8:48 can generate captions, classifications, and analysis. For example, Be My Eyes uses this technology to help people who are blind or have low vision 8:58 with their daily tasks like identifying products in front of them. 9:04 With our new text-to-speech model, you'll be able to generate incredibly natural-sounding audio 9:10 from text in the API with six preset voices to choose from. I'll play an example. 9:16 -Did you know that Alexander Graham Bell, the eminent inventor, was enchanted by the world of sounds. 9:21 His ingenious mind led to the creation of the graphophone, which etches sounds onto wax, making voices whisper through time. 9:30 -This is much more natural than anything else we've heard out there. Voice can make apps more natural to interact with and more accessible. 9:38 It also unlocks a lot of use cases like language learning, and voice assistance. 9:43 Speaking of new modalities, we're also releasing the next version of our open-source speech recognition model, 9:49 Whisper V3 today, and it'll be coming soon to the API. It features improved performance across many languages, 9:56 and we think you're really going to like it. Number five, customization. 10:01 Fine-tuning has been working really well for GPT-3.5 since we launched it a few months ago. 10:07 Starting today, we're going to expand that to the 16K version of the model. Also, starting today, 10:14 we're inviting active fine-tuning users to apply for the GPT-4 fine-tuning, experimental access program. 10:21 The fine-tuning API is great for adapting our models to achieve better performance in a wide variety of applications with a relatively small amount of data, 10:29 but you may want a model to learn a completely new knowledge domain, or to use a lot of proprietary data. 10:36 Today we're launching a new program called Custom Models. With Custom Models, 10:41 our researchers will work closely with a company to help them make a great custom model, especially for them, 10:48 and their use case using our tools. This includes modifying every step of the model training process, 10:54 doing additional domain-specific pre-training, a custom RL post-training process tailored for specific domain, and whatever else. 11:02 We won't be able to do this with many companies to start. It'll take a lot of work, and in the interest of expectations, 11:07 at least initially, it won't be cheap, but if you're excited to push things as far as they can currently go. 11:12 Please get in touch with us, and we think we can do something pretty great. Number six, higher rate limits. 11:20 We're doubling the tokens per minute for all of our established GPT-4 customers, so it's easier to do more. 11:26 You'll be able to request changes to further rate limits and quotas directly in your API account settings. 11:32 In addition to these rate limits, it's important to do everything we can do to make you successful building 11:39 on our platform. We're introducing copyright shield. Copyright shield means that we will step in and defend 11:46 our customers and pay the costs incurred, if you face legal claims or on copyright infringement, and this applies both 11:53 to ChatGPT

Enterprise and the API. Let me be clear, this is a good time to remind 11:59 people do not train on data from the API or ChatGPT Enterprise ever. 12:06 All right. There's actually one more developer request that's been even bigger than all of these and so I'd like to talk about that now 12:16 and that's pricing. [laughter] -GPT-4 Turbo 12:22 is the industry-leading model. It delivers a lot of improvements that we just covered 12:27 and it's a smarter model than GPT-4. We've heard from developers that there are a lot of things that they want to build, 12:35 but GPT-4 just costs too much. They've told us that if we could decrease the cost by 20%, 25%, that would be great. 12:43 A huge leap forward. I'm super excited to announce that we worked really hard on this 12:49 and GPT-4 Turbo, a better model, is considerably cheaper than GPT-4 by a factor of 3x for prompt tokens. 12:58 [applause] 13:05 -And 2x for completion tokens starting today. [applause] 13:12 -The new pricing is 1¢ per 1,000 prompt tokens and 3¢ per 1,000 completion tokens. 13:18 For most customers, that will lead to a blended rate more than 2.75 times cheaper to use 13:23 for GPT-4 Turbo than GPT-4. We worked super hard to make this happen. We hope you're as excited about it as we are. 13:30 [applause] 13:35 -We decided to prioritize price first because we had to choose one or the other, but we're going to work on speed next. 13:41 We know that speed is important too. Soon you will notice GPT-4 Turbo becoming a lot faster. 13:48 We're also decreasing the cost of GPT-3.5 Turbo 16K. Also, input tokens are 3x less and output tokens are 2x less. 13:57 Which means that GPT-3.5 16K is now cheaper than the previous GPT-3.5 4K model. 