aws startups

모든 스타트업을 위한 생성형 AI

AWS 기반 생성형 AI로 손쉽게 구축 및 규모 조정

목차

소개

스타트업을 위한 생성형 AI의 힘과 전망

기계 학습(ML) 패러다임 전환의 씨앗은 몇십 년 전부터 존재해 왔습니다. 하지만 확장 가능한 컴퓨팅 용량을 이용할 수 있게 되고, 데이터가 폭증하고, ML 기술이 급속히 발전하면서 모든 업종의 고객이 본격적으로 자사 비즈니스 혁신을 도모하게 됐습니다. OpenAI의 ChatGPT나 Google의 Bard와 같은 생성형 AI 도구가 널리 주목을 끌었으며, 적극적인 투자자도 계속 늘어나는 추세입니다.

2023년 3월 피치북 보고서에 따르면, 다수의 벤처 캐피털리스트(VC)가 암호화폐나 블록체인과 같은 다른 기술 혁신에 대해 보여온 투자 열기는 식어가고 있습니다. 전반적으로 2023년 1분기 VC 자금은 전년 동기 대비 53% 감소했지만, 생성형 AI는 밝은 전망을 보이고 있습니다. 2022년 한 해에만 이 부문에 대한 투자는 21억 달러에 달했으며, 이는 2020년 대비 425% 증가한 수치입니다.[1]2023년까지 텍스트 기반 생성에 특화된 대규모 언어 모델 스타트업인 AI21 Labs는 1억 5,500만 달러 규모의 시리즈 C 투자를 유치했으며, 자체 레이크하우스 생성형 AI 프로그램을 갖춘 데이터 분석 플랫폼 개발사인 데이터브릭스는 Nvidia와 T. Rowe Price를 비롯한 투자자로부터 5억 달러 이상의 자금을 조달했습니다. 그리고 의료 분야에서 사용할 생성형 AI 코파일럿을 만든 Corti는 2023년 하반기에 6천만 달러 규모의 시리즈 B를 모금했습니다.[2]

생성형 AI의 검증된 이점을 감안하면 이런 VC 수치는 전혀 놀랍지 않습니다. 창업자는 이 기술을 사용하여 시장 출시 기간을 단축하고, 혁신을 촉진하고, 고객 경험을 개인화하고, 비용을 최적화할 수 있습니다. 이제 인공 지능(AI)이 생산성을 강화해 줄 것이라 굳게 믿는 사업주(전체의 60% 돌파)의 대열에 합류할 때가 되었습니다.[3]

이 eBook은 자사에 생성형 AI 솔루션을 통합하는 데 관심이 있는 스타트업 리더를 위한 가이드입니다. eBook에는 생성형 AI를 활용한 경험이 있는 스타트업의 사례는 물론, 모든 단계의 스타트업이 생성형 AI 여정의 파트너로 Amazon Web Services(AWS)를 선택한 이유도 담겨 있습니다. 우선, 이 기술의 기본적인 요소부터 살펴보겠습니다.

[1] 'Vertical Snapshot: Generative AI(업종별 스냅샷: 생성형 AI)', PitchBook, 2023년 3월
[2] "Robbins, J., 'Generating less momentum? Generative AI deal count dips in Q3(모멘텀 생성 약화? 3분기 생성형 AI 거래 수 감소),' PitchBook, 2023년 10월
[3] Haan, K., '24 Top AI Statistics and Trends In 2023(2023년에 예상되는 24가지 AI 통계 및 트렌드),' Forbes, 2023년 4월
[4] 'Generative AI Market(생성형 AI 마켓),' Polaris Market Research, 2023년 1월

aws startups

전략적 중요성

전 세계 다양한 산업 분야의 스타트업이 더 빠르게 혁신하고, 직원 생산성을 높이고, 창의성을 발휘하고, 비즈니스 프로세스를 최적화하기 위해 생성형 AI를 활용하는 방안을 모색하고 있습니다.

그러나 대부분의 스타트업은 이러한 이점을 얻기 위한 경로를 확실히 알지 못합니다.

대부분의 창업자는 경쟁사가 우위를 점하기 전에 신속하게 생성형 AI에 투자해야 한다는 필요성에 공감합니다. 그러나 기술 도입, 용도, 결과 실현 및 측정 전략을 제대로 개발한 조직은 거의 없습니다.

스타트업이 어떻게 지금 바로 생성형 AI의 비즈니스 가치를 실현하는지 살펴보세요. 시장의 발전 속도에 발맞추고 경쟁에서 앞서나갈 수 있을 것입니다.

