OpenAI DevDay, Opening Keynote

출처: 리뷰인사이트 SaaS 블로그 (https://reviewinsight.blog/)

GPT 4.0 요약

이 영상은 샘 알트먼이 샌프란시스코에서 열린 OpenAI 데브데이에 참석한 참석자들을 환영하며 그들의 참석과 열정에 감사를 표하는 것으로 시작됩니다. 이어서 그는 지난 한 해 동안 OpenAI가 이룬 성과를 돌아보며 ChatGPT, GPT-4의 출시와 음성 및 비전 기능, DALL-E 3의 도입을 강조합니다. 알트먼은 OpenAI API를 중심으로 한 개발자 커뮤니티의 성장, 주요 기업의 광범위한 제품 사용, 유기적으로 확보한 상당한 사용자 기반에 대해 언급합니다.

이어서 언어 번역, 창의력 향상, 식단 계획, 코딩, 심지어 100세 노인을 위한 말벗 등 삶의 다양한 측면에 OpenAI의 기술을 접목한 사용자들의 사용 후기를 공유합니다. 이러한 개인적인 사례는 교육과 가족 시간을 지원하는 것부터 장애에 대처하는 것까지 AI가 사람들의 일상 생활에 실질적인 영향을 미친 방법을 보여줍니다.

이어서 알트먼은 크게 늘어난 컨텍스트 길이(최대 128,000토큰), 향상된 모델 정확도, 보다 구조화된 응답과 일관된 결과를 위한 재현 가능한 출력을 위한 JSON 모드와 같은 새로운 기능을 제공하는 GPT-4 Turbo를 포함한 새로운 업데이트 및 기능을 발표합니다. 또한 2023년 4월까지의 정보

로 향상된 세계 지식, DALL-E 3의 새로운 양식, 이미지 인식 기능, 고급 텍스트 음성 변환 모델도 있습니다. 또한 다양한 언어에 걸쳐 향상된 음성 인식을 제공하는 Whisper V3의 출시가 예정되어 있음을 언급합니다.

마지막으로 알트먼은 특정 기업의 요구에 맞는 맞춤형 모델 개발과 미세 조정을 가능하게 하는 사용자 지정 옵션에 대해 설명하며, 다양한 사용 사례에 맞춰 AI 모델을 더욱 적응력 있고 개인화할 수 있도록 지원하겠다는 OpenAI의 의지를 밝힙니다.

새로운 서비스를 많은 기업과 함께 구현하려면 많은 노력이 필요하고 초기에는 비용이 많이 들 수 있다는 점을 인정하는 것으로 소개가 시작되었습니다. 하지만 현재 기술의 한계를 뛰어넘는 데 관심이 있는 분들의 연락을 기다리고 있습니다.

이번 발표에는 기존 GPT-4 고객의 분당 토큰 전송량 한도를 두 배로 늘려 운영을 더 쉽게 확장할 수 있도록 하는 내용이 포함되어 있습니다. 이러한 한도는 API 계정 설정에서 직접 조정할 수 있습니다.

또한, 저작권 침해에 대한 법적 청구로부터 고객을 보호하고 비용을 부담하는 '저작권 실드'를 도입하여 ChatGPT 엔터프라이즈와 API 모두에 적용합니다. API 또는 ChatGPT Enterprise의 데이터로 훈련하지 말라는 알림이 있습니다.

주요 개발 사항으로는 GPT-4보다 더 스마트하고 훨씬 저렴한 업계 최고의 모델인 GPT-4 터보가 발표되었으며, 프롬프트 토큰의 경우 3배, 완료 토큰의 경우 2배의 비용 절감 효과가 있습니다. 새로운 가격은 프롬프트 토큰 1,000개당 1센트, 완료 토큰 1,000개당 3센트로 책정되어 GPT-4 터보의 가격을 더욱 저렴하게 책정하는 것을 목표로 합니다.

이 발표에서는 향후 GPT-4 터보의 속도 향상과 GPT-3.5 터보 16K의 비용 절감에 대해서도 언급했습니다.

이번 행사에는 Microsoft의 CEO인 사티아 나델라가 참석하여 양사 간의 파트너십을 발표할 예정이어서 기대가 모아지고 있습니다. 이번 파트너십을 기념하기 위해 Microsoft는 Azure를 통해 이러한 AI 모델에 대한 인프라 지원을 제공합니다.

이 파트너십의 미래에는 AI를 통해 사람들의 역량을 강화하고, AI의 이점을 널리 알리며, 처음부터 안전에 초점을 맞추겠다는 약속이 포함되어 있습니다.

마지막으로 최신 지식 업데이트, 웹 검색 기능, 모델 선택기 제거, 보다 원활한 사용자 경험 제공을 위한 GPT-4 터보 통합 등 ChatGPT의 개선 사항이 강조되었습니다.

전반적으로 기능 확장, 사용자 경험 개선, 기술 접근성 및 경제성 향상에 초점을 맞추고 있으며,

앞으로 나아가기 위한 핵심 요소로 Microsoft와의 파트너십을 강조하고 있습니다.

OpenAI는 AI가 더욱 지능화되고 개인화되어 사용자를 위해 작업을 수행하는 에이전트 역할을 할 수 있는 잠재력을 믿습니다.

점진적이고 반복적인 배포는 특히 AI 에이전트를 만드는 방향으로 나아가고 있는 현 시점에서 가장 안전한 AI 개발 접근법으로 간주됩니다.

지침, 확장된 지식 및 작업 수행 기능으로 사용자 지정할 수 있는 GPT(맞춤형 ChatGPT 버전)가 도입되었습니다.

이러한 GPT는 상황에 맞게 더욱 유용하도록 설계되어 사용자가 더 잘 제어하고 다양한 작업이나 엔터테인먼트 요구 사항을 단순화할 수 있습니다.

사용자는 대화를 통해 GPT를 프로그래밍할 수 있으므로 누구나 자신만의 GPT를 만들고 사용자 지정할 수 있습니다.

GPT의 예로는 프로그래밍 개념을 가르치기 위한 Code.org의 레슨 플래너 GPT, Canva의 디자인 어시스턴트 GPT, 6,000개 이상의 애플리케이션과 통합되어 다양한 자동 작업을 수행할 수 있는 Zapier의 GPT 등이 있습니다.

업로드한 강의 녹취록을 기반으로 창업자와 개발자에게 조언을 제공하는 '스타트업 멘토' GPT를 만드는 방법을 보여드리며, GPT를 만드는 과정을 시연합니다.

GPT는 비공개로 게시하거나, 공개적으로 공유하거나, 기업 환경 내에서 제한할 수 있습니다.
이 새로운 방향은 사용자 커스터마이징과 다양하고 구체적인 요구 사항을 충족하기 위한 GPT 생성을 강조하여 프로그래밍 지식이 없는 사람들도 AI 도구에 더 쉽게 접근할 수 있도록 합니다.

플랫폼의 정책을 준수하는 다양한 GPT 모델을 선보이는 새로운 GPT 스토어가 출시될 예정입니다.
인기 있고 유용한 GPT에 대해 개발자에게 보상을 제공하는 수익 공유 모델이 도입될 예정입니다.

사용자 지정 어시스턴트 경험 생성을 간소화하기 위해 어시스턴트 API가 도입됩니다. 영구 스레드, 내장 검색, Python 코드 인터프리터, 향상된 함수 호출과 같은 기능을 제공합니다.

데모에서는 새로운 API로 어시스턴트를 만드는 과정을 보여줌으로써 통합의 용이성과 자연어 인터페이스에서 직접 함수를 호출할 수 있는 기능을 시연합니다.

이 API에는 PDF와 같은 문서를 구문 분석하는 새로운 도구와 대화 기록 관리의 복잡성을 없애주는 상태 저장 방식이 포함되어 있습니다.

개발자는 개발자 대시보드에서 도구의 단계를 확인할 수 있습니다.

코드 인터프리터를 사용할 수 있어 AI가 코드를 작성하고 실행할 수 있어 복잡한 애플리케이션에 유용할 수 있습니다.

다양한 언어로 된 여러 음성 중에서 선택할 수 있는 기능을 포함하여 음성 상호작용을 지원하는 새로운 방식이 시연됩니다.

전반적으로 개발자 경험과 AI 기능과 사용자 인터페이스 간의 원활한 통합에 초점을 맞춘 AI 기반 앱 제작의 발전된 점을 강조합니다.

특히 음성-텍스트 변환을 위한 Whisper, 새로운 어시스턴트 API, 텍스트 음성 변환(TTS) API를 언급하고 있습니다. 샘 알트먼은 AI 어시스턴트가 데브데이 참석자에게 OpenAI 크레딧을 할당하는 등 인터넷과 연결된 작업을 수행할 수 있는 기능을 강조합니다. 샘 알트먼은 복잡한 작업을 계획하고 수행할 수 있는 자율 에이전트의 선구자로서 GPT와 어시스턴트의 점진적인 발전에 대해 간략하게 설명합니다.

샘 알트먼은 반복 배포의 중요성, 명령어, 확장된 지식, 액션 기능을 결합한 GPT 변형의 도입, 어시스턴트 API의 출시에 대해 강조합니다. 향상된 기능을 갖춘 새로운 GPT-4 터보 모델과 Microsoft와의 파트너십 강화에 대한 언급도 있습니다.

샘 알트먼은 OpenAI 팀의 재능과 노고에 감사를 표합니다. 개인의 역량 강화와 주체성을 약속하는 기술 및 사회 혁명으로서 AI의 잠재력에 대해 논의합니다. 이 연설은 청중들에게 미래를 위한 구축과 창조에 AI를 활용하도록 격려하고, 내년에 더욱 발전할 AI에 대한 기대감으로 마무리됩니다.