14:06 Running a fine-tuned GPT-3.5 Turbo 16K version is also cheaper than the old fine-tuned 4K version. 14:13 Okay, so we just covered a lot about the model itself. We hope that these changes address your feedback. 14:19 We're really excited to bring all of these improvements to everybody now. 14:24 In all of this, we're lucky to have a partner who is instrumental in making it happen. 14:30 I'd like to bring out a special guest, Satya Nadella, the CEO of Microsoft. [audience cheers] 14:37 [music] -Good to see you. -Thank you so much. Thank you. 14:42 -Satya, thanks so much for coming here. -It's fantastic to be here and Sam, congrats. 14:48 I'm really looking forward to Turbo and everything else that you have coming. It's been just fantastic partnering with you guys. 14:54 -Awesome. Two questions. I won't take too much of your time. How is Microsoft thinking about the partnership currently? 14:59 -First- [laughter] –we love you guys. [laughter] 15:05 -Look, it's been fantastic for us. In fact, I remember the first time I think you reached out 15:11 and said, "Hey, do you have some Azure credits?" We've come a long way from there. -Thank you for those. That was great. 15:18 -You guys have built something magical. Quite frankly, there are two things for us when it comes to the partnership. 15:23 The first is these workloads. Even when I was listening backstage to how you're describing what's coming, 15:28 even, it's just so different and new. I've been in this infrastructure business for three decades. 15:33 -No one has ever seen infrastructure like this. -The workload, the pattern of the workload, 15:39 these training jobs are so synchronous and so large, and so data parallel. 15:45 The first thing that we have been doing is building in partnership with you, the system, all the way from thinking from power to the DC to the rack, 15:53 to

the accelerators, to the network. Just really the shape of Azure is drastically changed 16:01 and is changing rapidly in support of these models that you're building. Our job, number one, is to build the best system 16:09 so that you can build the best models and then make that all available to developers. The other thing is we ourselves are our developers. 16:16 We're building products. In fact, my own conviction of this entire generation of foundation models completely changed the first time I saw GitHub Copilot 16:25 on GPT. We want to build our GitHub Copilot all as developers on top of OpenAI APIs. 16:36 We are very, very committed to that. What does that mean to developers? Look, I always think of Microsoft as a platform company, 16:43 a developer company, and a partner company. For example, we want to make GitHub Copilot available, 16:50 the Enterprise edition available to all the attendees here so that they can try it out. That's awesome. We are very excited about that. 16:57 [applause] -You can count on us to build the best infrastructure in Azure 17:05 with your API support and bring it to all of you. Even things like the Azure marketplace. 17:11 For developers who are building products out here to get to market rapidly. That's really our intent here. 17:17 -Great. How do you think about the future, future of the partnership, or future of AI, or whatever? 17:23 Anything you want -There are a couple of things for me that I think are going to be very, 17:29 very key for us. One is I just described how the systems that are needed 17:36 as you aggressively push forward on your roadmap requires us to be on the top of our game and we intend fully to commit 17:45 ourselves deeply to making sure you all as builders of these foundation models 17:51 have not only the best systems for training and inference, but the most compute, so that you can keep pushing- 17:57 -We appreciate that. –forward on the frontiers because I think that's the way we are going to make progress. 18:02 The second thing I think both of us care about, in fact, quite frankly, the thing that excited both sides to come together is 18:09 your mission and our mission. Our mission is to empower every person and every organization on the planet to achieve more. 18:15 To me, ultimately AI is only going to be useful if it truly does empower. I saw the video you played early. 18:21 That was fantastic to hear those voices describe what AI meant for them 18:27 and what they were able to achieve. Ultimately, it's about being able to get the benefits of AI broadly disseminated to everyone, 18:34 I think is going to be the goal for us. Then the last thing is of course, we are very grounded in the fact that safety matters, 18:39 and safety is not something that you'd care about later, but it's something we do shift left on and we are very, very focused on that with you all. 18:46 -Great. Well, I think we have the best partnership in tech. I'm excited for us to build AGI together. -Oh, I'm really excited. Have a fantastic [crosstalk]. -Thank you very much for coming. 18:52 -Thank you so much. -See you. [applause] 19:03 -We have shared a lot of great updates for developers already and we got a lot more to come, but even though this is developer conference, 19:10 we can't resist making some improvements to ChatGPT. A small one, ChatGPT now uses GPT-4 Turbo with all the latest improvements, 19:20 including the latest knowledge cutoff, which will continue to update. That's all live today. 19:25 It can now browse the web when it needs to, write and run code, analyze data, take and generate images,