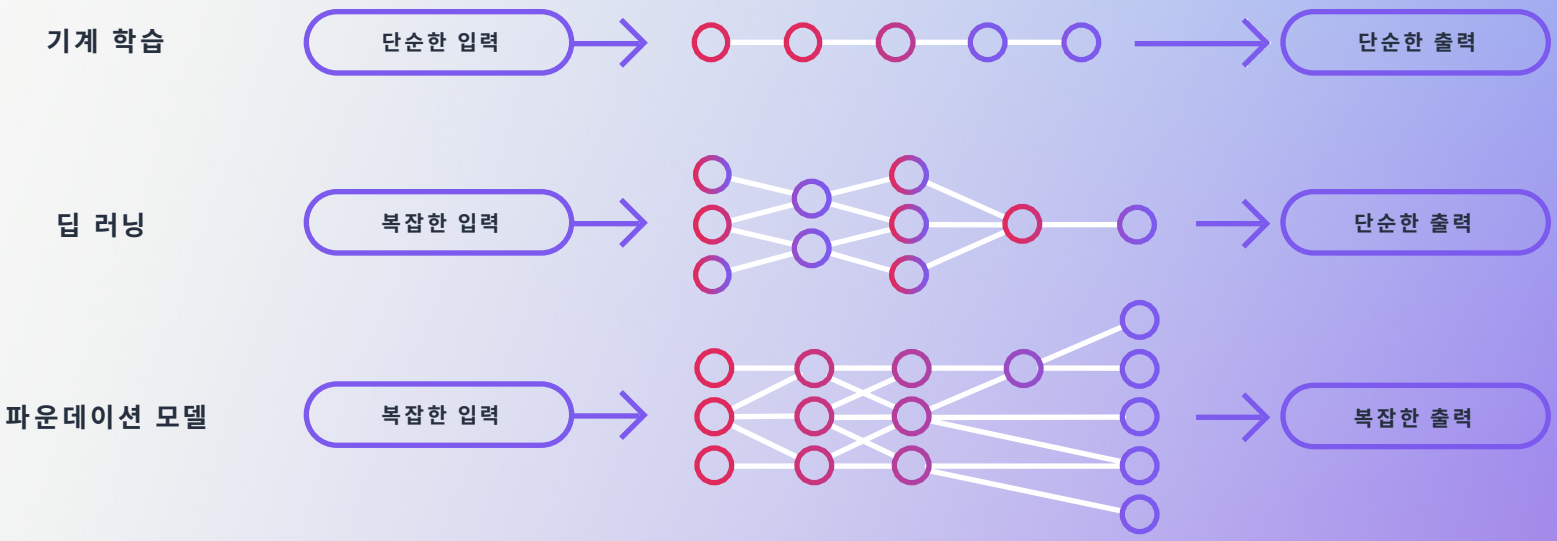
생성형 AI의 이해

스타트업이 생성형 AI의 비즈니스 가치를 완전히 실현하려면 먼저 AI 기술의 작동 원리를 근본적으로 이해해야 합니다. 생성형 AI는 대화, 스토리, 이미지, 동영상, 음악을 포함하여 새로운 콘텐츠와 아이디어를 생성할 수 있는 알고리즘을 의미합니다. 생성형 AI는 방대한 양의 데이터에 대해 사전 훈련을 마친 초대형 ML 모델을 기반으로 구동되는데, 이 ML 모델을 보통 **파운데이션 모델(FM)**이라고 합니다.

기존 ML에서는 숫자 값과 같은 간단한 입력을 예측 값과 같은 간단한 출력에 매핑했습니다. 딥 러닝의 출현으로, 동영상이나 이미지와 같은 복잡한 입력을 수행하고 비교적 간단한 출력(예: 이미지의 고양이 존재 여부)에 매핑할 수 있게 되었습니다.

생성형 AI를 사용하면 대량의 복잡한 데이터를 활용하여 보다 진보한 방식으로 정보를 포착하고 제공할 수 있습니다. 즉, 긴 문서를 요약하여 핵심 정보를 추출하는 등 복잡한 입력을 복잡한 출력에 매핑할 수 있게 된 것입니다.

텍스트 기반 생성형 AI 시스템은 **대규모 언어 모델(LLM)**이라는 특정 유형의 FM을 사용합니다. LLM은 다양한 분야를 넘나들며 광범위한 작업을 수행할 수 있습니다. 예를 들어, 코드를 작성하거나, 수학 문제를 풀거나, 대화에 참여하거나, 문서의 정보를 분석하여 질문에 답할 수 있습니다.