원문 번역(by deepl)

00:00
[음악] -좋은 아침입니다. 오늘 함께 해주셔서 감사합니다. 무대에 오신 것을 환영합니다, 샘 알트먼. [음악] [박수] -좋은 아침입니다. 첫 번째 OpenAI 데브데이에 오신 것을 환영합니다. 여러분을 모시게 되어 기쁘고 이 열기가 정말 대단합니다. [박수] - 샌프란시스코에 오신 것을 환영합니다. 샌프란시스코는 첫날부터 우리의 고향이었습니다.
00:33
샌프란시스코는 저희와 기술 업계 전반에 중요한 도시입니다. 샌프란시스코에서 계속 성장할 수 있기를 기대합니다. 오늘 발표할 멋진 소식들이 많지만, 먼저 지난 1년 동안 우리가 해온 일들에 대해 잠시 이야기하고 싶습니다. 약 1년 전인 11월 30일, 저희는 "로키 리서치 프리뷰"로 ChatGPT를 출시했고, 꽤 잘 진행되었습니다.
00:58
3월에는 그 후속 조치로 여전히 세계에서 가장 성능이 뛰어난 모델인 GPT-4를 출시했습니다. [박수] - 지난 몇 달 동안 음성 및 시각 기능을 출시하여 이제 ChatGPT가 보고, 듣고, 말할 수 있게 되었습니다. [박수] -매번 박수를 칠 필요가 없을 정도로 많은 기능이 있습니다. [웃음] -최근에는 세계에서 가장 진보된 이미지 모델인 DALL-E 3를 출시했습니다.
01:28
물론 ChatGPT 내에서 사용할 수 있습니다. 기업 고객을 위해 엔터프라이즈급 보안 및 개인 정보 보호, 더 빠른 속도의 GPT-4 액세스, 더 긴 컨텍스트 창 등을 제공하는 ChatGPT Enterprise를 출시했습니다. 현재 약 2백만 명의 개발자가 다양한 사용 사례를 위해 API를 구축하여 놀라운 작업을 수행하고 있으며, 포춘지 선정 500대 기업 중 92% 이상이 저희 제품을 사용하고 있으며, 현재 ChatGPT의 주간 활성 사용자는 약 1억 명에 달합니다.
02:02
[박수] - 놀라운 점은 전적으로 입소문만으로 달성했다는 것입니다. 사람들이 유용하다고 생각하면 친구들에게 알려주죠. OpenAI는 현재

세계에서 가장 진보되고 가장 널리 사용되는 AI 플랫폼이지만, 숫자가 모든 것을 말해주지는 않습니다. 정말 중요한 것은 사람들이 제품을 어떻게 사용하는지, 사람들이 AI를 어떻게 사용하는지, 그래서 짧은 동영상을 보여드리겠습니다.
02:30
-저는 사실 아버지께 타갈로그어로 편지를 쓰고 싶었습니다. 낭만적이지 않은 방식으로 부모님께 사랑한다는 말을 전하고 싶고, 저를 의지해도 된다는 말을 전하고 싶지만 필리핀 문화와 타갈로그어 문법에 따라 자녀와 부모 관계에 대한 존중이 담긴 방식으로 전하고 싶어요.
02:53
이 말을 타갈로그어로 번역하면 "널 정말 사랑해, 어떤 길이든 너와 함께할 거야"라는 뜻이 되죠. -어떤 가능성을 보고 "와우"라고 생각했어요. 가끔 어떤 일에 대해 확신이 서지 않을 때 ChatGPT를 통해 '내가 생각하는 게 바로 이거야'라고 생각하면 자신감이 생기는 것 같아요.
03:08
-가장 먼저 제 마음을 사로잡은 것은 여러분과 레벨이 같다는 점이었습니다. 많은 사람들이 어려움을 겪는 일이죠. 모든 크리에이티브가 자신의 이야기를 들어주는 사람만 있다면 얼마든지 할 수 있다는 생각이 들었습니다. -이것은 병든 헤모글로빈을 표현한 것입니다. -ChatGPT로 제작하셨나요? -ChatGPT와 함께 만들었습니다.

03:31
-"여기 제 냉장고 사진이에요." 같은 일상적인 활동에 사용하기 시작했습니다. 제가 뭘 놓치고 있는지 알려주실 수 있나요? 장보러 가는데 비건 식단을 따르는 레시피가 정말 필요하거든요."와 같이요. -코드 인터프리터에 액세스하자마자 "와, 이거 정말 멋지다"라는 생각이 들었습니다.
03:46
" 스프레드시트를 만들 수 있었죠. 무엇이든 할 수 있었죠. -약 3개월 전 제 100번째 생일에 Chatty를 발견했습니다. Chatty는 매우 친절하고, 인내심이 강하며, 지식이 풍부하고, 매우 빠릅니다. 정말 멋진 일이었어요. -저는 4.0 학생이지만 자녀도 네 명이나 있습니다. ChatGPT를 사용하기 시작하면서

ChatGPT에게 그런 질문을 할 수 있다는 것을 깨달았습니다.
04:14
답을 알려줄 뿐만 아니라 설명도 해줍니다. 과외가 더 이상 필요하지 않았어요. 제 삶을 되찾았어요. 가족을 위한 시간과 저를 위한 시간이 생겼습니다. -저는 왼쪽 몸 전체에 만성 신경 질환이 있어서 신경 손상이 있습니다. 뇌수술을 받았어요. 왼손을 제한적으로 사용할 수 있습니다.
04:35
이제 음성 입력을 통합할 수 있습니다. 그리고 앞뒤로 대화할 수 있는 최신 인터페이스는 저에게 최고의 인터페이스입니다. 여기 있습니다. [음악] [박수] - 저희는 사람들이 이 기술을 어떻게 사용하고 있는지에 대한 이야기를 듣는 것을 좋아합니다. 이것이 바로 우리가 이 모든 일을 하는 이유입니다.
05:04
이제 새로운 소식으로 넘어가 보겠습니다. [청중 환호] -먼저 저희가 개선한 몇 가지 사항에 대해 이야기한 다음, 앞으로 나아갈 방향에 대해 이야기하겠습니다. 지난 한 해 동안 저희는 전 세계 개발자들과 많은 시간을 보냈습니다. 많은 피드백을 들었습니다.

05:24
그 피드백을 바탕으로 오늘 여러분께 보여드릴 내용을 준비했습니다. 오늘은 새로운 모델인 GPT-4 터보를 출시합니다. [박수] -GPT-4 터보는 여러분께서 요청하셨던 많은 사항들을 해결해 줄 것입니다. 새로운 기능을 살펴보겠습니다. 이 부분에서는 크게 여섯 가지에 대해 이야기하겠습니다. 첫 번째, 컨텍스트 길이입니다.
05:52
많은 사람들이 훨씬 더 긴 컨텍스트 길이가 필요한 작업을 하고 있습니다. GPT-4는 최대 8K, 경우에 따라 최대 32K의 컨텍스트 길이를 지원했지만, 많은 분들이 원하는 작업에는 충분하지 않다는 것을 알고 있습니다. GPT-4 Turbo는 최대 128,000개의 컨텍스트 토큰을 지원합니다. [박수] - 이는 표준 책 300페이지에 해당하는 분량으로, 8K 컨텍스트보다 16배 더 깁니다.
06:21

컨텍스트 길이가 길어질 뿐만 아니라 모델이 긴 컨텍스트에서 훨씬 더 정확하다는 것을 알 수 있습니다. 둘째, 더 많은 제어 기능. 유니티는 개발자가 모델의 응답과 출력에 대해 더 많은 제어 기능을 필요로 한다는 의견을 많이 들었습니다. 여러 가지 방법으로 이 문제를 해결했습니다. 모델이 유효한 JSON으로 응답하도록 보장하는 JSON 모드라는 새로운 기능이 있습니다.
06:47
이는 개발자들의 요청이 많았던 기능입니다. API 호출이 훨씬 쉬워질 것입니다. 함수 호출 기능도 훨씬 개선되었습니다. 이제 한 번에 여러 함수를 호출할 수 있으며, 일반적으로 지침을 더 잘 따를 수 있습니다. 재현 가능한 출력이라는 새로운 기능도 도입합니다. 시드 매개변수를 전달하면 모델이 일관된 출력을 반환하도록 만들 수 있습니다.
07:10
물론 이 기능을 사용하면 모델 동작을 더 세밀하게 제어할 수 있습니다. 이 기능은 오늘 베타 버전으로 출시됩니다. [박수] - 앞으로 몇 주 안에 API에서 로그 프로브를 볼 수 있는 기능을 출시할 예정입니다. [박수] - 알겠습니다. 세 번째, 더 나은 세상 지식. 여러분은 이러한 모델이 세상에 대한 더 나은 지식에 액세스할 수 있기를 바라며, 저희도 마찬가지입니다.
07:36
플랫폼에서 검색 기능을 출시합니다. 외부 문서나 데이터베이스의 지식을 구축 중인 모든 것에 가져올 수 있습니다. 지식 차단 기능도 업데이트 중입니다. 저희도 여러분만큼이나, 어쩌면 그보다 더 많이, 세상에 대한 GPT-4의 지식이 2021년에 종료된다는 사실에 짜증이 납니다. 다시는 이런 일이 발생하지 않도록 노력하겠습니다.
07:55
GPT-4 Turbo는 2023년 4월까지 세계에 대한 지식을 보유하고 있으며, 시간이 지남에 따라 계속 개선해 나갈 것입니다. 네 번째, 새로운 양식. 아무도 예상치 못한 DALL-E 3, 비전 기능이 탑재된 GPT-4 터보, 그리고 새로운 텍스트 음성 변환 모델이 모두 오늘 API에 추가됩니다. [박수] - 이미지와 디자인을 프로그래밍 방식으로 생성하기 위해 DALL-E 3를 사용하기 시작한 고객이 몇 명 있습니다.

08:31
오늘, 코카콜라는 고객이 DALL-E 3를 사용하여 디왈리 카드를 생성할 수 있는 캠페인을 시작하며, 물론 개발자가 애플리케이션을 오용하지 않도록 보호하는 안전 시스템도 제공합니다. 이러한 도구는 API에서 사용할 수 있습니다. 이제 GPT-4 Turbo는 API를 통해 이미지를 입력으로 받아 캡션, 분류 및 분석을 생성할 수 있습니다.
08:52
예를 들어, 비 마이 아이즈는 이 기술을 사용하여 시각 장애가 있거나 시력이 낮은 사람들이 눈앞에 있는 제품을 식별하는 것과 같은 일상적인 작업을 돕습니다. 새로운 텍스트 음성 변환 모델을 사용하면 API의 텍스트에서 6개의 사전 설정된 음성 중에서 선택할 수 있는 놀랍도록 자연스러운 오디오를 생성할 수 있습니다.
09:14
예를 들어보겠습니다. -저명한 발명가인 알렉산더 그레이엄 벨이 소리의 세계에 매료되었다는 사실을 알고 계셨나요? 그의 독창적인 생각은 밀랍에 소리를 새겨 시간이 지나도 속삭이는 소리를 내는 축음기를 만들게 되었습니다. -지금까지 우리가 들어왔던 그 어떤 소리보다 훨씬 자연스러운 소리입니다.
09:34
음성을 사용하면 앱과 더 자연스럽게 상호 작용하고 더 쉽게 접근할 수 있습니다. 또한 언어 학습 및 음성 지원과 같은 많은 사용 사례를 열어줍니다. 새로운 방식에 대해 말씀드리자면, 오늘 오픈 소스 음성 인식 모델인 Whisper V3의 다음 버전도 출시되며, 곧 API에 제공될 예정입니다.
09:54
여러 언어에 걸쳐 향상된 성능을 제공하는 이 버전은 여러분이 정말 좋아할 것입니다. 다섯 번째, 사용자 지정. 몇 달 전 출시 이후 GPT-3.5에 대한 미세 조정이 정말 잘 이루어지고 있습니다. 오늘부터 이 기능을 16K 버전으로 확장할 예정입니다. 또한 오늘부터 적극적인 미세 조정 사용자를 대상으로 GPT-4 미세 조정 실험 액세스 프로그램을 신청할 수 있도록