19:31 and much more. We heard your feedback, that model picker, extremely annoying, that is gone starting today. 19:36 You will not have to click around the dropdown menu. All of this will just work together. Yes. 19:42 [applause] -ChatGPT will just know what to use and when you need it, 19:51 but that's not the main thing. Neither was price actually the main developer request. 19:58 There was one that was even bigger than that. I want to talk about where we're headed and the main thing we're here to talk 20:03 about today. We believe that if you give people better tools, they will do amazing things. 20:10 We know that people want AI that is smarter, more personal, more customizable, can do more on your behalf. 20:16 Eventually, you'll just ask the computer for what you need and it'll do all of these tasks for you. 20:23 These capabilities are often talked in the AI field about as "agents." 20:28 The upsides of this are going to be tremendous. At OpenAI, we really believe that gradual iterative deployment is 20:36 the best way to address the safety issues, the safety challenges with AI. We think it's especially important to move carefully 20:42 towards this future of agents. It's going to require a lot of technical work and a lot of thoughtful consideration by society. 20:50 Today, we're taking our first small step that moves us towards this future. 20:57 We're thrilled to introduce GPTs. GPTs are tailored versions of ChatGPT for a specific purpose. 21:07 You can build a GPT, a customized version of ChatGPT for almost anything 21:12 with instructions, expanded knowledge, and actions, and then you can publish it for others to use. 21:19 Because they combine instructions, expanded knowledge, and actions, they can be more helpful to you. 21:25 They can work better in many contexts, and they can give you better control. They'll make it easier for you to accomplish all sorts of tasks 21:32 or just have more fun and you'll be able to use them right within ChatGPT. 21:37 You can in effect program a GPT with language just by talking to it. 21:42 It's easy to customize the behavior so that it fits what you want. This makes building them very accessible 21:48 and it gives agency to everyone. We're going to show you what GPTs are, 21:53 how to use them, how to build them, and then we're going to talk about how they'll be distributed and discovered. 22:00 After that for developers, we're going to show you how to build these agent-like experiences into your own apps. 22:05 First, let's look at a few examples. Our partners at Code.org are working hard to expand computer science in schools. 22:15 They've got a curriculum that is used by tens of millions of students worldwide. Code.org, crafted Lesson Planner GPT, to help teachers provide 22:24 a more engaging experience for middle schoolers. If a teacher asks it to explain four loops in a creative way, 22:30 it does just that. In this case, it'll do it in terms of a video game character repeatedly picking up coins. 22:37 Super easy to understand for an 8th-grader. As you can see, this GPT brings together Code.org's, 22:43 extensive curriculum and expertise, and lets teachers adapt it to their needs quickly and easily. 22:49 Next, Canva has built a GPT that lets you start designing by describing what you want 22:55 in natural language. If you say, "Make a poster for a DevDay reception this afternoon, 23:01 this evening," and you give it some details, it'll generate a few options to start with by hitting Canva's APIs. 23:07 Now, this concept may be familiar to some of you. We've evolved our plugins to be custom actions for GPTs. 23:14