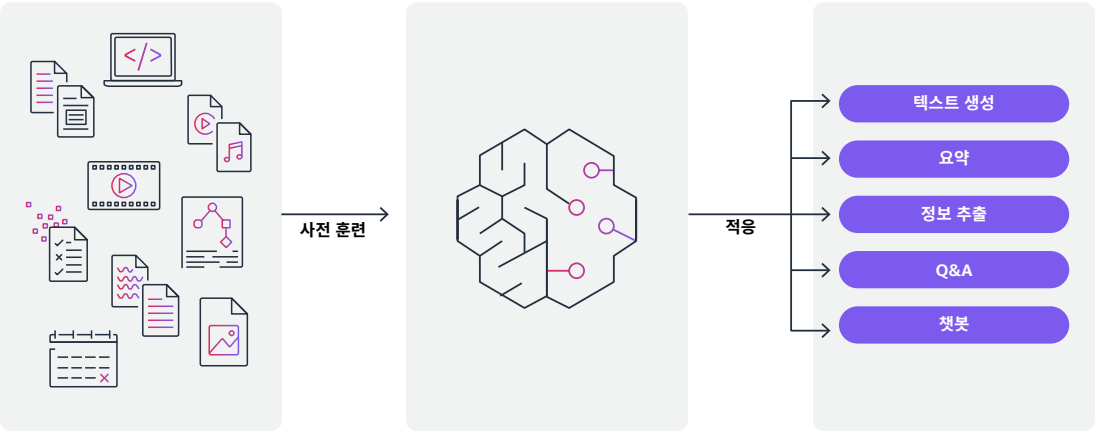


데이터를 차별화 요소로 만들기

스타트업의 비즈니스 요구 사항에 맞는 생성형 AI 애플리케이션을 구축하고자 한다면 조직의 데이터는 전략적 자산이라 할 수 있습니다. 스타트업 내 독점 데이터로 FM을 사용자 지정하고 미세 조정하여 '즉시 사용 가능한 FM' 대비 보다 차별화된 경험을 제공할 수 있습니다. 예를 들어, 쇼핑객 선호도를 추적하는 식료품 체인이라면 FM을 사용자 지정하여 경쟁사의 제품과 완전히 차별화된 더 나은 추천 엔진을 제작할 수 있습니다.

또한, 스타트업은 사용자 지정 FM을 활용하여 브랜드 특유의 분위기와 스타일을 구현하는 고유 콘텐츠를 쉽게 만들 수도 있습니다. 가령, 내부 배포용 일일 활동 보고서를 자동 생성해야 하는 테크 스타트업은 과거 보고서를 포함하는 독점 데이터로 FM을 사용자 지정할 수 있습니다. 그런 다음 FM은 이러한 보고서를 읽는 방법과 보고서를 생성하는 데 사용한 데이터를 학습하여 조직의 요구 사항을 더욱 잘 반영하는 보고서를 제공할 수 있었습니다.

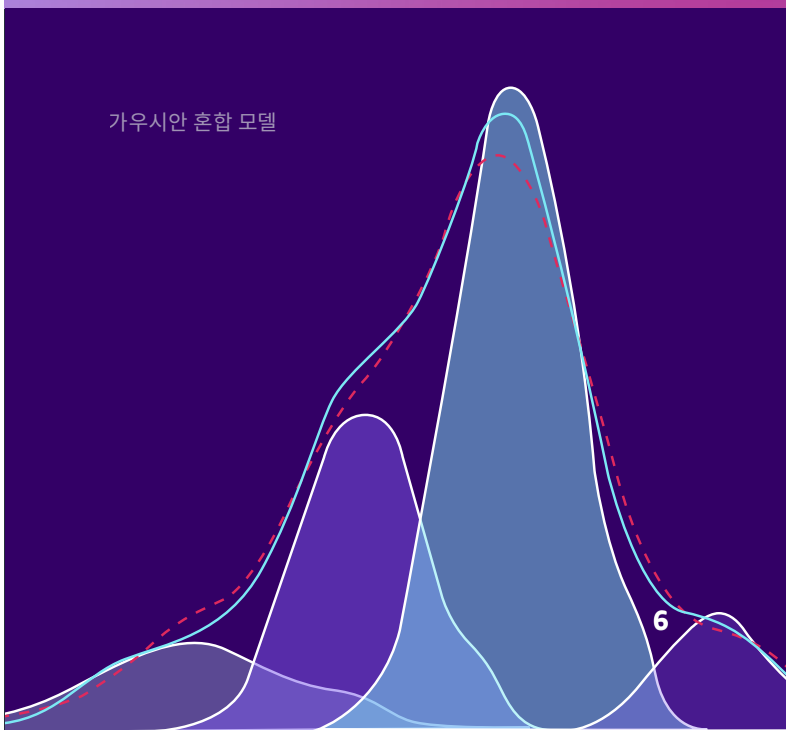
이제 기술의 작동 원리를 대략적으로 알아보았으니, 스타트업에서 생성형 AI를 어떻게 업무에 활용할 수 있을지 살펴보겠습니다.



aws startups

생성형 AI의 역사적 순간:

생성형 AI 애플리케이션을 만드는 데 사용되는 오늘날의 FM은 AI 혁신의 오랜 역사를 토대로 구축되었습니다. 생성형 AI 기능을 갖춘 초기 모델 중 2가지는 1950년대에 개발된 은닉 마르코프 모델(HMM)과 가우시안 혼합 모델(GMM)입니다. HMM은 알려진 데이터를 사용하여 알려지지 않은 데이터를 합리적으로 추측합니다(예: 결과를 기반으로 카드 플레이어가 부정행위를 하는지 예측). GMM은 데이터 그룹(예: 음악 재생 목록)과 해당 데이터 내의 하위 그룹(예: 장르)을 검사하여 알려지지 않은 정보(예: '이 노래는 랩입니다')를 추론합니다. 이 두 모델은 지금도 사용됩니다.