초대합니다.
10:21
미세 조정 API는 비교적 적은 양의 데이터로 다양한 애플리케이션에서 더 나은 성능을 달성하도록 모델을 조정하는 데 유용하지만, 완전히 새로운 지식 영역을 학습하거나 많은 독점 데이터를 사용하도록 모델을 조정하고 싶을 수도 있습니다. 오늘은 커스텀 모델이라는 새로운 프로그램을 소개합니다.
10:40
커스텀 모델을 통해 저희 연구진은 기업과 긴밀히 협력하여 특히 그 기업에 맞는 훌륭한 커스텀 모델을 만들 수 있도록 돕고, 도구를 사용하여 그 기업의 사용 사례를 지원합니다. 여기에는 모델 학습 프로세스의 모든 단계 수정, 추가적인 도메인별 사전 학습, 특정 도메인에 맞춘 맞춤형 RL 사후 학습 프로세스 등이 포함됩니다.
11:02
많은 회사에서 이 작업을 시작하기는 어려울 것입니다. 많은 작업이 필요하고 적어도 초기에는 비용이 저렴하지 않을 것이지만, 현재 가능한 범위까지 일을 추진하는 데 흥미를 느끼신다면 문의해 주세요. 저희에게 연락해 주시면 멋진 일을 해낼 수 있을 것 같습니다.
11:18
여섯 번째, 수수료 한도 상향. 기존 GPT-4 고객의 분당 토큰을 두 배로 늘려서 더 많은 작업을 더 쉽게 할 수 있도록 했습니다. API 계정 설정에서 직접 추가 전송률 한도 및 할당량 변경을 요청할 수 있습니다. 이러한 속도 제한 외에도, 저희는 여러분이 플랫폼에서 성공적으로 구축할 수 있도록 최선을 다하고 있습니다.
11:42
저작권 보호 기능을 도입합니다. 저작권 보호는 법적 청구 또는 저작권 침해에 직면한 경우 당사가 개입하여 고객을 방어하고 발생한 비용을 지불하는 것을 의미하며, 이는 ChatGPT Enterprise와 API 모두에 적용됩니다. 다시 한 번 말씀드리지만, 지금은 API나 ChatGPT Enterprise의 데이터로 트레이닝을 하지 않는다는 점을 다시 한 번 상기시켜드리고 싶습니다.

12:06
좋아요. 사실 이 모든 것보다 더 큰 개발자 요청이 하나 더 있었기 때문에 지금 그 얘기를 하고 싶은데 바로 가격입니다. [웃음] -GPT-4 Turbo는 업계를 선도하는 모델입니다. 방금 말씀드린 많은 개선 사항을 제공하며 GPT-4보다 더 스마트한 모델입니다.
12:32
개발자들로부터 구축하고 싶은 것이 많지만 GPT-4는 비용이 너무 많이 든다는 이야기를 들었습니다. 개발자들은 비용을 20%, 25%만 줄일 수 있다면 정말 좋을 것 같다고 말했습니다. 엄청난 도약이죠. 저희는 정말 열심히 노력했고, 더 나은 모델인 GPT-4 터보는 프롬프트 토큰의 경우 GPT-4보다 3배나 저렴하다는 사실을 발표하게 되어 매우 기쁩니다.
12:58
[박수] -오늘부터 완주 토큰은 2배로 인상됩니다. [박수] -새로운 가격은 프롬프트 토큰 1,000개당 1센트, 완료 토큰 1,000개당 3센트입니다. 대부분의 고객에게 이는 GPT-4보다 GPT-4 터보의 혼합 요금이 2.75배 이상 저렴해지는 결과를 가져올 것입니다. 저희는 이를 실현하기 위해 정말 열심히 노력했습니다.
13:28
여러분도 저희만큼이나 기대가 크시길 바랍니다. [박수] - 둘 중 하나를 선택해야 했기 때문에 가격을 우선순위에 두기로 결정했지만, 다음에는 속도에 대해 연구할 것입니다. 속도도 중요하다는 것을 알고 있습니다. 곧 GPT-4 터보가 훨씬 빨라질 것입니다. 또한 GPT-3.5 터보 16K의 가격도 낮추고 있습니다.
13:54
또한 입력 토큰은 3배, 출력 토큰은 2배 더 적습니다. 즉, GPT-3.5 16K는 이제 이전 GPT-3.5 4K 모델보다 저렴해졌습니다. 미세 조정된 GPT-3.5 터보 16K 버전을 실행하는 것도 기존의 미세 조정된 4K 버전보다 저렴합니다. 지금까지 모델 자체에 대해 많은 부분을 다루었습니다. 이번 변경 사항이 여러분의 피드백에 도움이 되길 바랍니다.
14:19
이제 모든 분들께 이 모든 개선 사항을 제공하게 되어 정말 기쁩니다. 이 모든 과정에서 저희는 운 좋게도 이를 실현하는 데 중요한 역할을 한

파트너를 만나게 되었습니다. 특별 게스트로 Microsoft의 CEO인 사티아 나델라를 모시고 싶습니다. [청중 환호] [음악] -반갑습니다. -정말 감사합니다.
14:41
감사합니다. -사티아, 와주셔서 정말 감사합니다. -여기 오게 돼서 정말 기쁘고 샘, 축하해요. 터보와 앞으로의 모든 일정이 정말 기대됩니다. 여러분과 함께 일하게 되어 정말 환상적이었어요. -멋지네요. 두 가지 질문이 있습니다. 시간을 많이 뺏지 않겠습니다. Microsoft는 현재 이 파트너십에 대해 어떻게 생각하고 있나요? -먼저, [웃음] 여러분을 사랑합니다.
15:06
[저희에게는 정말 환상적이었습니다. 사실 처음 연락을 주셨을 때 "혹시 Azure 크레딧이 있으세요?"라고 말씀하셨던 게 기억납니다. 거기서부터 먼 길을 왔네요. -정말 고마워요. 정말 멋졌어요. -여러분은 마법 같은 것을 만들었습니다. 솔직히 파트너십에 관해서는 두 가지가 있습니다.
15:24
첫 번째는 이러한 워크로드입니다. 무대 뒤에서 앞으로의 작업에 대해 설명하는 것을 듣고 있는데도 너무 다르고 새롭습니다. 저는 30년 동안 이 인프라 비즈니스에 종사해 왔습니다. -아무도 이런 인프라를 본 적이 없습니다. -워크로드, 워크로드 패턴, 이러한 교육 작업은 매우 동기적이고 규모가 크며 데이터가 병렬로 처리됩니다.
15:45
우리가 가장 먼저 한 일은 전원부터 DC, 랙, 가속기, 네트워크에 이르기까지 모든 것을 고객과 협력하여 시스템을 구축하는 것입니다. Azure의 형태가 크게 바뀌고 있으며 여러분이 구축하는 모델을 지원하기 위해 빠르게 변화하고 있습니다.
16:06
가장 중요한 것은 최고의 모델을 구축할 수 있도록 최고의 시스템을 구축한 다음 개발자가 이 모든 것을 사용할 수 있도록 하는 것입니다. 다른 하나는 우리 자신이 개발자입니다. 우리는 제품을 만들고 있습니다. 사실 저는 GPT에서 GitHub Copilot을 처음 본 순간 이 모든 세대의 기반 모델에 대한 확신이 완전히 바뀌었습니다.
16:29

저희는 모든 개발자가 OpenAI API를 기반으로 GitHub Copilot을 구축하기를 원합니다. 저희는 이를 위해 최선을 다하고 있습니다. 개발자에게는 어떤 의미인가요? 저는 항상 Microsoft를 플랫폼 회사, 개발자 회사, 파트너 회사로 생각합니다. 예를 들어, 여기 참석한 모든 개발자가 사용해 볼 수 있도록 엔터프라이즈 에디션인 GitHub Copilot을 제공하고자 합니다.
16:55
멋지네요. 정말 기대됩니다. [박수] - API 지원을 통해 Azure에서 최고의 인프라를 구축하여 여러분 모두에게 제공할 수 있습니다. Azure 마켓플레이스 같은 것들도요. 여기서 제품을 빌드하는 개발자가 빠르게 시장에 출시할 수 있습니다. 이것이 바로 우리의 의도입니다.

17:17
-좋아요. 미래, 파트너십의 미래, AI의 미래 등에 대해 어떻게 생각하시나요? 무엇이든 물어보세요 -저에게 매우 중요한 두 가지가 있다고 생각합니다. 하나는 방금 말씀드린 대로 로드맵을 공격적으로 추진하기 위해 필요한 시스템을 갖추기 위해서는 우리가 최고의 자리에 있어야 하며, 이러한 기반 모델을 구축하는 여러분 모두가 훈련과 추론을 위한 최고의 시스템을 갖출 수 있도록 전적으로 헌신할 계획입니다,
17:55
최고의 컴퓨팅 성능을 확보할 수 있도록 최선을 다할 것입니다. --계속 앞으로 나아가는 것이 우리가 발전할 수 있는 길이라고 생각하기 때문입니다. 두 번째로 우리 둘 다 중요하게 생각하는 것은, 사실 솔직히 말해서, 양쪽이 함께 모이게 된 것은 여러분의 사명과 우리의 사명입니다.
18:11
우리의 사명은 지구상의 모든 사람과 조직이 더 많은 것을 성취할 수 있도록 힘을 실어주는 것입니다. 궁극적으로 AI는 진정으로 힘을 실어줄 때만 유용할 것이라고 생각합니다. 아까 재생하신 동영상을 봤어요. AI가 자신에게 어떤 의미인지, 무엇을 성취할 수 있었는지 설명하는 목소리를 들으니 정말 환상적이었습니다.
18:29
궁극적으로는 AI의 혜택을 모든 사람에게 널리 보급하는 것이 저희의 목표가 될 것입니다. 그리고 마지막으로 안전이 중요하다는 사실에

기반하고 있으며, 안전은 나중에 신경 쓸 일이 아니라 우리가 왼쪽으로 이동하는 일이며 여러분과 함께 안전에 매우 집중하고 있습니다.
18:47
-좋아요. 저희는 기술 분야에서 최고의 파트너십을 맺고 있다고 생각합니다. AGI를 함께 만들게 되어 기대가 됩니다. -정말 기대됩니다. 환상적인 [크로스 토크] 되세요. -와주셔서 정말 감사합니다. -정말 감사합니다. -또 뵙겠습니다. [박수] -개발자를 위한 많은 훌륭한 업데이트를 이미 공유했고 앞으로 더 많은 업데이트가 있을 예정이지만, 개발자 컨퍼런스임에도 불구하고 ChatGPT에 대한 몇 가지 개선 사항을 소개하지 않을 수 없습니다.