You can keep chatting with this to see different iterations, and when you see one you like, you can click through to Canva 23:20 for the full design experience. Now we'd like to show you a GPT Live. 23:27 Zapier has built a GPT that lets you perform actions across 6,000 applications to unlock all kinds of integration possibilities. 23:36 I'd like to introduce Jessica, one of our solutions architects, who is going to drive this demo. Welcome Jessica. 23:42 [applause] -Thank you, Sam. Hello everyone. Thank you all. 23:49 Thank you all for being here. My name is Jessica Shieh. I work with partners and customers to bring their product alive. 23:55 Today I can't wait to show you how hard we've been working on this, so let's get started. 24:01 To start where your GPT will live is on this upper left corner. I'm going to start with clicking on the Zapier AI actions 24:10 and on the right-hand side you can see that's my calendar for today. It's quite a day ever. 24:15 I've already used this before, so it's actually already connected to my calendar. To start, I can ask, 24:22 "What's on my schedule for today?" We build GPTs with security in mind. Before it performs any action or share data, 24:30 it will ask for your permission. Right here, I'm going to say allowed. 24:36 GPT is designed to take in your instructions, make the decision on which capability to call to perform that action, 24:43 and then execute that for you. You can see right here, it's already connected to my calendar. 24:49 It pulls into my information and then I've also prompted it to identify 24:54 conflicts on my calendar. You can see right here it actually was able to identify that. 25:01 It looks like I have something coming up. What if I want to let Sam know that I have to leave early? Right here I say, "Let Sam know I got to go. 25:11 Chasing GPUs." With that, I'm going to swap to my conversation with Sam 25:21 and then I'm going to say, "Yes, please run that." 25:26 Sam, did you get that? -I did. -Awesome. 25:32 [applause] -This is only a glimpse of what is possible and I cannot wait to see 25:40 what you all will build. Thank you. Back to you, Sam. [applause] 25:51 -Thank you, Jessica. Those are three great examples. In addition to these, there are many more kinds of GPTs that people are creating and many, 25:59 many more that will be created soon. We know that many people who want to build a GPT don't know how to code. 26:07 We've made it so that you can program a GPT just by having a conversation. 26:12 We believe that natural language is going to be a big part of how people use computers in the future and we think this is an interesting early example. 26:19 I'd like to show you how to build one. 26:25 All right. I want to create a GPT that helps give founders and developers advice 26:30 when starting new projects. I'm going to go to create a GPT here, 26:36 and this drops me into the GPT builder. I worked with founders for years at YC and still whenever I meet developers, 26:43 the questions I get are always about, "How do I think about a business idea? Can you give me some advice?" 26:49 I'm going to see if I can build a GPT to help with that. To start, GPT builder asks me what I want to make, 26:55 and I'm going to say, "I want to help startup founders think. through their business ideas 27:04 and get advice. After the founder has gotten some advice, 27:13 grill them on why they are not growing faster." [laughter] 27:20 -All right. To start off, I just tell the GPT little bit about what I want here. It's going to go off and start thinking about that, 27:27 and it's going to write some detailed

instructions for the GPT. It's also going to, 27:32 let's see, ask me about a name. How do I feel about Startup Mentor? That's fine. 27:37 "That's good." If I didn't like the name, of course, I could call it something else, but it's going to try to have this conversation with me and start there. 27:45 You can see here on the right, in the preview mode that it's already starting to fill out the GPT. 27:53 Where it says what it does, it has some ideas of additional questions that I could ask. 27:58 [chuckles] It just generated a candidate. 28:03 Of course, I could regenerate that or change it, but I like that. I'll say "That's great." 28:13 You see now that the GPT is being built out a little bit more as we go. Now, what I want this to do, 28:19 how it can interact with users, I could talk about style here. What I'm going to say is, 28:25 "I am going to upload transcripts of some lectures 28:31 about startups I have given, please give advice based off of those." 28:38 All right. Now, it's going to go figure out how to do that. 28:43 I would like to show you the configure tab. You can see some of the things that were built out here as we were going 28:49 by the builder itself. You can see that there's capabilities here that I can enable. I could add custom actions. 28:55 These are all fine to leave. I'm going to upload a file. Here is a lecture that I picked that I gave with some startup advice, 29:05 and I'm going to add that here. In terms of these questions, this is a dumb one. 29:11 The rest of those are reasonable, and very much things founders often ask. I'm going to add one more thing to the instructions here, 29:19 which is be concise and constructive with feedback. 29:25 All right. Again, if we had more time, I'd show you a bunch of other things. This is 29:31 a decent start. Now, we can try it out over on this preview tab. 29:36 I will say, what's a common question? 