생성형 AI의 비즈니스 기능

다양한 단계의 스타트업이 여러 방법으로 생산성을 높이고 비즈니스 가치를 창출하기 위해 생성형 AI를 사용하고 있습니다.

혁신과 창의성
제안 받기, 프로토타입 생성 및 혁신적인 개념 살펴보기

가상 비서
사람과 같은 반응으로 고객 경험 향상

코드 생성
AI 코딩 컴패니언으로 개발자 생산성 57% 향상 **Amazon CodeWhisperer**[5]

대화형 검색
모든 회사 정보에서 인사이트 추출

고객 센터 분석
고객 통화에서 인사이트 요약 및 추출

콘텐츠 생성
텍스트, 이미지, 비디오 및 음악 생성

개인화
개인화된 권장 사항 개선 및 맞춤형 콘텐츠 생성

생성형 AI의 역사적 순간:

또 다른 초기 생성형 AI 모델로는 MIT 교수가 1964~1966년에 개발한 챗봇(구 명칭: '채터봇')인 ELIZA가 있습니다. 영화 피그말리온과 마이 페어 레이디의 주인공 'Eliza Doolittle'과 이름이 같은 ELIZA 프로그램은 인간의 상호 작용으로부터 '배우며' 더욱 정교해졌습니다. ELIZA의 가장 유명한 사용 사례는 초기 정신 의학 인터뷰를 수행하는 치료사의 행동을 모방하는 것입니다(사용자는 환자 역할).

[5] Amazon이 Amazon CodeWhisperer 평가판 중 수행한 '생산성 챌린지'에서 수집한 데이터

생성형 AI를 위한 비즈니스 고려 사항

스타트업에 가장 유용한 생성형 AI의 기능을 식별하고 비즈니스 프로세스에 구현하는 전략을 수립하면서, 생성형 AI 애플리케이션 개발에 사용할 FM을 결정해야 합니다.

FM을 지원하기 위해 사용할 인프라도 신중히 고려해야 합니다. 성능 요건을 충족하는 경제적인 인프라는 비즈니스에서 사용하는 모델에 도움이 됩니다.

생성형 AI 애플리케이션 생성에 사용되는 FM을 평가할 때는 다음을 제공하는 모델을 찾아보세요.

1. 보안 및 개인 정보 보호 기능이 내장되어 생성형 AI 애플리케이션을 쉽게 구축하고 확장할 수 있는 방법
2. 자체 모델을 훈련하고 대규모로 추론을 실행할 수 있는 고성능의 저비용 인프라
3. 업무 수행 방식을 혁신하는 생성형 AI 기반의 애플리케이션
4. 차별화 요소로서의 데이터

책임 있는 AI, 보안 및 개인정보 보호

방대한 크기와 개방형 특성을 가진 FM은 개발 주기 전체에서 정확성, 공정성, 지식 재산(IP) 고려 사항, 할루시네이션, 유해성, 개인정보 보호와 같은 책임 있는 AI 관련 우려를 정의, 측정 및 완화하는 데 있어 새로운 문제를 제기합니다. 예를 들어, 공정성의 문제를 살펴보겠습니다. 우리는 의사를 지칭할 때 남성 대명사와 여성 대명사를 같은 비율로 사용하라고 LLM에 요청할 수 있을까요? 프롬프트에서 의사가 수염이 있다고 설명하는 경우에도 동일하게 적용해야 할까요? 그리고 다른 직업의 경우도 마찬가지일까요? 미국여자프로농구(WNBA)는 어떨까요? LLM의 맥락에서 공정성을 단순히 정의하기란 어렵습니다. 그렇기에 새로운 접근 방식과 솔루션이 필요합니다.

생성형 AI 기술과 그 기술의 사용법이 계속 진화하면서, 새로운 문제가 발생하여 관심과 완화를 위한 노력이 추가로 필요할 것입니다. 이러한 과제를 해결하고 혁신을 촉진하고자 **학계, 산업계, 정부 파트너**가 함께 협력하여 생성형 AI가 책임감 있고 개인정보를 보호하는 안전한 방식으로 계속 진화할 수 있도록 새로운 솔루션과 개념을 모색하고 있습니다.