19:16
작은 부분이지만, ChatGPT는 이제 최신 지식 차단을 포함한 모든 최신 개선 사항이 포함된 GPT-4 Turbo를 사용하며, 이는 계속 업데이트될 예정입니다. 오늘 모두 라이브입니다. 이제 필요할 때 웹을 탐색하고, 코드를 작성 및 실행하고, 데이터를 분석하고, 이미지를 촬영 및 생성하는 등의 작업을 수행할 수 있습니다. 매우 성가셨던 모델 선택기가 오늘부터 사라진다는 여러분의 피드백을 들었습니다.
19:37
이제 드롭다운 메뉴를 클릭할 필요가 없습니다. 이 모든 것이 함께 작동합니다. 예. [박수] -ChatGPT는 무엇을 언제 사용해야 하는지 알아서 알려주지만, 그게 중요한 것은 아닙니다. 사실 가격도 개발자의 주요 요청사항이 아니었습니다. 그보다 더 큰 요구사항이 있었습니다. 지금부터 우리가 나아갈 방향과 오늘 이야기하고자 하는 주요 사항에 대해 말씀드리겠습니다.
20:06
저희는 사람들에게 더 나은 도구를 제공하면 놀라운 일을 할 수 있다고 믿습니다. 사람들은 더 똑똑하고, 더 개인화되고, 더 맞춤화되고, 사용자를 대신해 더 많은 일을 할 수 있는 AI를 원한다는 것을 잘 알고 있습니다. 결국에는 컴퓨터에게 필요한 것을 요청하기만 하면 컴퓨터가 이 모든 작업을 대신 수행하게 될 것입니다. 이러한 기능을 AI 분야에서는 흔히

'에이전트'라고 부릅니다.
20:29
" 이로 인한 장점은 엄청날 것입니다. OpenAI에서는 점진적인 반복 배포가 AI의 안전 문제, 즉 보안 문제를 해결하는 가장 좋은 방법이라고 믿습니다. 특히 에이전트의 미래를 향해 신중하게 나아가는 것이 중요하다고 생각합니다. 많은 기술적 작업과 사회의 신중한 고려가 필요할 것입니다.
20:51
오늘 우리는 이러한 미래를 향해 나아가기 위한 작은 첫 걸음을 내딛습니다. GPT를 소개하게 되어 매우 기쁩니다. GPT는 특정 목적에 맞는 ChatGPT의 맞춤형 버전입니다. 지침, 확장된 지식 및 행동이 포함된 거의 모든 용도의 맞춤형 ChatGPT 버전인 GPT를 구축한 다음 다른 사람들이 사용할 수 있도록 게시할 수 있습니다.

21:20
지침, 확장된 지식, 행동이 결합되어 있기 때문에 더 많은 도움이 될 수 있습니다. 여러 상황에서 더 잘 작동할 수 있고, 더 잘 제어할 수 있습니다. 모든 종류의 작업을 더 쉽게 수행하거나 더 재미있게 즐길 수 있도록 도와주며 ChatGPT 내에서 바로 사용할 수 있습니다.
21:38
대화하는 것만으로도 사실상 언어로 GPT를 프로그래밍할 수 있습니다. 원하는 목적에 맞게 동작을 쉽게 사용자 지정할 수 있습니다. 따라서 매우 쉽게 구축할 수 있으며 모든 사람에게 권한을 부여할 수 있습니다. GPT가 무엇인지, 어떻게 사용하는지, 어떻게 구축하는지, 그리고 어떻게 배포되고 발견되는지에 대해 이야기하겠습니다.
22:00
그 다음에는 개발자를 위해 이러한 에이전트와 유사한 경험을 앱에 구축하는 방법을 보여드리겠습니다. 먼저 몇 가지 예를 살펴보겠습니다. Code.org의 파트너들은 학교에서 컴퓨터 과학을 확대하기 위해 열심히 노력하고 있습니다. 이들은 전 세계 수천만 명의 학생들이 사용하는 커리큘럼을 보유하고 있습니다.
22:20

Code.org는 교사가 중학생에게 더 매력적인 경험을 제공할 수 있도록 수업 플래너 GPT를 만들었습니다. 교사가 네 개의 루프를 창의적인 방식으로 설명해 달라고 요청하면 바로 그렇게 해줍니다. 이 경우, 비디오 게임 캐릭터가 반복적으로 동전을 줍는다는 관점에서 설명합니다. 8학년이 이해하기 매우 쉽습니다.
22:40
보시다시피, 이 GPT는 Code.org의 광범위한 커리큘럼과 전문 지식을 통합하여 교사가 필요에 따라 빠르고 쉽게 적용할 수 있습니다. 다음으로, Canva는 자연어로 원하는 것을 설명하여 디자인을 시작할 수 있는 GPT를 만들었습니다. "오늘 오후, 오늘 저녁에 있을 개발자 데이 리셉션 포스터를 만들어줘"라고 말하고 몇 가지 세부 정보를 입력하면 Canva의 API를 통해 몇 가지 옵션을 생성하여 시작할 수 있습니다.
23:08
이 개념이 익숙하신 분도 계실 겁니다. 저희는 플러그인을 GPT를 위한 커스텀 액션으로 발전시켰습니다. 이 플러그인으로 계속 채팅하면서 다양한 반복을 확인하고 마음에 드는 것을 발견하면 Canva로 이동하여 전체 디자인 경험을 할 수 있습니다. 이제 GPT 라이브를 보여드리겠습니다.
23:28
Zapier는 6,000개의 애플리케이션에서 작업을 수행하여 모든 종류의 통합 가능성을 열어주는 GPT를 구축했습니다. 이 데모를 진행할 솔루션 아키텍트 중 한 명인 Jessica를 소개하겠습니다. 제시카를 환영합니다. [박수] - 감사합니다, Sam. 안녕하세요, 여러분. 모두 감사합니다. 모두 참석해 주셔서 감사합니다.
23:51
저는 제시카 쉬입니다. 저는 파트너 및 고객과 협력하여 제품에 생명을 불어넣는 일을 합니다. 오늘은 저희가 얼마나 열심히 노력해왔는지 보여드리고 싶어서 빨리 시작하고 싶네요. GPT를 시작할 위치는 이 왼쪽 상단 모서리입니다. Zapier AI 동작을 클릭하는 것으로 시작하겠습니다. 오른쪽에 오늘 제 캘린더가 표시됩니다.

24:14
정말 멋진 하루입니다. 전에도 사용해 본 적이 있어서 사실 이미 캘린더에 연결되어 있습니다. 먼저 "오늘 일정이 뭐야?"라고 물어볼 수 있습니다. 저희는 보안을 염두에 두고 GPT를 구축합니다. 어떤 작업을 수행하거나 데이터를 공유하기 전에 사용자의 허가를 요청합니다. 여기서는 허용이라고 답하겠습니다.
24:37
GPT는 사용자의 지시를 받아 해당 작업을 수행하기 위해 어떤 기능을 호출할지 결정한 다음 이를 실행하도록 설계되었습니다. 여기 보시다시피 이미 제 캘린더에 연결되어 있습니다. 내 정보를 가져온 다음 내 캘린더에서 충돌을 식별하라는 메시지도 표시합니다.
24:57
여기에서 실제로 이를 식별할 수 있었다는 것을 알 수 있습니다. 예정된 일정이 있는 것 같습니다. 샘에게 일찍 퇴근해야 한다고 알리려면 어떻게 해야 할까요? 바로 여기에 "샘에게 나 가야 한다고 알려줘. GPU를 쫓아갑니다." 그리고 샘과의 대화로 전환한 다음 "네, 실행해 주세요."라고 말합니다.
25:27
" 샘, 들었어? -네. -멋지네요. [박수] -이것은 가능성의 일부일 뿐이며 여러분 모두가 무엇을 만들지 빨리 보고 싶습니다. 고마워요. 다시 한 번, 샘. [박수] - 고마워요, 제시카. 세 가지 훌륭한 예입니다. 이 외에도 사람들이 만들고 있는 GPT의 종류는 훨씬 더 많으며, 곧 더 많은 GPT가 만들어질 것입니다.
26:02
GPT를 만들고자 하는 많은 사람들이 코딩 방법을 모른다는 것을 알고 있습니다. 저희는 대화만으로 GPT를 프로그래밍할 수 있도록 만들었습니다. 우리는 자연어가 앞으로 사람들이 컴퓨터를 사용하는 방식에서 큰 부분을 차지할 것이라고 믿으며, 이것이 흥미로운 초기 사례라고 생각합니다.
26:20
만드는 방법을 보여드리겠습니다. 좋아요. 저는 창업자와 개발자가 새로운 프로젝트를 시작할 때 조언을 줄 수 있는 GPT를 만들고 싶어요. 여기서 GPT를 만들려고 하는데, GPT 빌더로 이동합니다. 저는 YC에서 수년간

창업자들과 함께 일했지만 지금도 개발자들을 만날 때마다 항상 받는 질문은 "사업 아이디어에 대해 어떻게 생각하나요?"입니다. 조언 좀 해주실 수 있나요?" 이를 돕기 위해 GPT를 구축할 수 있는지 알아보려고 합니다.
26:53
먼저 GPT 빌더가 무엇을 만들고 싶은지 물어보면 저는 "스타트업 창업자의 사업 아이디어를 통해 생각을 돕고 조언을 얻고 싶습니다."라고 대답합니다. 창업자가 조언을 얻은 후에는 왜 더 빨리 성장하지 못하는지에 대해 질문합니다." [웃음] -알겠어요. 우선 GPT에 제가 원하는 바를 조금만 말씀드리겠습니다.
27:26
그러면 GPT는 이에 대해 생각하기 시작하고 GPT를 위한 몇 가지 세부 지침을 작성할 것입니다. 또한 이름에 대해 물어볼 것입니다. 스타트업 멘토에 대해 어떻게 생각하나요? 괜찮습니다. "좋아요." 물론 이름이 마음에 들지 않으면 다른 이름으로 부를 수도 있지만, 저와 대화를 시도하고 거기서부터 시작하려고 할 것입니다.
27:46
여기 오른쪽의 미리보기 모드에서 이미 GPT를 작성하기 시작하는 것을 볼 수 있습니다. 무엇을 하는지에 대한 설명과 함께 제가 추가로 질문할 수 있는 몇 가지 아이디어가 나와 있습니다. [웃음] 방금 후보자가 생성되었습니다. 물론 제가 다시 생성하거나 변경할 수도 있지만, 저는 그게 마음에 듭니다. 저는 "좋아요.