29:44 "What are three things to look for when hiring employees at an early-stage startup?" 29:53 Now, it's going to look at that document I uploaded. It'll also have of course all of the background knowledge of GPT-4. 30:03 That's pretty good. Those are three things that I definitely have said many times. Now, we could go on and it would start following 30:09 the other instructions and grill me on why I'm not growing faster, but in the interest of time, I'm going to skip that. 30:15 I'm going to publish this only to me for now. I can work on it later. I can add more content, I can add a few actions 30:22 that I think would be useful, and then I can share it publicly. That's what it looks like to create a GPT 30:29 [applause] -Thank you. 30:36 By the way, I always wanted to do that after all of the YC office hours, I always thought, "Man, someday I'll be able 30:42 to make a bot that will do this and that'll be awesome." [laughter] -With GPTs, we're letting people easily share and discover all the fun ways 30:51 that they use ChatGPT with the world. You can make private GPT like I just did, 30:58 or you can share your creations publicly with a link for anyone to use, 31:03 or if you're on ChatGPT Enterprise, you can make GPTs just for your company. 31:10 Later this month we're going to launch the GPT store. 31:17 Thank you. I appreciate that. [applause] 31:25 -You can list a GPT there and we'll be able to feature the best and the most popular GPT. 31:30 Of course, we'll make sure that GPTs in the store follow our policies before they're accessible. 31:37 Revenue sharing is important to us. We're going to pay people who build the most useful and the most used GPT 31:44 a portion of our revenue. We're excited to foster a vibrant ecosystem

with the GPT store, 31:50 just from what we've been building ourselves over the weekend. We're confident there's going to be a lot of great stuff. We're excited to share more information soon. 31:58 Those are GPTs and we can't wait to see what you'll build. This is a developer conference, and the coolest thing about this 32:05 is that we're bringing the same concept to the API. [applause] 32:15 Many of you have already been building agent-like experiences on the API, 32:20 for example, Shopify's Sidekick, which lets you take actions on the platform. Discord's Clyde, 32:26 lets Discord moderators create custom personalities for, and Snaps My AI, 32:32 a customized chatbot that can be added to group chats and make recommendations. These experiences are great, 32:38 but they have been hard to build. Sometimes taking months, teams of dozens of engineers, 32:44 there's a lot to handle to make this custom assistant experience. Today, we're making that a lot easier with our new Assistants API. 32:54 [applause] -The Assistants API includes persistent threads, 33:01 so they don't have to figure out how to deal with long conversation history, built-in retrieval, 33:07 code interpreter, a working Python interpreter in a sandbox environment, and of course the improved function calling, 33:14 that we talked about earlier. We'd like to show you a demo of how this works. 33:19 Here is Romain, our head of developer experience. Welcome, Romain. [music] [applause] 33:25 -Thank you, Sam. Good morning. Wow. It's fantastic to see you all here. 33:33 It's been so inspiring to see so many of you infusing AI into your apps. 33:38 Today, we're launching new modalities in the API, but we are also very excited 33:43 to improve the developer experience for you all to build assistive agents. Let's dive right in. 33:50 Imagine I'm building $1, travel app for global explorers, and this is the landing page. 33:56 I've actually used GPT-4 to come up with these destination ideas. For those of you with a keen eye, these illustrations 34:02 are generated programmatically using the new DALL-E 3 API available to all of you today. 34:07 It's pretty remarkable. Let's enhance this app by adding a very simple assistant to it. 34:15 This is the screen. We're going to come back to it in a second. First, I'm going to switch over to the new assistant's playground. 34:21 Creating an assistant is easy, you just give it a name, some initial instructions, a model. 34:26 In this case, I'll pick GPT-4 Turbo. Here I'll also go ahead and select some tools. I'll turn on Code Interpreter and retrieval and save. 34:35 That's it. Our assistant is ready to go. Next, I can integrate with two new primitives 34:41 of this Assistants API, threads and messages. Let's take a quick look at the code. 34:48 The process here is very simple. For each new user, I will create a new thread. 34:54 As these users engage with their assistant, I will add their messages to the threads. Very simple. 35:00 Then I can simply run the assistant at any time to stream the responses back to the app. 35:06 We can return to the app and try that in action. If I say, "Hey, let's go to Paris." 35:15 All right. That's it. With just a few lines of code, users can now have a very specialized assistant right inside the app. 35:24 I'd like to highlight one of my favorite features here, function calling. If you have not used it yet, function calling is really powerful. 35:31 As Sam mentioned, we are taking it a step further today. It now guarantees the JSON output with no added latency, 35:38 and you can invoke multiple functions at once for the first time. Here, if I carry