데이터 개인정보 보호와 보안은 생성형 AI를 책임감 있게 확장하는 데 매우 중요합니다. 모델을 사용자 지정하고 미세 조정할 때 스타트업은 데이터의 사용처와 사용 방법을 파악할 필요가 있습니다. 스타트업은 보유한 비공개 데이터가 공개 모델을 학습하는 데 사용되지 않으며 고객 데이터가 비공개로 유지된다는 확신을 가져야 합니다. 스타트업이 비즈니스 애플리케이션을 실행할 수 있으려면 처음부터 보안, 확장성, 개인정보 보호가 기본으로 제공되어야 합니다.

블로그 게시물 Responsible AI in the Generative Era 보기
더 알아보기 ›

생성형 AI의 역사적 순간:

2014년, 최초의 생성적 적대 신경망(GAN)이 개발되면서 생성형 AI는 큰 돌파구를 찾았습니다. GAN은 두 모델('생성자'와 '구분자')이 경쟁하는 제로섬 게임입니다. 생성자는 점점 더 '진짜'로 보이는 콘텐츠를 제조하는 반면, 구분자는 상대의 기술을 분석하여 가짜를 더 효과적으로 가려냅니다. AI의 훈련에 다른 AI를 사용하는 참신한 접근법은 혁신 그 자체였으며, GAN은 이렇게 디지털 이미지의 새로운 시대를 열었습니다.

AWS로 생성형 AI의 성공을 이끄는 비결

AWS와 함께하면 스타트업에서 생성형 AI의 온전한 비즈니스 가치를 실현할 수 있습니다. 애플리케이션을 재구성하고, 새로운 형태의 고객 경험을 만들고, 전례 없는 수준으로 생산성을 높이고, 궁극적으로 스타트업을 혁신하세요.

경험과 전문성

AWS의 핵심 장점 중 하나는 20년 이상의 집중 투자를 통해 쌓아온 풍부한 AI 자산입니다. 실제로 10만이 넘는 고객이 현재 AWS를 AI에 사용하고 있습니다.

AWS를 뒷받침하는 Amazon은 ML 기능을 활용하여 전자 상거래 추천 엔진을 구동하고, 주문 처리 센터에서 로봇 픽업 경로를 최적화합니다. 나아가 ML은 Amazon의 공급망, 예측 및 용량 계획을 알려줍니다.

딥 러닝은 Amazon Prime Air 드론 배송 시스템과 Amazon Go 컴퓨터 비전(CV) 기술에도 적용되어 고객이 기존의 결제 과정 없이 물건을 골라 매장을 나가는 혁신적인 소매 경험을 선사합니다. 또한, 30개가 넘는 여러 ML 시스템에서 사용할 수 있는 Alexa는 매주 수십억 차례의 작업을 수행하며 고객을 지원합니다.

수천 명의 전담 ML 엔지니어를 보유한 Amazon과 AWS에서 AI는 기업 문화에 깊숙이 통합되어 미래를 만들어 나가고 있습니다.

현재

10만

여 고객이 AI를 위해 AWS 사용

스타트업이 AWS로 구축하는 이유

모든 단계의 스타트업은 여러 가지 이유로 AWS에서 생성형 AI 및 기타 AI 애플리케이션을 구축하기로 결정합니다. 고객의 말을 빌리자면, AWS를 기반으로 구축할 때 이점은 다음과 같습니다.

보안 및 개인정보 보호 기능이 내장된 생성형 AI 애플리케이션을 구축하고 확장하는 가장 쉬운 방법

Amazon Bedrock은 고객이 FM을 사용하여 생성형 AI 기반 애플리케이션을 구축하고 규모를 조정하는 가장 쉬운 방법입니다. Bedrock은 API를 통해 **Amazon Titan** FM 및 AI21 Labs, Anthropic, Cohere, Stability AI, Meta와 같은 주요 AI 기업의 모델에 액세스할 수 있도록 지원합니다. Bedrock을 사용하는 고객은 현재 지원되는 가장 유연하고 안전한 클라우드 컴퓨팅 환경으로 설계된 AWS의 이점을 활용할 수 있습니다. **Agents for Amazon Bedrock**은 개발자가 독점 지식 소스를 기반으로 최신 답변을 제공하고 광범위한 사용 사례의 작업을 완수하는 생성형 AI 애플리케이션을 쉽게 만들도록 지원하는 완전관리형 기능입니다.

생성형 AI를 위한 고성능 저비용 인프라

AWS는 수년 동안 ML 워크로드에 대해 최고 수준의 성능과 비용 최적화를 제공하는 실리콘 개발에 투자해 왔습니다. 그 결과로 탄생한 **AWS Trainium**과 **AWS Inferentia**는 클라우드에서 모델 훈련과 추론 실행을 가장 저렴한 비용으로 제공합니다. AWS가 개발한 **Amazon Elastic Compute Cloud**(Amazon EC2) 인스턴스로 인해 사용자는 모델 훈련과 추론 실행 기능을 이용할 수 있습니다. 예를 들어, Trainium을 기반으로 한 **Amazon EC2 Trn1** 인스턴스는 훈련 비용을 최대 50% 절감하며[6], AWS Inferentia2를 기반으로 한 **Amazon EC2 Inf2** 인스턴스는 추론당 비용을 최대 40% 줄여줍니다[7].

