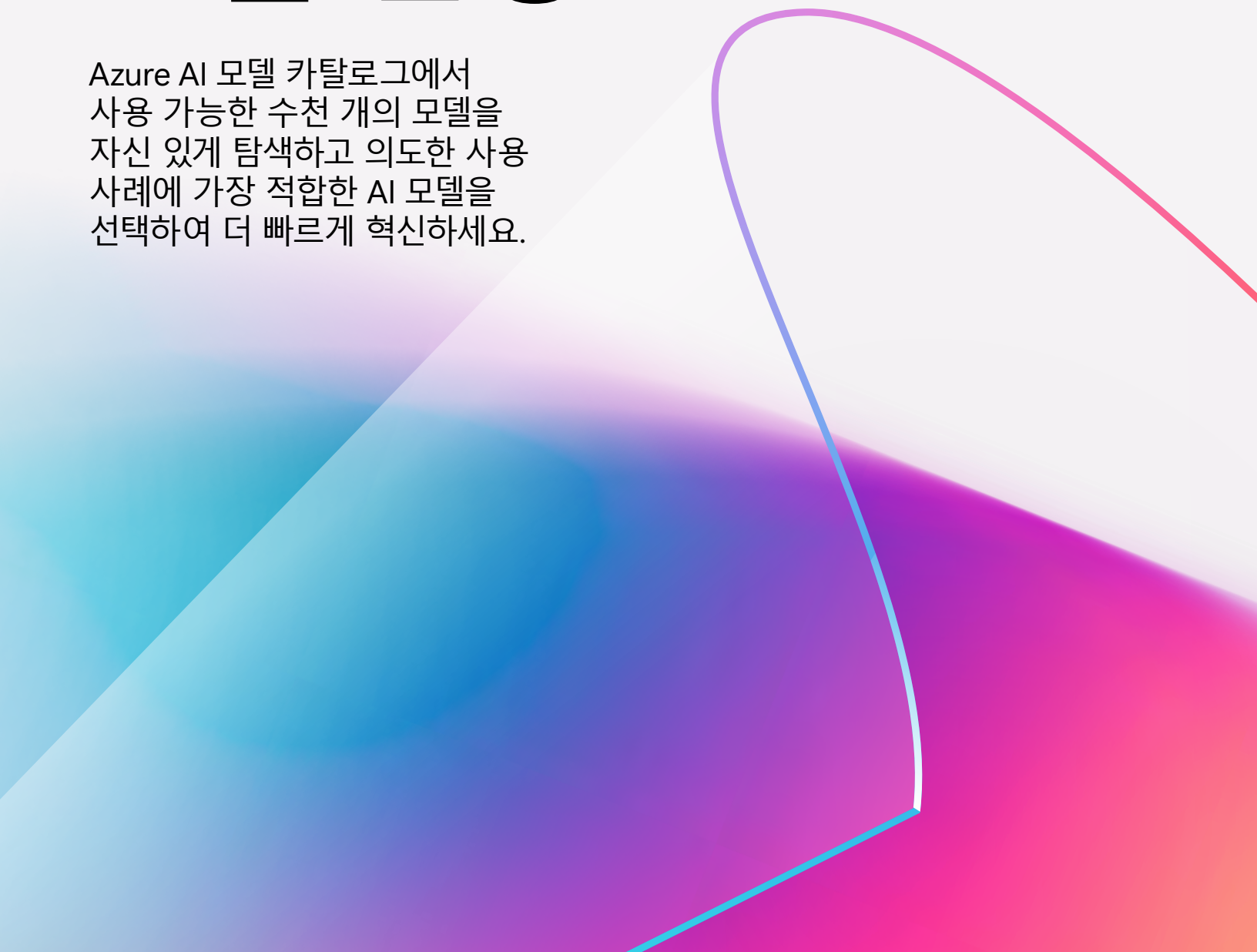
Microsoft Azure

적합한 AI 모델 선정

Azure AI 모델 카탈로그에서 사용 가능한 수천 개의 모델을 자신 있게 탐색하고 의도한 사용 사례에 가장 적합한 AI 모델을 선택하여 더 빠르게 혁신하세요.



목차

새로운 AI 모델이 출시되고 기존 모델은 계속 개선되고 있습니다

모델 혁신에는 부족함이 없습니다

지난 10년 동안 AI는 대규모 기반 모델의 등장에 힘입어 제조, 소매유통업, 재무, 의료 서비스 등의 분야에 크게 진출했습니다. 이러한 모델은 광범위한 데이터 세트에 대해 세심하게 훈련되었으며 자연어 처리, 컴퓨터 비전, 콘텐츠 제작을 위한 생성형 AI 등 다양한 작업에 적용할 수 있을 만큼 다재다능합니다. 그 결과, 인간 지능의 전유물로 여겨졌던 복잡한 작업도 이제 기계가 처리할 수 있게 되었습니다. 이러한 발전은 번역 앱과 크리에이티브 콘텐츠 도구부터 가상 채팅 도우미에 이르기까지 다양한 실용적인 애플리케이션을 지원하여 오늘날 우리가 기술과 상호 작용하는 방식을 크게 향상시킵니다.