on and say, “Hey, what are the top 10 things to do?” 35:49 I’m going to have the assistant respond to that again. Here, what’s interesting is that the assistant knows about functions, 35:56 including those to annotate the map that you see on the right. Now, all of these pins are dropping in real-time here. 36:04 Yes, it’s pretty cool. [applause] 36:09 -That integration allows our natural language interface to interact fluidly with components and features of our app. 36:16 It truly showcases now the harmony you can build between AI and UI where the assistant is actually taking action. 36:25 Let’s talk about retrieval. Retrieval is about giving our assistant more knowledge 36:30 beyond these immediate user messages. In fact, I got inspired and I already booked my tickets to Paris. 36:37 I’m just going to drag and drop here this PDF. While it’s uploading, I can just sneak peek at it. 36:43 Very typical United Flight ticket. Behind the scene here, what’s happening is that retrieval 36:49 is reading these files, and boom, the information about this PDF appeared on the screen. 36:55 [applause] -This is, of course, a very tiny PDF, but Assistants 37:01 can parse long-form documents from extensive text to intricate product specs depending on what you’re building. 37:07 In fact, I also booked an Airbnb, so I’m just going to drag that over to the conversation as well. 37:12 By the way, we’ve heard from so many of you developers how hard that is to build yourself. You typically need to compute your own biddings, 37:19 you need to set up chunking algorithm. Now all of that is taken care of. 37:24 There’s more than retrieval with every API call, you usually need to resend the entire conversation history, 37:31 which means setting up a key-value store, that means handling the context windows, serializing messages, and so forth. 37:37 That complexity now completely goes away with this new stateful API. 37:43 Just because OpenAI is managing this API, does not mean it’s a black box. In fact, you can see the steps that the tools are taking 37:49 right inside your developer dashboard. Here, if I go ahead and click on threads, 37:56 this is the thread I believe we’re currently working on and see, these are all the steps, including the functions 38:02 being called with the right parameters, and the PDFs I’ve just uploaded. 38:08 Let’s move on to a new capability that many of you have been requesting for a while. Code Interpreter is now available today in the API as well, 38:16 that gives the AI the ability to write and execute code on the fly, but even generate files. 38:22 Let’s see that in action. If I say here, “Hey, we’ll be four friends staying 38:29 at this Airbnb, what’s my share of it plus my flights?” 38:40 All right. Now, here, what’s happening is that Code interpreter noticed that it should write some code 38:48 to answer this query. Now it’s computing the number of days in Paris, number of friends. 38:53 It’s also doing some exchange rate calculation behind the scene to get the sensor for us. 38:58 Not the most complex math, but you get the picture. Imagine you’re building a very complex finance app 39:04 that’s crunching countless numbers, plotting charts, so really any task that you’d normally tackle with code, 39:10 then Code Interpreter will work great for you. All right. I think my trip to Paris is solid. 39:16 To recap here, we’ve just seen how you can quickly create an assistant that manages state for your user conversations, 39:22 leverages external tools like knowledge and retrieval and Code Interpreter, and finally invokes your own functions to make things