차별화 요소로서의 데이터

AWS를 사용하면 스타트업의 데이터를 전략적 자산으로 사용하여 FM을 사용자 지정하고 더욱 차별화된 경험을 구축할 수 있습니다. 일반 생성형 AI 애플리케이션과 비즈니스와 고객을 진정으로 이해하는 애플리케이션 간의 차이를 보여주는 것은 바로 데이터입니다. 또한 가장 포괄적인 데이터 및 AI 서비스 세트로 데이터를 사용하여 AWS에서 FM을 안전하게 사용자 지정하고 비즈니스, 데이터 및 고객에 전문화된 모델을 구축할 수 있습니다.

업무 수행 방식을 혁신하는 생성형 AI 기반의 애플리케이션

AWS는 고객이 생성형 AI를 사용하여 작업을 완료하는 방식을 혁신하는 강력한 신규 애플리케이션을 구축하고 있습니다. **Amazon CodeWhisperer**를 사용하여 코딩을 간소화하고, **Amazon QuickSight Generative BI**를 사용하여 비즈니스 인텔리전스를 간소화하며, **AWS HealthScribe**를 사용하여 의료 기술 스타트업의 임상 효율성을 개선하는 목적별 대화형 에이전트로 생산성을 강화합니다. 보안, 개인 정보 보호, 책임 있는 AI를 가장 우선시하고 기존 데이터 소스 및 애플리케이션으로 쉽게 사용자 지정 및 통합할 수 있어 스타트업은 번거로운 작업 없이 생성형 AI를 신속하게 활용할 수 있습니다.

스타트업을 위한 책임 있는 AI 관련 추가 리소스:

추가 생성형 AI 리소스 ›

추가 AI 리소스 ›

aws startups

[6] 동급 Amazon EC2 인스턴스 대비

[7] 이전 세대 AWS Inferentia 기반 인스턴스 대비

AWS 생성형 AI 서비스

다음을 포함한 여러 AWS 기술을 통해 생성형 AI 애플리케이션을 지원합니다.

Amazon Bedrock ›

FM으로 생성형 AI 애플리케이션을 구축하고 확장합니다. Bedrock은 다음과 같이 다양한 FM을 지원합니다.

- **Amazon Titan**: 텍스트 요약, 생성, 분류, 개방형 Q&A, 정보 추출, 임베딩 및 검색용 FM
- **AI21 Labs Jurassic-2 Multilingual LLM**: 다양한 언어로 텍스트를 생성하는 FM
- **Anthropic Claude 2**: 정직하고 책임 있는 AI 시스템을 훈련하기 위한 연구에 기반한 대화, 질문 잡변 및 워크플로 자동화용 LLM
- **Stability AI Stable Diffusion**: 독특하고 사실적인 고품질 이미지, 아트, 로고 및 디자인 생성
- **Cohere Command + Embed:** 비즈니스 애플리케이션용 텍스트 생성 모델 및 100개 이상의 언어로 검색, 클러스터링 또는 분류하기 위한 임베딩 모델
- **Meta Llama 2**: 질문과 답변, 독해와 같은 자연어 작업을 위해 사전 훈련되고 미세 조정된 LLM

AWS Trainium: 이 ML 모델 가속기[8]를 통해 비용을 최대 50% 절감하며 모델 훈련 시간 단축

AWS Inferentia2: 이 가속기를 사용하여 추론당 비용을 최대 40% 절감하고 고성능 FM 추론 실행[9]

Amazon CodeWhisperer: 개인 개발자는 무료로 사용할 수 있는 AI 코딩 도우미를 통해 보안을 강화하는 동시에 57% 더 빠르게 애플리케이션 개발[10]

Amazon QuickSight 생성형 BI: Amazon QuickSight의 생성형 BI 기능을 사용하여 기존의 다단계 비즈니스 인텔리전스(BI) 작업을 직관적이고 강력한 자연어 환경으로 전환합니다.

Amazon SageMaker: 관리형 인프라와 도구로 자체 FM을 구축하여 확장 가능하고, 신뢰할 수 있으며, 안전한 모델 구축, 훈련, 배포 가속화

Amazon SageMaker JumpStart: ML을 빠르게 시작할 수 있도록 알고리즘, 모델 및 ML 솔루션에 대한 액세스를 제공하는 ML 허브입니다. SageMaker JumpStart를 통해 ML 실무자는 **공개적으로 사용 가능한 다양한 FM** 중에서 선택할 수 있습니다. ML 실무자는 네트워크 격리 환경에서 전용 SageMaker 인스턴스에 FM을 배포하고, 모델 훈련과 배포를 위해 SageMaker를 사용하여 모델을 사용자 지정할 수 있습니다.