이제 선택의 문제입니다

모델 공급자가 AI 모델 포트폴리오를 계속 출시하고 발전시키면서 기술 의사 결정권자는 새로운 과제에 직면하고 있습니다: 모델 혁신의 선두에 서서 성과 목표와 예산 제약을 충족하면서 원하는 사용 사례를 구현하는 데 적합한 모델을 자신 있게 비교, 선택, 최적화할 수 있습니다. 현실은 하나의 모델이 모든 모델에 적합하지 않으며, 오늘날에는 수천 가지의 모델이 존재합니다. 끊임없이 확장되는 가능성의 세계를 탐색하는 것은 압도적일 수 있지만, 특정 애플리케이션에 어떤 AI 모델을 사용할지 정보에 입각한 선택을 하기 위한 필수 단계입니다.

한 곳에서 모델을 평가, 테스트 및 모니터링하여 정보에 입각한 선택을 할 수 있습니다

AI 혁신가들은 개발 프로세스 전반에 걸쳐 AI 모델을 비교, 선택, 최적화하는 더 나은 접근 방식을 요구하고 있습니다. 이러한 이유로 Microsoft는 최신 오픈 소스 및 기초 모델을 한곳에 모아 보다 유연한 모델 선택이 가능하도록 Azure AI 모델 카탈로그를 만들었습니다. 개발자는 1,600개 이상의 모델을 살펴볼 수 있으며, 여기에는 Meta의 Llama 3와 같은 최신 혁신 기술뿐만 아니라 OpenAI, Mistral, NVIDIA, Cohere와 같은 AI 개발 분야의 리더와의 전략적 파트너십을 통해 탄생한 모델도 포함되어 있습니다.

기업은 단일 사용자 경험(UX)에서 오늘날 시장을 선도하는 주요 모델 제품군을 모두 실험해볼 수 있습니다. Microsoft의 주요 목표는 독점, 오픈 소스, 자사, 타사 등 Azure AI에서 최고의 모델을 제공하는 것입니다. 하지만 모델 선택 과정도 더욱 원활하고 접근하기 쉽게 만들고자 합니다. 멀티모달, 짧은 지연 시간, 비용 효율성 등 가장 중요한 사항에 따라 Azure AI 모델 카탈로그에서 옵션을 보다 확실하게 좁히고 사용 사례에 가장 적합한 것을 선택할 수 있는 배포 방법이 있습니다. 이 eBook에서는 Azure AI 모델 카탈로그의 상위 모델 공급자로부터 인사이트를 수집하여 사용자 지정 AI 솔루션에 적합한 모델을 선택하기 위한 권장 사항과 모범 사례를 제공합니다. 또한 실제 조직이 AI 모델을 선택하게 된 주요 결정 요인이 무엇인지 조사하고 있습니다.

다음 질문에 대한 모델 제공자의 답변을 들어보세요

모델 선택이 중요한 이유

올바른 모델을 선택하는 것은 비용과 성능에만 영향을 미치는 것이 아니라 장기적인 성공의 토대를 마련하는 것입니다.

모델 선택에 가장 큰 영향을 미치는 요소는 무엇인가요?

선도적인 모델 제공업체는 모델 선택 프로세스에 가장 큰 영향을 미치는 요소의 포괄적인 목록을 작성합니다.

개발 수명 주기 중 모델 선택은 언제 평가해야 하나요?

조직이 개발 라이프사이클의 각 단계에서 모델을 비교, 선택 및 최적화하기 위해 고려할 수 있는 상위 세 가지 질문을 알아보세요.

실제 조직이 AI 모델을 선택하게 된 주요 결정 요인을 알아보세요

Mistral Large를 선택해야 하는 이유

유럽의 한 선도적인 핀테크 조직은 고객 지원을 간소화하기 위해 Mistral Large를 뱅킹 플랫폼에 통합했습니다.

Meta Llama 컬렉션을 선택해야 하는 이유

Persado는 Llama 컬렉션으로 혁신을 이루어 ML, NLP, 딥러닝 모델을 통해 금융 서비스 마케팅을 혁신합니다.

Command R+ 및 Rerank를 선택해야 하는 이유

Atomicwork는 원활한 셀프 서비스 솔루션으로 생산성과 IT 지원을 혁신하는 AI 디지털 어시스턴트를 구축합니다.

GPT-4V를 선택해야 하는 이유

EY는 다양한 전문 작업을 지원하기 위해 Azure AI Foundry 및 Azure OpenAI Service를 활용하여 생산성과 혁신을 향상합니다.