28:13
" 이제 GPT가 조금씩 더 구축되고 있는 것을 보셨을 겁니다. 이제 제가 원하는 기능, 사용자와 상호 작용하는 방법, 스타일에 대해 이야기할 수 있습니다. 제가 말씀드리고자 하는 것은 제가 했던 스타트업 관련 강연의 녹취록을 업로드할 예정이니 이를 바탕으로 조언을 부탁드립니다.
28:39
" 좋아요. 이제 그 방법을 알아볼 차례입니다. 구성 탭을 보여드리겠습니다. 빌더 자체에 의해 여기에 구축된 몇 가지 사항을 볼 수 있습니다. 여기에 제가 활성화할 수 있는 기능이 있는 것을 볼 수 있습니다. 사용자 지정

작업을 추가할 수 있습니다. 이것들은 모두 그대로 두어도 괜찮습니다.
28:58
파일을 업로드하겠습니다. 여기 제가 스타트업에 대한 조언과 함께 제가 골라낸 강의가 있는데 여기에 추가하겠습니다. 이 질문들에 관해서는 이것은 멍청한 질문입니다. 나머지는 창업자들이 자주 묻는 합리적인 질문입니다. 여기에 한 가지를 더 추가하자면, 피드백을 간결하고 건설적으로 하라는 것입니다.
29:26
좋아요. 다시 한 번 말씀드리지만 시간이 더 있었다면 더 많은 것을 보여드리고 싶습니다. 이 정도면 괜찮은 시작입니다. 이제 이 미리보기 탭에서 사용해 볼 수 있습니다. 가장 많이 받는 질문은 무엇일까요? "초기 단계의 스타트업에서 직원을 채용할 때 살펴봐야 할 세 가지 사항은 무엇인가요?" 이제 제가 업로드한 문서를 살펴볼 것입니다.
29:56
물론 여기에는 GPT-4에 대한 모든 배경 지식도 포함되어 있습니다. 꽤 괜찮네요. 제가 여러 번 말씀드린 세 가지입니다. 이제 다른 설명에 이어서 왜 더 빨리 성장하지 못하는지에 대한 질문을 할 수도 있지만 시간 관계상 생략하겠습니다.
30:16
지금은 저에게만 공개하겠습니다. 나중에 작업할 수 있습니다. 더 많은 콘텐츠를 추가하고 유용하다고 생각되는 몇 가지 작업을 추가한 다음 공개적으로 공유할 수 있습니다. 이것이 바로 GPT를 만드는 모습입니다
[박수] - 감사합니다. 그건 그렇고, 저는 항상 YC 근무 시간이 끝나면 '언젠가 이걸 할 수 있는 봇을 만들면 정말 멋질 거야'라고 생각했습니다.
30:45
"[웃음] - GPT를 통해 사람들이 ChatGPT를 사용하는 모든 재미있는 방법을 전 세계와 쉽게 공유하고 발견할 수 있도록 하고 있습니다. 방금 제가 한 것처럼 비공개 GPT를 만들 수도 있고, 누구나 사용할 수 있도록 링크를 통해 공개적으로 공유할 수도 있으며, ChatGPT Enterprise를 사용하는 경우 회사만을 위한 GPT를 만들 수도 있습니다.

31:11

이번 달 말에 GPT 스토어가 출시될 예정입니다. 감사합니다. 감사합니다.
[박수] - GPT를 등록하면 가장 인기 있는 최고의 GPT를 소개할 수 있습니다. 물론, 스토어에 등록된 GPT가 트위터의 정책을 준수하는지 확인한 후에 액세스할 수 있도록 할 것입니다. 저희는 수익 분배를 중요하게 생각합니다.
31:40
가장 유용하고 가장 많이 사용되는 GPT를 만든 사람들에게 수익의 일부를 지급할 예정입니다. 주말 동안 우리가 직접 구축한 것만으로도 GPT 스토어를 통해 활기찬 생태계를 조성할 수 있게 되어 기쁩니다. 멋진 콘텐츠가 많이 나올 것이라고 확신합니다. 곧 더 많은 정보를 공유할 수 있게 되어 기쁩니다.
31:58
여러분들이 무엇을 만들지 벌써부터 기대가 됩니다. 이번 개발자 컨퍼런스의 가장 멋진 점은 API에도 동일한 개념을 적용한다는 점입니다.
[박수] 많은 분들이 이미 API에서 플랫폼에서 작업을 수행할 수 있는 Shopify의 사이드킥과 같은 에이전트와 유사한 경험을 구축해 왔습니다.
32:25
Discord의 클라이드는 Discord 운영자가 사용자 지정 캐릭터를 만들 수 있는 기능이며, 스냅 마이 AI는 그룹 채팅에 추가하고 추천을 할 수 있는 맞춤형 챗봇입니다. 이러한 경험은 훌륭하지만 구축하기는 어려웠습니다. 수십 명의 엔지니어가 팀을 이루어 수개월이 걸리기도 하는 등 맞춤형 어시스턴트 경험을 만들기 위해 처리해야 할 사항이 많습니다.
32:49
이제 새로운 어시스턴트 API를 통해 훨씬 더 쉽게 만들 수 있습니다. [박수] - 어시스턴트 API에는 영구 스레드가 포함되어 있어 긴 대화 기록, 내장 검색, 코드 인터프리터, 샌드박스 환경에서 작동하는 Python 인터프리터, 그리고 앞서 설명한 개선된 함수 호출을 처리하는 방법을 알아낼 필요가 없습니다.
33:17
이 기능이 어떻게 작동하는지 데모를 보여드리겠습니다. 개발자 경험 책임자인 Romain을 소개합니다. 어서 오세요, 로맹. [음악] [박수] - 고마워요,

Sam. 좋은 아침입니다. 와우. 여러분을 만나게 되어 정말 반갑습니다. 많은 분들이 앱에 AI를 도입하는 모습을 보니 정말 고무적이었습니다. 오늘 API에서 새로운 양식을 출시하지만, 여러분 모두가 보조 에이전트를 구축할 수 있도록 개발자 환경을 개선하게 되어 매우 기쁩니다.
33:48
바로 시작해 보겠습니다. 전 세계 탐험가를 위한 1달러짜리 여행 앱을 만들고 있고 이것이 랜딩 페이지라고 가정해 보겠습니다. 저는 실제로 GPT-4를 사용하여 이러한 목적지 아이디어를 생각해 냈습니다. 예리한 눈을 가진 분들을 위해, 이 그림은 오늘 여러분 모두가 사용할 수 있는 새로운 DALL-E 3 API를 사용하여 프로그래밍 방식으로 생성되었습니다.
34:08
꽤 놀랍습니다. 이 앱에 아주 간단한 어시스턴트를 추가하여 앱을 향상시켜 보겠습니다. 이것이 화면입니다. 잠시 후에 다시 설명하겠습니다. 먼저 새로운 어시스턴트의 플레이그라운드로 전환하겠습니다.
어시스턴트를 만드는 방법은 간단합니다. 이름과 몇 가지 초기 지침, 모델만 지정하면 됩니다.
34:27
이 경우에는 GPT-4 Turbo를 선택하겠습니다. 여기에서 몇 가지 도구도 선택하겠습니다. 코드 인터프리터와 검색을 켜고 저장하겠습니다. 그게 다입니다. 어시스턴트가 준비되었습니다. 다음으로 이 어시스턴트 API의 두 가지 새로운 기본 요소인 스레드와 메시지를 통합할 수 있습니다. 코드를 간단히 살펴봅시다.
34:49
이 과정은 매우 간단합니다. 새로운 사용자마다 새 스레드를 생성합니다. 이 사용자가 어시스턴트와 소통하면 그 메시지를 스레드에 추가합니다. 아주 간단하죠. 그런 다음 언제든지 어시스턴트를 실행하여 응답을 앱으로 다시 스트리밍할 수 있습니다. 앱으로 돌아가서 실제로 사용해 볼 수 있습니다.
35:10
"이봐, 파리에 가자."라고 말하면. 알았어요. 그거예요. 이제 사용자는 몇 줄의 코드만으로 앱 내에서 매우 전문적인 비서를 사용할 수 있습니다.

여기서 제가 가장 좋아하는 기능 중 하나인 함수 호출을 강조하고 싶습니다. 아직 사용해 보지 않으셨다면 함수 호출은 정말 강력합니다.
35:31
샘이 언급했듯이 오늘 한 단계 더 발전했습니다. 이제 지연 시간 없이 JSON 출력을 보장하며, 처음으로 여러 함수를 한 번에 호출할 수 있습니다. 여기서 제가 계속해서 "이봐요, 해야 할 일 10가지가 뭐죠?"라고 말해보겠습니다. 어시스턴트가 다시 응답하도록 하겠습니다.
35:54
여기서 흥미로운 점은 어시스턴트가 오른쪽에 표시된 지도에 주석을 다는 기능을 포함하여 여러 기능을 알고 있다는 것입니다. 이제 이 모든 핀이 여기에 실시간으로 떨어지고 있습니다. 네, 정말 멋집니다. [박수] -이 통합을 통해 자연어 인터페이스가 앱의 구성 요소 및 기능과 원활하게 상호 작용할 수 있습니다.

36:17
이제 어시스턴트가 실제로 작업을 수행할 때 AI와 UI가 어떻게 조화를 이룰 수 있는지 진정으로 보여줍니다. 검색에 대해 이야기해 보겠습니다. 검색은 어시스턴트에게 즉각적인 사용자 메시지를 넘어 더 많은 지식을 제공하는 것입니다. 실제로 저는 영감을 받아 이미 파리행 티켓을 예약했습니다. 이 PDF를 여기로 끌어다 놓을게요.
36:40
업로드되는 동안 살짝 들여다볼 수 있습니다. 아주 전형적인 유나이티드 항공권입니다. 이 장면 뒤에서 검색이 이 파일을 읽고 있는 동안 이 PDF에 대한 정보가 화면에 나타납니다. [박수] - 물론 이것은 매우 작은 PDF이지만 어시스턴트는 구축하는 내용에 따라 광범위한 텍스트부터 복잡한 제품 사양에 이르기까지 긴 형식의 문서를 파싱할 수 있습니다.
37:07
사실 저도 에어비앤비를 예약했기 때문에 그 얘기를 대화로 끌어오려고 합니다. 그런데 많은 개발자들로부터 직접 구축하는 것이 얼마나 어려운지 들었습니다. 일반적으로 입찰가를 직접 계산해야 하고, 청킹 알고리즘도 설정해야 합니다. 이제 이 모든 것이 자동으로 처리됩니다.

37:25
API를 호출할 때마다 검색만 하는 것이 아니라 전체 대화 내역을 다시 보내야 하므로 키-값 저장소를 설정하고 컨텍스트 창을 처리하고 메시지를 직렬화하는 등의 작업을 수행해야 합니다. 이제 새로운 스테이트풀 API를 사용하면 이러한 복잡성이 완전히 사라집니다. OpenAI가 이 API를 관리한다고 해서 블랙박스라는 의미는 아닙니다.
37:47
실제로 개발자 대시보드에서 도구가 수행하는 단계를 바로 확인할 수 있습니다. 여기서 스레드를 클릭하면 현재 작업 중인 스레드를 확인할 수 있고, 올바른 매개변수로 호출되는 함수와 방금 업로드한 PDF를 포함한 모든 단계를 확인할 수 있습니다.
38:08
이제 많은 분들이 오랫동안 요청해 오셨던 새로운 기능으로 넘어가 보겠습니다. 이제 API에서도 코드 인터프리터를 사용할 수 있게 되어, AI가 코드를 즉시 작성하고 실행할 수 있을 뿐만 아니라 파일 생성까지 할 수 있게 되었습니다. 실제로 작동하는 모습을 살펴봅시다. "친구 4명이 이 에어비앤비 숙소에 묵는데, 내 몫과 내 항공편을 합쳐서 얼마지?"라고 입력해 보겠습니다. 좋아요.
38:43
이제 코드 인터프리터가 이 쿼리에 응답하기 위해 코드를 작성해야 한다는 것을 알아챘습니다. 이제 파리에서의 체류 일수와 친구 수를 계산하고 있습니다. 또한 센서를 가져 오기 위해 백그라운드에서 환율 계산도 수행합니다. 아주 복잡한 계산은 아니지만 이해가 되실 겁니다.
39:01
수많은 숫자를 계산하고 차트를 그리는 매우 복잡한 금융 앱을 만들고 있다고 상상해 보세요. 일반적으로 코드로 처리하는 모든 작업을 코드 인터프리터가 훌륭하게 처리할 수 있습니다. 좋아요. 파리 여행이 잘 마무리된 것 같아요. 지금까지 사용자 대화에 대한 상태를 관리하고, 지식 및 검색, 코드 인터프리터와 같은 외부 도구를 활용하고, 마지막으로 자체 함수를 호출하여 작업을 수행하는 어시스턴트를 빠르게 만드는 방법을 살펴봤습니다.
39:32

하지만 오늘 출시하는 새로운 양식과 결합된 함수 호출의 가능성을 실제로 보여드리고 싶었던 것이 하나 더 있습니다. 개발자 데이를 준비하면서 이 이벤트에 대한 모든 것을 알고 있는 작은 맞춤형 어시스턴트를 만들었는데, 오늘 하루 종일 돌아다니면서 채팅 인터페이스를 사용하는 대신 음성을 사용하면 어떨까 하는 생각이 들었습니다. 오른쪽에서 보실 수 있도록 제 휴대폰을 화면에 띄워 보겠습니다.
39:58
멋지네요. 오른쪽에는 마이크 입력을 받는 아주 간단한 Swift 앱이 보입니다. 왼쪽에는 실제로 터미널 로그를 불러와서 백그라운드에서 무슨 일이 일어나는지 볼 수 있도록 하겠습니다. 한 번 해보겠습니다.
안녕하세요, 저는 지금 기조 연설 무대에 서 있습니다. 데브 데이 참석자들에게 인사해 주시겠어요? -안녕하세요, 개발자 데이에 오신 것을 환영합니다.