[8] AWS Trainium은 동급 Amazon EC2 인스턴스에 비해 훈련 비용을 최대 50% 절감
[9] AWS Inferentia는 동급 Amazon EC2 인스턴스에 비해 추론당 비용을 최대 40% 절감
[10] Amazon이 Amazon CodeWhisperer 평가판 중 수행한 '생산성 챌린지'에서 수집한 데이터

고객 사례

생성형 AI로 무엇이 가능한지 증명해 보이고 있는 스타트업

고객 성공 사례

크고 작은 스타트업이 자사 비즈니스에 생성형 AI를 통합해 혁신의 속도를 높이고 경쟁사보다 유리한 경쟁 우위를 구축하고 있습니다. AWS가 이러한 혁명적인 기술을 잘 활용하도록 스타트업을 어떻게 돕고 있는지 네 가지 예를 들어 소개합니다.

고객 성공 사례

AWS 솔루션으로 성공을 위한 도약을 이룬 InsightFinder

AI 기반 예측형 관측 플랫폼 스타트업 **InsightFinder**는 이 플랫폼을 이용하는 학생과 교사의 수가 급속도로 늘어나면서 규모 조정 문제에 직면하게 되었습니다. 회사에 전송되는 알림을 필터링할 내부 인프라가 없었던 것입니다. InsightFinder는 자사 엔진을 **Amazon CloudWatch** 데이터와 연결하면서 필수적인 인사이트를 신속하고 간편하게 받아볼 수 있었습니다.

사례 읽기 ›

aws startups

InsightFinder

"많은 AI 기술 기업이 하드웨어 리소스에 거액을 투자해야 한다고 생각합니다. 하지만 우리는 [AWS와 함께한 결과] 고성능 엔진을 구축할 수 있었고, 그러면서도 적당한 비용을 초과하지 않았습니다."

Helen Gu, InsightFinder Founder

고객 성공 사례

AWS 기계 학습 솔루션을 사용해 최신 사기 행위 방지 앱을 구축한 Fraud.net

사기 행위 및 규정 준수 플랫폼인 **Fraud.net**은 수많은 대출 기관, 은행, 결제 처리 기관, 디지털 상거래 기업과 그 고객에게 피해를 주는 사기 행위 발생률 문제를 해결한다는 취지로 창립했습니다. Fraud.net은 이 목표를 이루는 데 가장 큰 방해가 되는 장애물은 데이터에 대한 투명성 부족이라는 사실을 깨달았습니다. 그래서 신속하게 배포할 수 있고 확장 가능하며 안전한 플랫폼을 구축해, 이 플랫폼에서 사기 행위 데이터를 통합하고 실행 가능한 인사이트를 도출하고자 했습니다. 이 스타트업은 AWS의 이벤트 기반 아키텍처를 활용해 이벤트 수에 따라 규모를 확장하거나 축소할 수 있게 했습니다. 또한 컴퓨팅에는 Amazon EC2와 Lambda, 고도로 확장 가능한 객체 스토리지에는 Amazon S3 등 AWS 솔루션을 사용했습니다. 이러한 솔루션 덕분에 고객 레벨, 기관 레벨, 기관 간 레벨 등 세 가지 레벨의 데이터를 통합하고 분석할 수 있었습니다.

사례 읽기 ›

aws startups

Fraud.net

"AWS의 도움을 받아 초당 수천 개의 트랜잭션을 처리하게 되었습니다. 3, 4년 전만 해도 사실상 불가능했던 규모죠."

Whitney Anderson, Fraud.net Co-Founder & CEO

고객 성공 사례

SageMaker의 DeepSpeed를 사용해 지연 시간이 짧은 GPT-J 추론을 실현한 Mantium

AI 애플리케이션 구축 및 관리를 위한 클라우드 플랫폼 제공업체인 글로벌 기업 **Mantium**은 크고 작은 기업을 도와 AI 애플리케이션을 구축하고, 전에는 불가능했던 수준으로 자동화를 빠르고 손쉽게 실행하고 있습니다. 하지만 Mantium은 한 가지 난관에 직면했습니다. 바로 오픈 소스 모델은 프로덕션 등급 성능에 맞춰 고안되는 경우가 드물다는 것입니다. GPT-J처럼 최신형 텍스트 생성을 지원하는 생성형 사전 훈련 트랜스포머에는 대응 지연 시간이 가장 중대한 장애물입니다. 이 때문에 프로덕션 배포가 실용성이 떨어지고, 심하면 실행이 불가능할 수도 있습니다. Mantium에서는 DeepSpeed 추론 엔진을 활용해 최적화된 CUDA 커널을 Hugging Face 트랜스포머 GPT-J 구현에 주입하여 GPT-J를 기반으로 텍스트 생성 속도를 대폭 높일 수 있었습니다.