모델 선택이 중요한 이유

빠르게 진화하는 AI 환경에서 어떤 모델을 선택하느냐에 따라 AI 프로젝트와 제품의 성패가 갈릴 수 있습니다. 대규모 독점 모델은 광범위한 기능과 고성능을 갖춘 프로토타입에는 매우 효과적일 수 있지만, 고객이 원하는 만큼 신속하게 응답하지 못할 수도 있습니다. 또는 높은 컴퓨팅 리소스가 필요한 복잡한 작업을 도입하는 순간 컴퓨팅 비용이 급증하는 소규모 스타트업의 예산에는 적합하지 않을 수도 있습니다. 결론은 모든 상황에 적합한 단일 모델은 없다는 것입니다. 그렇기 때문에 Meta, Cohere, Mistral, Microsoft와 같은 선도적인 모델 제공업체들은 AI의 잠재력을 최대한 활용하기 위해 올바른 모델을 선택하는 것이 매우 중요하다고 강조합니다.

모델 선택이 중요한 이유를 묻는 질문에 전문가들은 다음과 같이 비즈니스에 가장 강력한 영향을 미친다고 생각하는 것으로 응답했습니다

성능과 정확성: 강점 활용

AI 모델마다 고유한 아키텍처와 학습 방법이 있으며, 각 모델마다 뚜렷한 장단점이 있습니다. 예를 들어, 어떤 모델은 긴 텍스트를 처리하고 문맥에 맞는 응답을 생성하는 데 탁월한 반면, 어떤 모델은 사용자의 의도를 이해하거나 복잡한 추론 작업을 수행하는 데 능숙합니다. 조직은 적절한 모델을 선택함으로써 이러한 강점을 활용하여 특정 사용 사례에 맞게 조정함으로써 궁극적으로 성능과 정확성을 향상시킬 수 있습니다.

유연성과 적응성: 성공을 위한 커스터마이징

모델 선택은 AI 제품의 유연성과 적응성을 결정하는 데 중추적인 역할을 합니다. 일부 모델은 특정 작업이나 도메인에 맞게 미세 조정할 수 있어 목표 영역에서 사용자 지정 및 성능 향상이 가능합니다. 이러한 적응성을 통해 AI 솔루션이 대상 고객, 사용 사례 및 산업에 최적화되어 보다 효과적이고 효율적인 결과물을 도출할 수 있습니다.

확장성 및 성능: 효율성 균형 유지

모델 선택은 AI 제품의 확장성과 성능에 큰 영향을 미칩니다. 특정 모델은 효율성과 속도를 고려하여 설계되어 대용량 실시간 애플리케이션에 이상적입니다. 다른 제품은 더 정교한 기능을 제공하지만 더 많은 연산 능력이 필요하고 응답 시간이 느릴 수 있습니다. AI 제품이 사용자 경험을 저하시키지 않으면서 효과적으로 확장하려면 정확성과 성능 간의 적절한 균형을 유지하는 것이 필수적입니다.

비용 및 리소스 관리: 투자 최적화

올바른 모델을 선택하는 것은 비용 및 리소스 관리에 즉각적인 영향을 미칩니다. 모델마다 계산 요구 사항이 다르므로 전체 배포 및 유지 관리 비용에 영향을 미칠 수 있습니다. 프로젝트의 예산 및 리소스 제약 조건에 맞는 모델을 선택함으로써 조직은 AI 투자를 최적화하고 비용 효율적인 솔루션을 달성할 수 있습니다.

미래 대비와 혁신: 앞서가기

궁극적으로 올바른 모델을 선택하면 장기적인 성공을 거둘 수 있습니다. 강력한 연구와 탄탄한 기반이 뒷받침되는 모델을 선택하면 새로운 발전이 등장할 때마다 AI 제품을 지속적으로 개선하고 적용할 수 있습니다. 이러한 접근 방식은 제품 수명 주기를 연장할 뿐만 아니라 역동적인 시장에서 조직이 경쟁력을 유지하는 데 도움이 됩니다.

모델 선택에 가장 큰 영향을 미치는 요소는 무엇인가요?

비즈니스에 미치는 영향을 고려한 모델 선택의 중요성을 확인했습니다. 그렇다면 이러한 선택을 이끄는 요인은 무엇일까요? 다른 모델 제공업체와 협력하여 조직이 옵션 범위를 좁히고 사용 사례에 적합한 모델을 찾기 위해 참조할 수 있는 포괄적인 분류 체계를 작성했습니다. 작업의 특성부터 계산 리소스, 해석 가능성 등에 이르기까지 각 요소는 AI 애플리케이션의 성공을 결정하는 데 중추적인 역할을 합니다.