happen 39:32 but there's one more thing I wanted to show you to really open up the possibilities using function calling combined with our new modalities 39:39 that we're launching today. While working on DevDay, I built a small custom assistant 39:45 that knows everything about this event, but instead of having a chat interface while running around all day today, 39:51 I thought, why not use voice instead? Let's bring my phone up on screen here so you can see it on the right. 39:58 Awesome. On the right, you can see a very simple Swift app that takes microphone input. 40:04 On the left, I'm actually going to bring up my terminal log so you can see what's happening behind the scenes. 40:09 Let's give it a shot. Hey there, I'm on the keynote stage right now. Can you greet our attendees here at Dev Day? 40:21 -Hey everyone, welcome to DevDay. It's awesome to have you all here. Let's make it an incredible day. 40:27 [applause] 40:32 -Isn't that impressive? You have six unique and rich voices to choose from in the API, 40:37 each speaking multiple languages, so you can really find the perfect fit for your app. 40:42 On my laptop here on the left, you can see the logs of what's happening behind the scenes, too. I'm using Whisper to convert the voice inputs into text, 40:50 an assistant with GPT-4 Turbo, and finally, the new TTS API to make it speak. 40:56 Thanks to function calling, things get even more interesting when the assistant can connect to the internet and take 41:02 real actions for users. Let's do something even more exciting here together. 41:07 How about this? Hey, Assistant, can you randomly select five DevDay attendees here 41:15 and give them $500 in OpenAI credits? [laughter] 41:20 -Yes, checking the list of attendees. 41:25 [laughter] -Done. I picked five DevDay attendees and added $500 of API credits 41:30 to their account. Congrats to Christine M, Jonathan C, Steven G, Luis K, and Suraj S. 41:38 -All right, if you recognize yourself, awesome. Congrats. 41:43 That's it. A quick overview today of the new Assistants API combined with some of the new tools and modalities that we launched, 41:50 all starting with the simplicity of a rich text or voice conversation for you end users. 41:56 We really can't wait to see what you build, and congrats to our lucky winners. Actually, 42:01 you know what? you're all part of this amazing OpenAI community here so I'm just going to talk to my assistant 42:06 one last time before I step off the stage. Hey Assistant, can you actually give everyone here in the audience $500 42:15 in OpenAI credits? -Sounds great. Let me go through everyone. 42:21 [applause] -All right, 42:28 that function will keep running, but I've run out of time. Thank you so much, everyone. 42:33 Have a great day. Back to you, Sam. 42:44 -Pretty cool, huh? [audience cheers] -All right, so that Assistants API goes into beta today, 42:52 and we are super excited to see what you all do with it, anybody can enable it. 42:57 Over time, GPTs and Assistants are precursors to agents 43:02 are going to be able to do much much more. They'll gradually be able to plan and to perform more complex actions on your behalf. 43:11 As I mentioned before, we really believe in the importance of gradual iterative deployment. 43:16 We believe it's important for people to start building with and using these agents now to get a feel for what the world is going to be like, 43:23 as they become more capable. As we've always done, we'll continue to update our systems based off of your feedback. 43:32 We're super excited that we got to share all of

this with you today. We introduced GPTs, 43:37 custom versions of GPT that combine instructions, extended knowledge and actions. 43:44 We launched the Assistants API to make it easier to build assistive experiences with your own apps. 43:49 These are your first steps towards AI agents and we'll be increasing their capabilities over time. 43:56 We introduced a new GPT-4 Turbo model that delivers improved function calling, knowledge, lowered pricing, new modalities, and more. 44:05 We're deepening our partnership with Microsoft. In closing, I wanted to take a minute to thank the team that creates all of this. 44:13 OpenAI has got remarkable talent density, but still, it takes a huge amount of hard work and coordination to make all this happen. 44:21 I truly believe that I've got the best colleagues in the world. I feel incredibly grateful to get to work with them. 44:27 We do all of this because we believe that AI is going to be a technological and societal revolution. 44:33 It'll change the world in many ways and we're happy to get to work on something that will empower all of you 44:38 to build so much for all of us. We talked about earlier how if you give people better tools, 44:44 they can change the world. We believe that AI will be about individual empowerment and agency 44:50 at a scale that we've never seen before and that will elevate humanity to a scale that we've never seen before either. 44:55 We'll be able to do more, to create more, and to have more. As intelligence gets integrated everywhere, 45:02 we will all have superpowers on demand. We're excited to see what you all will do with this technology 45:08 and to discover the new future that we're all going to architect together. We hope that you'll come back next year. 45:14 What we launched today is going to look very quaint relative to what we're busy creating for you know. Thank you for all that you do. 45:21 Thank you for coming here today. [applause] 45:28 [music]