40:24
여러분을 만나게 되어 정말 반갑습니다. 멋진 하루를 만들어 봅시다. [박수] - 인상적이지 않나요? API에는 각각 여러 언어를 구사하는 6개의 독특하고 풍부한 음성을 선택할 수 있으므로 앱에 꼭 맞는 음성을 찾을 수 있습니다. 여기 왼쪽에 있는 제 노트북에서 백그라운드에서 일어나는 일에 대한 로그도 볼 수 있습니다.
40:47
음성 입력을 텍스트로 변환하기 위해 Whisper를 사용하고, GPT-4 Turbo가 포함된 어시스턴트를 사용하고, 마지막으로 새로운 TTS API를 사용하여 말하도록 하고 있습니다. 함수 호출 덕분에 어시스턴트가 인터넷에 연결하여 사용자를 위해 실제 작업을 수행할 수 있게 되면 상황이 훨씬 더 흥미로워집니다. 여기서 더 흥미로운 것을 함께 만들어 봅시다.
41:08
이건 어때요? 어시스턴트, 여기 데브데이 참석자 중 5명을 무작위로 뽑아서 OpenAI 크레딧 500달러를 줄 수 있나요? [웃음] - 네, 참석자 명단 확인 중입니다. [네, 참석자 명단 확인 중입니다. 개발자 회의 참석자 5명을 뽑아 그들의 계정에 500달러의 API 크레딧을 추가했습니다. Christine M,

Jonathan C, Steven G, Luis K, Suraj S에게 축하를 보냅니다.
41:38
-자신을 알아보신다면 정말 멋지십니다. 축하합니다. 여기까지입니다. 오늘은 최종 사용자를 위한 간단한 서식 있는 텍스트 또는 음성 대화부터 시작하여 저희가 출시한 몇 가지 새로운 도구 및 양식과 결합된 새로운 어시스턴트 API에 대한 간략한 개요를 살펴보았습니다. 여러분의 멋진 결과물을 기대하며, 행운의 당첨자 여러분께도 축하의 인사를 전합니다.
42:00
사실, 여러분 모두 이 놀라운 OpenAI 커뮤니티의 일원이기 때문에 무대에서 내려오기 전에 제 어시스턴트에게 마지막으로 한 마디만 할게요. 어시스턴트, 여기 청중 모두에게 OpenAI 크레딧 500달러를 주실 수 있나요? -좋아요. 모두 살펴볼게요. [박수] -알겠습니다, 그 기능은 계속 실행할 수 있지만 시간이 다 됐어요.
42:32
여러분, 정말 감사합니다. 좋은 하루 되세요. 또 봐요, 샘. -꽤 멋지지 않나요? [청중 환호] -오케이, 오늘부터 어시스턴트 API가 베타 버전으로 제공되며, 여러분 모두가 이를 통해 무엇을 할 수 있을지 매우 기대됩니다. 시간이 지남에 따라 GPT와 어시스턴트는 상담원이 훨씬 더 많은 일을 할 수 있도록 하는 전조입니다.
43:06
점차적으로 상담원을 대신하여 더 복잡한 작업을 계획하고 수행할 수 있게 될 것입니다. 앞서 말씀드렸듯이 저희는 점진적인 반복 배포의 중요성을 정말 중요하게 생각합니다. 사람들이 지금부터 이러한 에이전트를 구축하고 사용하면서 더 많은 기능을 갖추게 되면 어떤 세상이 펼쳐질지 미리 느껴보는 것이 중요하다고 생각합니다.
43:26
항상 그래왔듯이 앞으로도 여러분의 피드백을 바탕으로 시스템을 지속적으로 업데이트할 것입니다. 오늘 이 모든 것을 여러분과 공유할 수 있게 되어 매우 기쁩니다. 지침, 확장된 지식, 행동이 결합된 맞춤형 버전인 GPT를 도입했습니다. 또한, 자체 앱으로 보조 경험을 더 쉽게 구축할 수 있도록 어시스턴트 API를 출시했습니다.
43:50

이는 AI 에이전트를 향한 첫걸음이며 시간이 지남에 따라 기능을 늘려나갈 예정입니다. 향상된 함수 호출, 지식, 저렴한 가격, 새로운 양식 등을 제공하는 새로운 GPT-4 터보 모델을 도입했습니다. Microsoft와의 파트너십도 더욱 공고히 하고 있습니다. 마지막으로 이 모든 것을 만들어낸 팀에게 감사의 말씀을 전하고 싶습니다.

44:14

OpenAI는 놀라운 인재 밀도를 가지고 있지만, 이 모든 것을 실현하기 위해서는 엄청난 노력과 조율이 필요합니다. 저는 정말 세계 최고의 동료들이 있다고 믿습니다. 그들과 함께 일할 수 있다는 사실에 정말 감사함을 느낍니다. 우리가 이 모든 일을 하는 이유는 AI가 기술적, 사회적 혁명이 될 것이라고 믿기 때문입니다.

44:33

AI는 다양한 방식으로 세상을 변화시킬 것이며, 여러분 모두가 우리 모두를 위해 많은 것을 구축할 수 있도록 힘을 실어줄 수 있는 일을 할 수 있게 되어 기쁩니다. 앞서 사람들에게 더 나은 도구를 제공하면 세상을 바꿀 수 있다고 이야기했습니다. 저희는 AI가 이전에는 볼 수 없었던 규모의 개인 역량 강화와 주체성에 관한 것이며, 이는 인류를 이전에는 볼 수 없었던 규모로 끌어올릴 것이라고 믿습니다.

44:56

우리는 더 많은 일을 하고, 더 많은 것을 창조하고, 더 많은 것을 가질 수 있게 될 것입니다. 인텔리전스가 모든 곳에 통합되면서 우리 모두는 필요에 따라 초능력을 갖게 될 것입니다. 여러분 모두가 이 기술로 무엇을 할 수 있을지, 그리고 우리 모두가 함께 설계할 새로운 미래를 발견할 수 있기를 기대합니다. 내년에 다시 찾아주시길 바랍니다.

45:15

오늘 우리가 출시한 것은 우리가 바쁘게 만들고 있는 것에 비해 매우 기이하게 보일 것입니다. 여러분의 모든 노력에 감사드립니다. 오늘 이 자리에 와주셔서 감사합니다. [박수] [음악]

00:00

[music] -Good morning. Thank you for joining us today. Please welcome to the stage, Sam Altman. [music] [applause] -Good morning. Welcome to our first-ever OpenAI DevDay. We're thrilled that

you're here and this energy is awesome. [applause] -Welcome to San Francisco. San Francisco has been our home since day one.
00:33
The city is important to us and the tech industry in general. We're looking forward to continuing to grow here. We've got some great stuff to announce today, but first, I'd like to take a minute to talk about some of the stuff that we've done over the past year. About a year ago, November 30th, we shipped ChatGPT as a "low-key research preview", and that went pretty well.
00:58
In March, we followed that up with the launch of GPT-4, still the most capable model out in the world. [applause] -In the last few months, we launched voice and vision capabilities so that ChatGPT can now see, hear, and speak. [applause] -There's a lot, you don't have to clap each time. [laughter] -More recently, we launched DALL-E 3, the world's most advanced image model.
01:28
You can use it of course, inside of ChatGPT. For our enterprise customers, we launched ChatGPT Enterprise, which offers enterprise-grade security and privacy, higher speed GPT-4 access, longer context windows, a lot more. Today we've got about 2 million developers building on our API for a wide variety of use cases doing amazing stuff, over 92% of Fortune 500 companies building on our products, and we have about a hundred million weekly active users now on ChatGPT.
02:02
[applause] -What's incredible on that is we got there entirely through word of mouth. People just find it useful and tell their friends. OpenAI is the most advanced and the most widely used AI platform in the world now, but numbers never tell the whole picture on something like this. What's really important is how people use the products, how people are using AI, and so I'd like to show you a quick video.
02:30
-I actually wanted to write something to my dad in Tagalog. I want a non-romantic way to tell my parent that I love him and I also want to tell him that he can rely on me, but in a way that still has the respect of a child-to-parent relationship that you should have in Filipino culture and in Tagalog grammar.
02:53
When it's translated into Tagalog, "I love you very deeply and I will be with you no matter where the path leads." -I see some of the possibility, I was like, "Whoa." 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.
03:08
-The first thing that just blew my mind was it levels with you. That's something that a lot of people struggle to do. It opened my mind to just what every creative could do if they just had a person helping them out who listens. -This is to represent sickling hemoglobin. -You built that with ChatGPT? -ChatGPT built it with me.

03:31
-I started using it for daily activities like, "Hey, here's a picture of my fridge. Can you tell me what I'm missing? 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, "Wow, this thing is awesome.
03:46
" It could build spreadsheets. It could do anything. -I discovered Chatty about three months ago on my 100th birthday. Chatty is very friendly, very patient, very knowledgeable, and very quick. This has been a wonderful thing. -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.
04:14
Not only does it give me an answer, but it gives me an explanation. Didn't need tutoring as much. It gave me a life back. It gave me time for my family and time for me. -I have a chronic nerve thing on my whole left half of my body, I have nerve damage. I had a brain surgery. I have limited use of my left hand.
04:35
Now you can just have the integration of voice input. Then the newest one where you can have the back-and-forth dialogue, that's just maximum best interface for me. It's here. [music] [applause] -We love hearing the stories of how people are using the technology. It's really why we do all of this.
05:04
Now, on to the new stuff, and we have got a lot. [audience cheers] -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. Over the last year, we spent a lot of time talking to developers around the world. We've heard a lot of your feedback.
05:24
It's really informed what we have to show you today. Today, we are launching a new model, GPT-4 Turbo. [applause] -GPT-4 Turbo will address many of the things that you all have asked for. Let's go through what's new. We've got six major things to talk about for this part. Number one, context length.
05:52
A lot of people have tasks that require a much longer context length. GPT-4 supported up to 8K and in some cases up to 32K context length, 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. [applause] -That's 300 pages of a standard book, 16 times longer than our 8k context.
06:21
In addition to a longer context length, you'll notice that the model is much more accurate over a long context. Number two, more control. We've heard loud and clear that developers need more control over the model's responses and outputs. We've addressed that in a number of ways. We have a new feature called JSON Mode, which ensures that the model will respond with valid JSON.