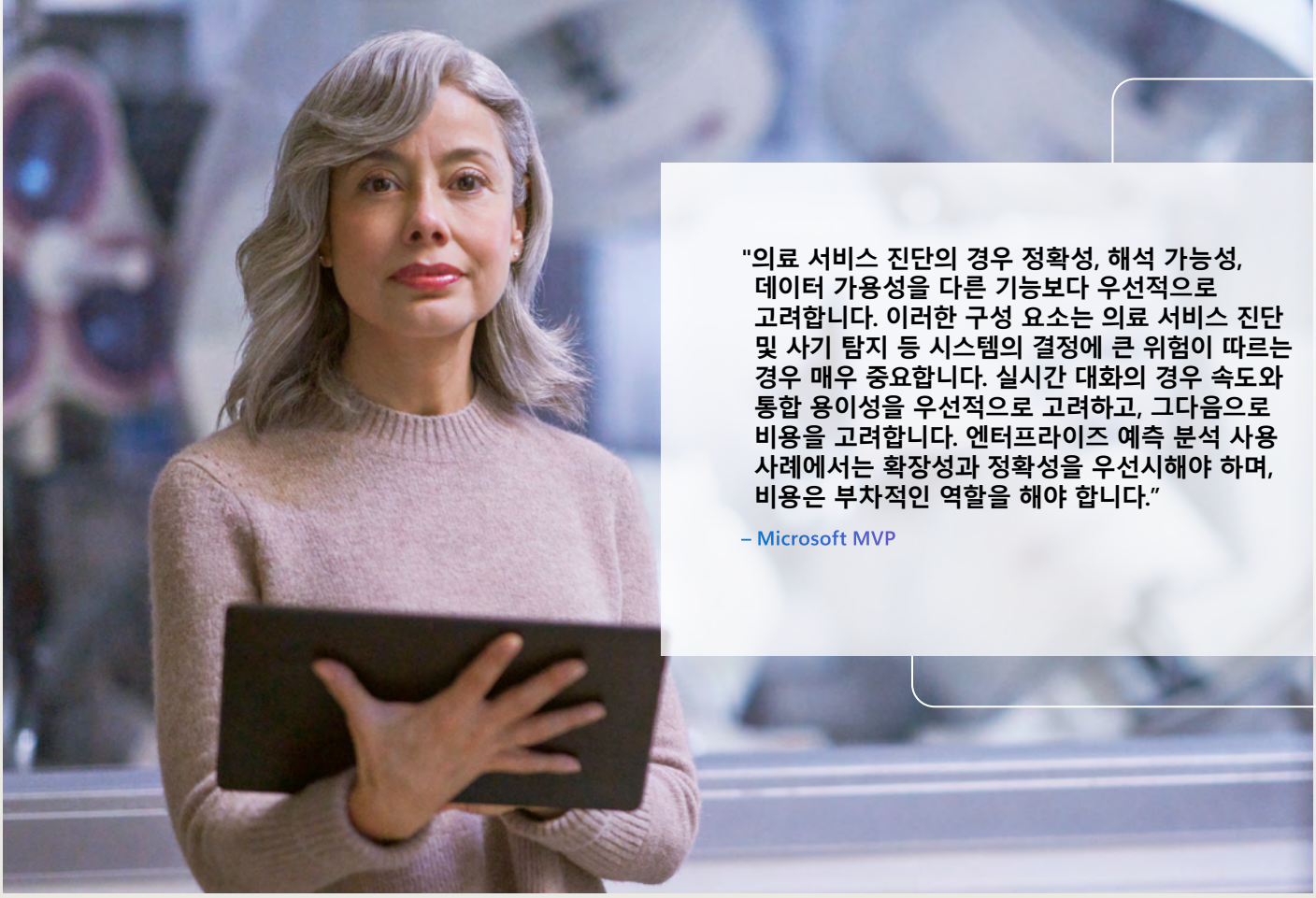
다음 분류법을 사용하여 옵션의 범위를 좁히고 사용 사례에 적합한 모델을 찾으세요

작업의 성격: 모델이 어떤 업무를 수행해야 하나요? 각기 다른 AI 모델은 각기 다른 작업에 탁월합니다. 예를 들어 컨벌루션 신경망(CNN)은 이미지 인식에 적합하고, 순환 신경망(RNN)은 자연어 처리(NLP)와 같은 순차적 데이터 처리에 적합합니다. 최적의 성능을 위해서는 작업의 특정 요구 사항을 이해하는 것이 필수적입니다. 이를 통해 NLP, 오디오, 컴퓨터 비전, 멀티모달 등 필요한 모델 유형에 집중할 수 있습니다.

컴퓨팅 리소스: 일부 모델은 실행에 상당한 연산 능력과 메모리가 필요합니다. 트랜스포머와 같은 대규모 모델에는 GPU나 TPU와 같은 특수 하드웨어가 필요한 경우가 많습니다. 모델 운영을 지원하는 데 필요한 물리적 및 가상 리소스를 이해하면 장기적으로 지갑을 보호하는 데 도움이 될 수 있습니다. 그렇기 때문에 사용 가능한 리소스를 평가하고 인프라 기능에 맞는 모델을 선택하는 것이 중요합니다.