사례 읽기 ›

MANTIUM

"DeepSpeed 추론 엔진은 SageMaker 추론 엔드포인트에 간단하게 통합할 수 있습니다. SageMaker를 사용하면 사용자 지정 추론 엔드포인트를 매우 쉽게 배포할 수 있고, DeepSpeed 통합은 종속성을 포함하고 코드를 몇 줄 쓰는 것으로 간단하게 해결되었습니다."

Joe Hoover, Mantium R&D Senior Applied Scientist

aws startups

고객 성공 사례

SageMaker로 복원력, 성능, 비용 절약을 달성한 Stability AI

FM은 언어, 이미지, 오디오, 비디오 등 분야마다 다양한 다운스트림 작업에 맞춰 변용할 수 있는 대규모 모델입니다. FM이 훈련하기 힘든 대상인 이유는 GPU나 Trainium 칩 수천 개를 포함한 고성능 컴퓨팅 클러스터가 있어야 하며, 그 클러스터를 효율적으로 관리할 소프트웨어 또한 필요하기 때문입니다. 커뮤니티 기반, 오픈 소스 AI 기업으로서 획기적인 기술 개발에 주력하는 **Stability AI**에서는 AWS를 클라우드 제공업체로 선택해 역대 가장 큰 규모의 GPU 클러스터를 퍼블릭 클라우드에 프로비저닝하기로 했습니다. SageMaker 관리형 인프라와 최적화 라이브러리를 사용하자, Stability AI의 모델 훈련은 전보다 복원력, 성능과 비용 효율성이 좋아졌습니다. 게다가 훈련 시간과 비용을 반 이상 줄일 수 있었습니다.

사례 읽기 ›

aws startups

stability.ai

"AWS는 전체 형식에 걸쳐 오픈 소스 기반 모델 규모를 조정하는 데 핵심적인 파트너가 되어 주었습니다. 이 모델을 SageMaker에 도입해 수만 명의 개발자와 수백만 명의 사용자가 이를 유익하게 활용할 수 있도록 지원하게 되어 기쁩니다."

Emad Mostaque, Stability AI Founder & CEO

고객 성공 사례

Runway, AWS와 함께 사내 연구 인프라 규모 조정

Runway는 AWS와 협력해 고성능 컴퓨팅(HPC) 클러스터의 규모를 조정하고 연구 인프라를 활용해 생성형 제품군 전반에 걸쳐 동급 최고의 사용자 경험을 제공했습니다. Runway의 Gen-2 시스템은 AWS에서 훈련해 텍스트, 이미지나 비디오 클립을 사용해 새로운 비디오를 만들 수 있습니다. Gen-2는 Runway의 다중 모달 생성 모델을 기반으로 개선점을 적용한 버전으로, 비디오 생성용 최첨단 AI 시스템이 얼마나 큰 발전을 이루었는지 보여줍니다.

사례 읽기 ›

runway

"이 획기적인 비디오 생성 모델을 개발하고 훈련하는 데 AWS가 중대한 역할을 했습니다. 앞으로도 생성형 AI로 무엇이 가능할지 길을 개척하는 과정을 계속 함께할 수 있었으면 좋겠습니다."

Cristóbal Valenzuela, Runway Co-Founder & CEO

다음 단계

생성형 AI 시작하기

이제 생성형 AI, 생성형 AI가 할 수 있는 일, 잠재적인 비즈니스 이점을 깊이 있게 이해했으니, 다음 단계로 목표를 명확히 정의하고 생성형 AI를 활용하는 사용 사례를 파악해야 합니다. 소규모로 실험하고, 단순하되 정확한 목표를 정해 시작하는 것이 가장 좋습니다. 빠르게 결과를 달성했다면 상위 직무와 외부로 활동 규모를 확장하면 됩니다.

전문가와 협업하여 데이터 가용성, 데이터 품질, 생성형 AI와 관련된 윤리적 영향과 같은 요소를 고려하기를 적극 권장합니다. 나아가 비용, 확장성, 에너지 소비에 상당한 영향을 미칠 수 있는 인프라도 사전에 고려해야 합니다. AWS 전문가와 협력하면 의사 결정 프로세스와 구현 단계 전반에 걸쳐 귀중한 지침을 얻을 수 있습니다.

지금 시작하세요!

생성형 AI는 역사상 가장 획기적인 기술의 하나가 될 것입니다. 인간의 창의력을 향상하고 혁신의 한계를 확장하며 결과를 극대화할 수 있기 때문입니다. AWS는 그 최전방에 서서, 공정하고 정확한 AI 서비스 개발을 위해 노력하고, 스타트업에 AI 애플리케이션을 책임 있게 구축하는 데 필요한 도구와 지침을 제공하고자 합니다. 이제 여러분의 스타트업도 시작할 때입니다.

스타트업을 위한 AWS 기반 생성형 AI에 대해 자세히 알아보기 ›