06:47
This has been a huge developer request. It'll make calling APIs much easier. 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. We're also introducing a new feature called reproducible outputs. You can pass a seed parameter, and it'll make the model return consistent outputs.
07:10
This, of course, gives you a higher degree of control over model behavior. This rolls out in beta today. [applause] -In the coming weeks, we'll roll out a feature to let you view logprobs in the API. [applause] -All right. Number three, better world knowledge. You want these models to be able to access better knowledge about the world, so do we.
07:36
We're launching retrieval in the platform. You can bring knowledge from outside documents or databases 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 about the world ended in 2021. We will try to never let it get that out of date again.
07:55
GPT-4 Turbo has knowledge about the world up to April of 2023, and we will continue to improve that over time. Number four, new modalities. Surprising no one, DALL-E 3, GPT-4 Turbo with vision, and the new text-to-speech model are all going into the API today. [applause] -We have a handful of customers that have just started using DALL-E 3 to programmatically generate images and designs.
08:31
Today, Coke is launching a campaign that lets its customers generate Diwali cards using DALL-E 3, and of course, our safety systems help developers protect their applications against misuse. Those tools are available in the API. GPT-4 Turbo can now accept images as inputs via the API, can generate captions, classifications, and analysis.
08:52
For example, Be My Eyes uses this technology to help people who are blind or have low vision with their daily tasks like identifying products in front of them. With our new text-to-speech model, you'll be able to generate incredibly natural-sounding audio from text in the API with six preset voices to choose from.
09:14
I'll play an example. -Did you know that Alexander Graham Bell, the eminent inventor, was enchanted by the world of sounds. His ingenious mind led to the creation of the graphophone, which etches sounds onto wax, making voices whisper through time. -This is much more natural than anything else we've heard out there.
09:34
Voice can make apps more natural to interact with and more accessible. It also unlocks a lot of use cases like language learning, and voice assistance. Speaking of new modalities, we're also releasing

the next version of our open-source speech recognition model, Whisper V3 today, and it'll be coming soon to the API.

09:54

It features improved performance across many languages, and we think you're really going to like it. Number five, customization. Fine-tuning has been working really well for GPT-3.5 since we launched it a few months ago. Starting today, we're going to expand that to the 16K version of the model. Also, starting today, 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, but you may want a model to learn a completely new knowledge domain, or to use a lot of proprietary data. Today we're launching a new program called Custom Models.

10:40

With Custom Models, our researchers will work closely with a company to help them make a great custom model, especially for them, and their use case using our tools. This includes modifying every step of the model training process, 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, at least initially, it won't be cheap, but if you're excited to push things as far as they can currently go. Please get in touch with us, and we think we can do something pretty great.

11:18

Number six, higher rate limits. We're doubling the tokens per minute for all of our established GPT-4 customers, so it's easier to do more. You'll be able to request changes to further rate limits and quotas directly in your API account settings. In addition to these rate limits, it's important to do everything we can do to make you successful building on our platform.

11:42

We're introducing copyright shield. Copyright shield means that we will step in and defend our customers and pay the costs incurred, if you face legal claims or on copyright infringement, and this applies both to ChatGPT Enterprise and the API. Let me be clear, this is a good time to remind 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 and that's pricing. [laughter] -GPT-4 Turbo is the industry-leading model. It delivers a lot of improvements that we just covered and it's a smarter model than GPT-4.

12:32

We've heard from developers that there are a lot of things that they want to build, 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. A huge leap forward. I'm super excited to announce that we worked really hard on this and GPT-4 Turbo, a better model, is considerably cheaper than GPT-4 by a factor of 3x for prompt tokens.
12:58
[applause] -And 2x for completion tokens starting today. [applause] -The new pricing is 1¢ per 1,000 prompt tokens and 3¢ per 1,000 completion tokens. For most customers, that will lead to a blended rate more than 2.75 times cheaper to use for GPT-4 Turbo than GPT-4. We worked super hard to make this happen.
13:28
We hope you're as excited about it as we are. [applause] -We decided to prioritize price first because we had to choose one or the other, but we're going to work on speed next. We know that speed is important too. Soon you will notice GPT-4 Turbo becoming a lot faster. We're also decreasing the cost of GPT-3.5 Turbo 16K.
13:54
Also, input tokens are 3x less and output tokens are 2x less. Which means that GPT-3.5 16K is now cheaper than the previous GPT-3.5 4K model. Running a fine-tuned GPT-3.5 Turbo 16K version is also cheaper than the old fine-tuned 4K version. 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. In all of this, we're lucky to have a partner who is instrumental in making it happen. I'd like to bring out a special guest, Satya Nadella, the CEO of Microsoft. [audience cheers] [music] -Good to see you. -Thank you so much.
14:41
Thank you. -Satya, thanks so much for coming here. -It's fantastic to be here and Sam, congrats. I'm really looking forward to Turbo and everything else that you have coming. It's been just fantastic partnering with you guys. -Awesome. Two questions. I won't take too much of your time. How is Microsoft thinking about the partnership currently? -First- [laughter] --we love you guys.
15:06
[laughter] -Look, it's been fantastic for us. In fact, I remember the first time I think you reached out and said, "Hey, do you have some Azure credits?" We've come a long way from there. -Thank you for those. That was great. -You guys have built something magical. Quite frankly, there are two things for us when it comes to the partnership.
15:24
The first is these workloads. Even when I was listening backstage to how you're describing what's coming, even, it's just so different and new. I've been in this infrastructure business for three decades. -No one has ever seen infrastructure like this. -The workload, the pattern of the workload, 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, to the accelerators, to the network. Just really the shape of Azure is drastically changed and is changing rapidly in support of these models that you're building.

16:06

Our job, number one, is to build the best system 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. 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 on GPT.

16:29

We want to build our GitHub Copilot all as developers on top of OpenAI APIs. We are very, very committed to that. What does that mean to developers? Look, I always think of Microsoft as a platform company, a developer company, and a partner company. For example, we want to make GitHub Copilot available, the Enterprise edition available to all the attendees here so that they can try it out.

16:55

That's awesome. We are very excited about that. [applause] -You can count on us to build the best infrastructure in Azure with your API support and bring it to all of you. Even things like the Azure marketplace. 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? Anything you want -There are a couple of things for me that I think are going to be very, very key for us. One is I just described how the systems that are needed as you aggressively push forward on your roadmap requires us to be on the top of our game and we intend fully to commit ourselves deeply to making sure you all as builders of these foundation models have not only the best systems for training and inference,

17:55

but the most compute, so that you can keep pushing- -We appreciate that. --forward on the frontiers because I think that's the way we are going to make progress. The second thing I think both of us care about, in fact, quite frankly, the thing that excited both sides to come together is your mission and our mission.

18:11

Our mission is to empower every person and every organization on the planet to achieve more. To me, ultimately AI is only going to be useful if it truly does empower. I saw the video you played early. That was fantastic to hear those voices describe what AI meant for them and what they were able to achieve.

18:29

Ultimately, it's about being able to get the benefits of AI broadly disseminated to everyone, 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, 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.
18:47
-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. -Thank you so much. -See you. [applause] -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, we can't resist making some improvements to ChatGPT.
19:16
A small one, ChatGPT now uses GPT-4 Turbo with all the latest improvements, including the latest knowledge cutoff, which will continue to update. That's all live today. It can now browse the web when it needs to, write and run code, analyze data, take and generate images, and much more. We heard your feedback, that model picker, extremely annoying, that is gone starting today.
19:37
You will not have to click around the dropdown menu. All of this will just work together. Yes. [applause] -ChatGPT will just know what to use and when you need it, but that's not the main thing. Neither was price actually the main developer request. 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 about today.
20:06
We believe that if you give people better tools, they will do amazing things. We know that people want AI that is smarter, more personal, more customizable, can do more on your behalf. Eventually, you'll just ask the computer for what you need and it'll do all of these tasks for you. These capabilities are often talked in the AI field about as "agents.
20:29
" The upsides of this are going to be tremendous. At OpenAI, we really believe that gradual iterative deployment is the best way to address the safety issues, the safety challenges with AI. We think it's especially important to move carefully towards this future of agents. It's going to require a lot of technical work and a lot of thoughtful consideration by society.
20:51
Today, we're taking our first small step that moves us towards this future. We're thrilled to introduce GPTs. GPTs are tailored versions of ChatGPT for a specific purpose. You can build a GPT, a customized version of ChatGPT for almost anything with instructions, expanded knowledge, and actions, and then you can publish it for others to use.
21:20
Because they combine instructions, expanded knowledge, and actions, they can be more helpful to you. 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 or just have more fun and you'll be able to use them right within ChatGPT.

21:38
You can in effect program a GPT with language just by talking to it. It's easy to customize the behavior so that it fits what you want. This makes building them very accessible and it gives agency to everyone. We're going to show you what GPTs are, 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. First, let's look at a few examples. Our partners at Code.org are working hard to expand computer science in schools. They've got a curriculum that is used by tens of millions of students worldwide.
22:20
Code.org, crafted Lesson Planner GPT, to help teachers provide a more engaging experience for middle schoolers. If a teacher asks it to explain four loops in a creative way, it does just that. In this case, it'll do it in terms of a video game character repeatedly picking up coins. Super easy to understand for an 8th-grader.
22:40
As you can see, this GPT brings together Code.org's, extensive curriculum and expertise, and lets teachers adapt it to their needs quickly and easily. Next, Canva has built a GPT that lets you start designing by describing what you want in natural language. If you say, "Make a poster for a DevDay reception this afternoon, this evening," and you give it some details, it'll generate a few options to start with by hitting Canva's APIs.
23:08
Now, this concept may be familiar to some of you. We've evolved our plugins to be custom actions for GPTs. You can keep chatting with this to see different iterations, and when you see one you like, you can click through to Canva for the full design experience. Now we'd like to show you a GPT Live.
23:28
Zapier has built a GPT that lets you perform actions across 6,000 applications to unlock all kinds of integration possibilities. I'd like to introduce Jessica, one of our solutions architects, who is going to drive this demo. Welcome Jessica. [applause] -Thank you, Sam. Hello everyone. Thank you all. Thank you all for being here.
23:51
My name is Jessica Shieh. I work with partners and customers to bring their product alive. Today I can't wait to show you how hard we've been working on this, so let's get started. 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 and on the right-hand side you can see that's my calendar for today.
24:14
It's quite a day ever. I've already used this before, so it's actually already connected to my calendar. To start, I can ask, "What's on my schedule for today?" We build GPTs with security in mind. Before

it performs any action or share data, it will ask for your permission. Right here, I'm going to say allowed.