"특히 프로덕션 환경이나 클라우드 플랫폼에서는 배포 규모에 따라 비용이 제한적인 요소가 될 수 있습니다. 일반적으로 스타트업과 비영리 단체는 비용 효율적이고 적절한 성능을 제공하는 클라우드 제공업체에서 제공하는 사전 학습된 모델을 선호합니다. 대기업의 경우 자체 데이터를 사용하여 자체 모델을 교육하는 등의 추가 옵션을 위한 충분한 예산이 있을 수 있습니다. 이를 위해서는 자체 인프라를 구축하거나 클라우드 컴퓨팅 서비스를 사용하고 가상 머신을 프로비저닝해야 합니다."

– Microsoft MVP



"의료 서비스 진단의 경우 정확성, 해석 가능성, 데이터 가용성을 다른 기능보다 우선적으로 고려합니다. 이러한 구성 요소는 의료 서비스 진단 및 사기 탐지 등 시스템의 결정에 큰 위험이 따르는 경우 매우 중요합니다. 실시간 대화의 경우 속도와 통합 용이성을 우선적으로 고려하고, 그다음으로 비용을 고려합니다. 엔터프라이즈 예측 분석 사용 사례에서는 확장성과 정확성을 우선시해야 하며, 비용은 부차적인 역할을 해야 합니다."

– Microsoft MVP

모델 배포: 필요한 컴퓨팅 리소스에 따라 고려 중인 모델의 배포 옵션, 액세스 방법, 확장성 및 지연 시간을 평가할 수 있습니다. 모델을 사용자 디바이스, 온-프레미스 서버 또는 클라우드에서 로컬로 실행할 수 있는지 고려하세요. SaaS API, 컨테이너화된 배포 또는 엣지 컴퓨팅을 통해 모델에 액세스하는 옵션이 있을 수 있습니다. 사용자 지정 배포에는 강력한 특수 하드웨어가 필요할 수도 있습니다.

모델 개방성: AI 시스템이 효과적일 뿐만 아니라 신뢰할 수 있고 윤리적이며 적응력을 갖추는 것이 얼마나 중요한가요? 모델 개방성이란 사용자, 개발자, 이해관계자가 AI 모델의 내부 작업, 데이터, 프로세스에 액세스하고 이해할 수 있는 정도를 말합니다. 완전한 투명성, 해석 가능성, 접근성 및 재현성을 원한다면 오픈소스가 가장 좋은 방법일 수 있습니다. 반면 오픈 계량 모델은 모델 무게만 제공하지만 더 비용 효율적이고 편리한 옵션이 될 수 있습니다. 독점 모델은 가장 낮은 수준의 개방성을 나타내며, 사용자가 코드, 데이터 및 모델 가중치에 액세스할 수 없습니다. 그러나 우수한 성능, 안정성, 전문 지원 서비스 등이 개방성보다 더 중요한 요소라면 독점 옵션이 더 실용적인 선택이 될 수 있습니다.

투명성 및 해석 가능성: 모델 개방성의 일부는 투명성과 해석 가능성입니다. 이러한 측면은 오픈 소스 시스템과 같은 가장 개방적인 모델도 의사 결정을 내리는 방식을 이해해야 하는 경우 필수적인 요소입니다. 또한 의료 및 금융과 같이 규제가 엄격한 산업에서는 특히 중요한 투명성 및 해석 가능성 수준이 다를 수 있습니다. 예를 들어, 의료 서비스에서는 의사 결정 트리 및 선형 회귀와 같은 투명한 모델이 내재된 해석 가능성으로 인해 선호되며, AI가 어떻게 예측을 내리고 데이터 소스, 데이터 품질 및 활용된 알고리즘에 대한 가시성을 높여줍니다.

성능 요구 사항: 모델이 얼마나 정확하고 빨라야 하나요? 의료 진단이나 금융 거래와 같이 위험도가 높은 애플리케이션에는 매우 높은 정확도와 짧은 지연 시간이 요구될 수 있습니다. 반면에 다른 애플리케이션에는 더 관대한 요구 사항이 있을 수 있습니다. 전반적인 성능 요구 사항을 이해하면 모델 예측의 정확성, 지연 시간, 시간에 따른 모델 성능의 일관성 및 견고성을 고려하여 어떤 종류의 모델이 더 적합한지 알 수 있는 단서를 얻을 수 있습니다.