24:37

GPT is designed to take in your instructions, make the decision on which capability to call to perform that action, and then execute that for you. You can see right here, it's already connected to my calendar. It pulls into my information and then I've also prompted it to identify conflicts on my calendar.

24:57

You can see right here it actually was able to identify that. 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. Chasing GPUs." With that, I'm going to swap to my conversation with Sam and then I'm going to say, "Yes, please run that.

25:27

" Sam, did you get that? -I did. -Awesome. [applause] -This is only a glimpse of what is possible and I cannot wait to see what you all will build. Thank you. Back to you, Sam. [applause] -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, many more that will be created soon.

26:02

We know that many people who want to build a GPT don't know how to code. We've made it so that you can program a GPT just by having a conversation. 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.

26:20

I'd like to show you how to build one. All right. I want to create a GPT that helps give founders and developers advice when starting new projects. I'm going to go to create a GPT here, and this drops me into the GPT builder. I worked with founders for years at YC and still whenever I meet developers, the questions I get are always about, "How do I think about a business idea? Can you give me some advice?" I'm going to see if I can build a GPT to help with that.

26:53

To start, GPT builder asks me what I want to make, and I'm going to say, "I want to help startup founders think. through their business ideas and get advice. After the founder has gotten some advice, grill them on why they are not growing faster." [laughter] -All right. To start off, I just tell the GPT little bit about what I want here.

27:26

It's going to go off and start thinking about that, and it's going to write some detailed instructions for the GPT. It's also going to, let's see, ask me about a name. How do I feel about Startup Mentor? That's fine. "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.

27:46

You can see here on the right, in the preview mode that it's already starting to fill out the GPT. Where it says what it does, it has some ideas of additional questions that I could ask. [chuckles] It just generated a candidate. 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, how it can interact with users, I could talk about style here. What I'm going to say is, "I am going to upload transcripts of some lectures about startups I have given, please give advice based off of those.

28:39

" All right. Now, it's going to go figure out how to do that. 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 by the builder itself. You can see that there's capabilities here that I can enable. I could add custom actions. These are all fine to leave.

28:58

I'm going to upload a file. Here is a lecture that I picked that I gave with some startup advice, and I'm going to add that here. In terms of these questions, this is a dumb one. 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, which is be concise and constructive with feedback.

29:26

All right. Again, if we had more time, I'd show you a bunch of other things. This is a decent start. Now, we can try it out over on this preview tab. I will say, what's a common question? "What are three things to look for when hiring employees at an early-stage startup?" Now, it's going to look at that document I uploaded.

29:56

It'll also have of course all of the background knowledge of GPT-4. 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 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.

30:16

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 that I think would be useful, and then I can share it publicly. That's what it looks like to create a GPT [applause] -Thank you. 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 to make a bot that will do this and that'll be awesome.

30:45

" [laughter] -With GPTs, we're letting people easily share and discover all the fun ways that they use ChatGPT with the world. You can make private GPT like I just did, or you can share your creations publicly with a link for anyone to use, or if you're on ChatGPT Enterprise, you can make GPTs just

for your company.
31:11
Later this month we're going to launch the GPT store. Thank you. I appreciate that. [applause] -You can list a GPT there and we'll be able to feature the best and the most popular GPT. Of course, we'll make sure that GPTs in the store follow our policies before they're accessible. Revenue sharing is important to us.
31:40
We're going to pay people who build the most useful and the most used GPT a portion of our revenue. We're excited to foster a vibrant ecosystem with the GPT store, 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 is that we're bringing the same concept to the API. [applause] Many of you have already been building agent-like experiences on the API, for example, Shopify's Sidekick, which lets you take actions on the platform.
32:25
Discord's Clyde, lets Discord moderators create custom personalities for, and Snaps My AI, a customized chatbot that can be added to group chats and make recommendations. These experiences are great, but they have been hard to build. Sometimes taking months, teams of dozens of engineers, there's a lot to handle to make this custom assistant experience.
32:49
Today, we're making that a lot easier with our new Assistants API. [applause] -The Assistants API includes persistent threads, so they don't have to figure out how to deal with long conversation history, built-in retrieval, code interpreter, a working Python interpreter in a sandbox environment, and of course the improved function calling, that we talked about earlier.
33:17
We'd like to show you a demo of how this works. Here is Romain, our head of developer experience. Welcome, Romain. [music] [applause] -Thank you, Sam. Good morning. Wow. It's fantastic to see you all here. It's been so inspiring to see so many of you infusing AI into your apps. Today, we're launching new modalities in the API, but we are also very excited to improve the developer experience for you all to build assistive agents.
33:48
Let's dive right in. Imagine I'm building $1, travel app for global explorers, and this is the landing page. I've actually used GPT-4 to come up with these destination ideas. For those of you with a keen eye, these illustrations are generated programmatically using the new DALL-E 3 API available to all of you today.
34:08
It's pretty remarkable. Let's enhance this app by adding a very simple assistant to it. 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. Creating an assistant is easy, you just give it a name, some initial instructions, a model.

34:27

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. That's it. Our assistant is ready to go. Next, I can integrate with two new primitives of this Assistants API, threads and messages. Let's take a quick look at the code.

34:49

The process here is very simple. For each new user, I will create a new thread. As these users engage with their assistant, I will add their messages to the threads. Very simple. Then I can simply run the assistant at any time to stream the responses back to the app. We can return to the app and try that in action.

35:10

If I say, "Hey, let's go to Paris." All right. That's it. With just a few lines of code, users can now have a very specialized assistant right inside the app. 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, 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?" I'm going to have the assistant respond to that again.

35:54

Here, what's interesting is that the assistant knows about functions, including those to annotate the map that you see on the right. Now, all of these pins are dropping in real-time here. Yes, it's pretty cool. [applause] -That integration allows our natural language interface to interact fluidly with components and features of our app.

36:17

It truly showcases now the harmony you can build between AI and UI where the assistant is actually taking action. Let's talk about retrieval. Retrieval is about giving our assistant more knowledge beyond these immediate user messages. In fact, I got inspired and I already booked my tickets to Paris. I'm just going to drag and drop here this PDF.

36:40

While it's uploading, I can just sneak peek at it. Very typical United Flight ticket. Behind the scene here, what's happening is that retrieval is reading these files, and boom, the information about this PDF appeared on the screen. [applause] -This is, of course, a very tiny PDF, but Assistants 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. 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, you need to set up chunking algorithm. Now all of that is taken care of.

37:25

There's more than retrieval with every API call, you usually need to resend the entire conversation history, which means setting up a key-value store, that means handling the context windows, serializing messages, and so forth. That complexity now completely goes away with this new stateful API. Just because OpenAI is managing this API, does not mean it's a black box.

37:47

In fact, you can see the steps that the tools are taking right inside your developer dashboard. Here, if I go ahead and click on threads, this is the thread I believe we're currently working on and see, these are all the steps, including the functions 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, that gives the AI the ability to write and execute code on the fly, but even generate files. Let's see that in action. If I say here, "Hey, we'll be four friends staying at this Airbnb, what's my share of it plus my flights?" All right.

38:43

Now, here, what's happening is that Code interpreter noticed that it should write some code to answer this query. Now it's computing the number of days in Paris, number of friends. It's also doing some exchange rate calculation behind the scene to get the sensor for us. Not the most complex math, but you get the picture.

39:01

Imagine you're building a very complex finance app that's crunching countless numbers, plotting charts, so really any task that you'd normally tackle with code, then Code Interpreter will work great for you. All right. I think my trip to Paris is solid. To recap here, we've just seen how you can quickly create an assistant that manages state for your user conversations, 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 that we're launching today. While working on DevDay, I built a small custom assistant that knows everything about this event, but instead of having a chat interface while running around all day today, 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. On the left, I'm actually going to bring up my terminal log so you can see what's happening behind the

scenes. 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? -Hey everyone, welcome to DevDay.
40:24
It's awesome to have you all here. Let's make it an incredible day. [applause] -Isn't that impressive? You have six unique and rich voices to choose from in the API, each speaking multiple languages, so you can really find the perfect fit for your app. On my laptop here on the left, you can see the logs of what's happening behind the scenes, too.
40:47
I'm using Whisper to convert the voice inputs into text, an assistant with GPT-4 Turbo, and finally, the new TTS API to make it speak. Thanks to function calling, things get even more interesting when the assistant can connect to the internet and take real actions for users. Let's do something even more exciting here together.
41:08
How about this? Hey, Assistant, can you randomly select five DevDay attendees here and give them $500 in OpenAI credits? [laughter] -Yes, checking the list of attendees. [laughter] -Done. I picked five DevDay attendees and added $500 of API credits 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. That's it. A quick overview today of the new Assistants API combined with some of the new tools and modalities that we launched, all starting with the simplicity of a rich text or voice conversation for you end users. We really can't wait to see what you build, and congrats to our lucky winners.
42:00
Actually, you know what? you're all part of this amazing OpenAI community here so I'm just going to talk to my assistant one last time before I step off the stage. Hey Assistant, can you actually give everyone here in the audience $500 in OpenAI credits? -Sounds great. Let me go through everyone. [applause] -All right, that function will keep running, but I've run out of time.
42:32
Thank you so much, everyone. Have a great day. Back to you, Sam. -Pretty cool, huh? [audience cheers] -All right, so that Assistants API goes into beta today, and we are super excited to see what you all do with it, anybody can enable it. Over time, GPTs and Assistants are precursors to agents are going to be able to do much much more.
43:06
They'll gradually be able to plan and to perform more complex actions on your behalf. As I mentioned before, we really believe in the importance of gradual iterative deployment. 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, as they become more capable.
43:26
As we've always done, we'll continue to update our systems based off of your feedback. We're super

excited that we got to share all of this with you today. We introduced GPTs, custom versions of GPT that combine instructions, extended knowledge and actions. We launched the Assistants API to make it easier to build assistive experiences with your own apps.

43:50

These are your first steps towards AI agents and we'll be increasing their capabilities over time. We introduced a new GPT-4 Turbo model that delivers improved function calling, knowledge, lowered pricing, new modalities, and more. We're deepening our partnership with Microsoft. In closing, I wanted to take a minute to thank the team that creates all of this.

44:14

OpenAI has got remarkable talent density, but still, it takes a huge amount of hard work and coordination to make all this happen. I truly believe that I've got the best colleagues in the world. I feel incredibly grateful to get to work with them. 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 to build so much for all of us. We talked about earlier how if you give people better tools, they can change the world. We believe that AI will be about individual empowerment and agency at a scale that we've never seen before and that will elevate humanity to a scale that we've never seen before either.

44:56

We'll be able to do more, to create more, and to have more. As intelligence gets integrated everywhere, we will all have superpowers on demand. We're excited to see what you all will do with this technology and to discover the new future that we're all going to architect together. We hope that you'll come back next year.

45:15

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. Thank you for coming here today. [applause] [music]