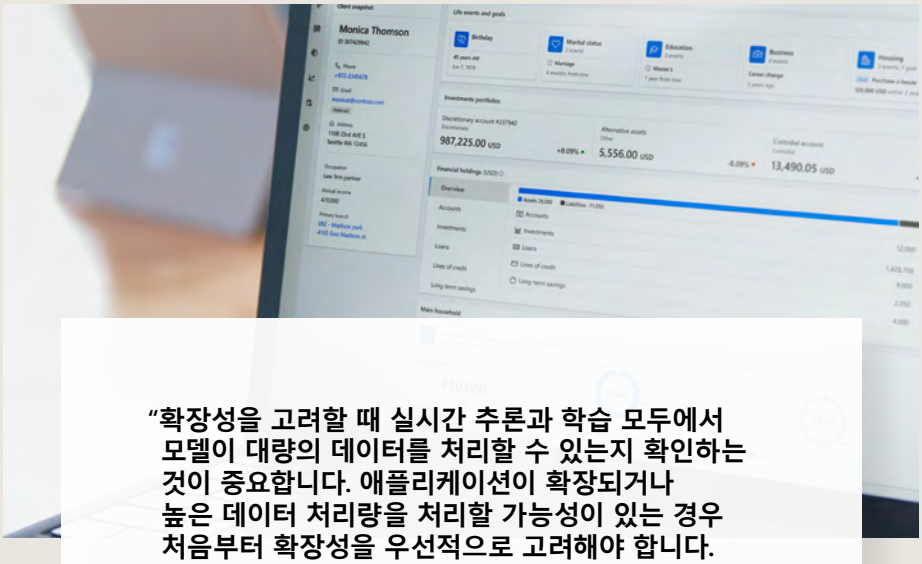
"자율 주행 및 대화형 AI와 같은 애플리케이션에서는 속도가 매우 중요합니다. 온라인 서비스에서 사용자 상호 작용과 같은 고속 데이터 처리를 처리할 때는 정확도와 함께 지연 시간과 추론 시간을 우선시하는 것이 중요하며, 특히 대규모 데이터 세트로 작업하거나 규모를 확장할 때는 더욱 그렇습니다."

– Microsoft MVP

모델 복잡성 및 정밀도: 규제가 엄격한 산업에서 사업을 운영하는 경우, 규정 준수, 예측의 정확성, 효과적인 위험 관리를 보장하기 위해 모델 복잡성과 정밀도가 중요한 요소일 가능성이 높습니다. 모델은 수십억에서 수조까지 매개 변수의 수가 다양합니다. 매개 변수가 높을수록 정밀도와 전반적인 성능이 향상됩니다. 그러나 이는 전산 비용 증가로 이어질 수 있습니다. 고정밀 모델은 오탐을 탐지하고 오류를 최소화하는 데 탁월합니다. 그러나 특히 훈련 중에 처리 능력이 향상되고 메모리 사용량이 증가하면 비용이 증가하는 경우가 많습니다. 정량화와 같은 기술은 성능과 리소스 요구 사항의 균형을 맞추는 데 도움이 될 수 있습니다.

보안 및 개인 정보 보호: 보안 프로토콜은 얼마나 엄격하나요? 보안은 대부분의 조직에서 최우선적으로 고려해야 할 사항이며, 일부 모델은 다른 모델보다 더 강력한 데이터 보호 기능을 제공합니다. 조직은 GDPR 및 HIPAA와 같은 규정을 준수하는 모델을 찾고, 보안 액세스 제어 및 악의적인 입력으로 모델을 속일 수 있는 공격에 대한 보호 등 배포 프로세스의 보안도 평가해야 합니다.

사용자 맞춤화 및 확장성: 필요에 따라 모델을 확장할 수 있나요? 사용자 맞춤 및 확장이 가능한 모델을 선택하면 장기적인 실행 가능성과 성공을 보장할 수 있습니다. 사용자 맞춤화로 특정 요구 사항에 맞게 모델을 조정하고 기존 시스템에 원활하게 통합할 수 있으며, 시간이 지남에 따라 요구 사항이 변화함에 따라 조정할 수 있습니다. 확장성은 증가하는 데이터와 사용자 요구를 효율적으로 처리할 수 있는 모델을 보장하여 향후 성장을 위한 유연성을 제공합니다. 조직은 모델의 성능뿐만 아니라 적응성과 장기적인 가치에 대해서도 평가해야 합니다.



"확장성을 고려할 때 실시간 추론과 학습 모두에서 모델이 대량의 데이터를 처리할 수 있는지 확인하는 것이 중요합니다. 애플리케이션이 확장되거나 높은 데이터 처리량을 처리할 가능성이 있는 경우 처음부터 확장성을 우선적으로 고려해야 합니다. 이를 평가하려면 모델의 문서화, 계산 복잡성, 대규모 데이터 세트에서의 성능을 검토하세요."

– Microsoft MVP

개발 수명 주기 중 모델 선택은 언제 평가해야 하나요?

올바른 모델 선택은 사용 사례를 해결할 뿐만 아니라 전체 워크플로를 혁신합니다.

개발자는 Azure AI Foundry의 Azure AI 모델 카탈로그를 활용하여 지능형 AI 애플리케이션을 원활하게 프로토타이핑, 최적화 및 운영할 수 있습니다. 그렇다면 이 개발 수명 주기에서 모델 선택은 언제, 어디서 고려될까요? 프로토타입 제작 단계에서 개발자에게 처음부터 최적의 모델을 선택하도록 과도한 요구를 하는 대신, 모델 선택에 대한 보다 반복적이고 적응력 있는 접근 방식을 옹호합니다. 대신 팀은 간단하지만 강력한 일련의 질문을 스스로에게 던짐으로써 각 개발 단계를 통해 프로젝트를 진행하거나 되돌릴 수 있습니다: AI가 제 사용 사례를 해결할 수 있나요? 내 사용 사례에 가장 적합한 모델은 무엇인가요? 실제 워크로드에 맞게 확장할 수 있나요? 이 세 가지 질문은 개발자가 맞춤형 AI 솔루션을 제공하는 데 필요한 모든 항목을 충족하는 올바른 모델을 찾는 데 도움이 되는 전환점이 됩니다.

모델 선택 및 개선 단계

시작하기

- 작업 우선순위
- 플래그십 모델로 테스트

심층 분석

- 추론 속도
- 예산
- 정밀도 수준
- 배포 옵션
- 성과
- 리소스 제약

확장

- 스트레스 테스트
- 규정 준수 및 보안
- 안정성 및 가동 시간
- 적응성
- 사용자 환경
- 비용 효율성

프로토타입

AI 모델을 탐색하고 테스트하여 실현 가능성 판단하기

질문 1

AI가 제 사용 사례를 해결할 수 있나요?

프로토타입 제작을 막 시작했을 때 가장 먼저 대답해야 할 중요한 질문입니다. 하지만 이 질문에 답하려면 사용 사례를 정의해야 합니다. 어떤 문제를 해결하려고 하나요? AI 솔루션이 어떤 작업을 수행하기를 원하시나요? 주요 사용 사례를 고려하고, 그 발견을 프로토타입을 구축할 주력 모델을 찾는 기준으로 삼으세요. 이 플래그십 모델을 고수하지 않을 수도 있지만, 실현 가능성을 평가하기 위한 좋은 출발점이 될 수 있습니다.

주요 사용 사례에 따라 프로토타입의 프론티어 또는 플래그십 모델을 선택하는 것으로 시작하세요

고객 리뷰를 분석하여 제품에 대한 전반적인 감정을 파악하고 싶으신가요? GPT-4와 같은 AI 모델은 언어 번역, 감정 분석, 텍스트 요약에 능숙합니다. 아니면 이미지 및 동영상 분석을 지원하는 모델을 찾고 계신가요? 아마도 DALL-E가 가장 적합할 것입니다. 사용 사례를 정의한 다음, 프로토타입에 사용할 최고의 인텔리전스를 갖춘 가장 진보된 모델을 선택하세요.

텍스트: GPT-4, Copilot, Llama3 또는 Mistral을 시작점으로 고려하세요.

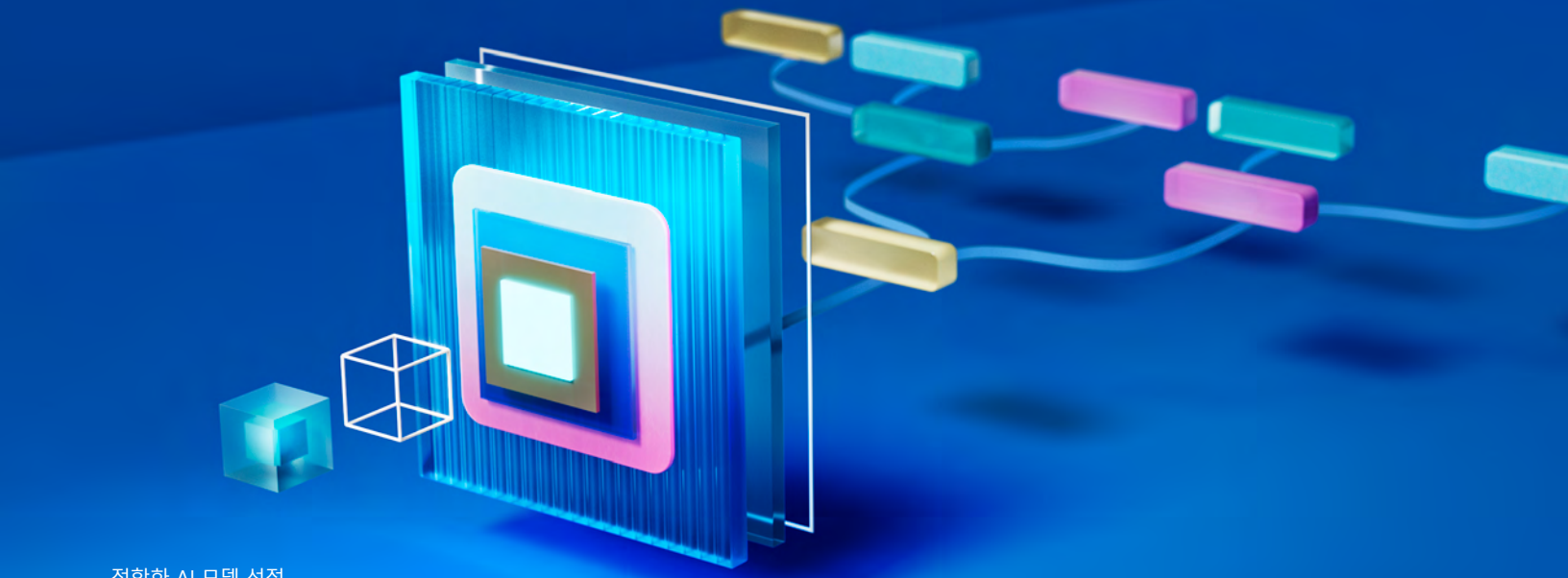
오디오: 텍스트 음성 변환 기능의 경우 OpenAI Whisper를 고려해 보세요.

이미지: OpenAI DALL-E, Phi-3 Vision 또는 Stability AI Stable Diffusion을 살펴보세요.

멀티모달: OpenAI GPT-4o로 시작하는 것이 좋습니다.

코드 생성: CodeLlama

임베딩: Cohere Embed



최적화

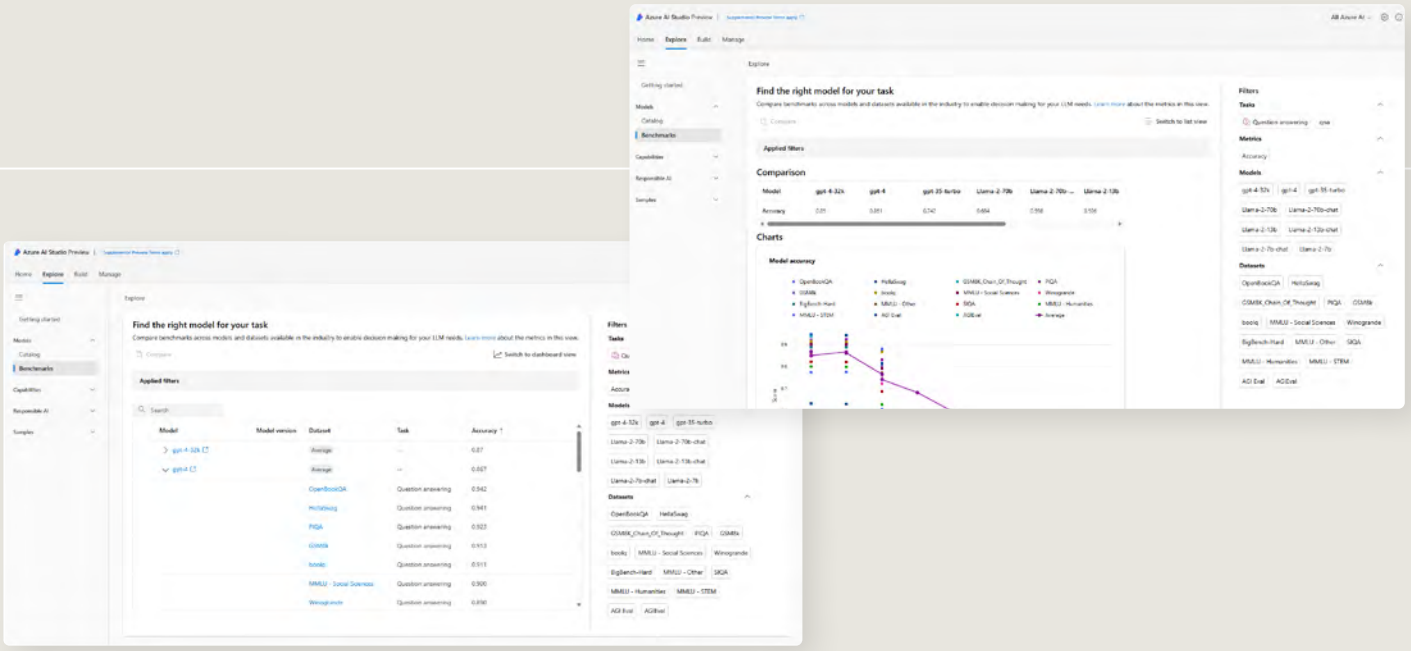
특정 작업 또는 도메인에 맞게 모델 채택

질문 2

내 사용 사례에 가장 적합한 모델은 무엇인가요?

프로토타입이 있고 실현 가능성이 입증되었습니다. 이제 다음 단계로 넘어갈 준비가 되었습니다. 최적화. 이 단계에서는 프로토타입을 제작할 준비를 해야 하므로 더 많은 질문과 고려 사항이 생길 수 있습니다. 예를 들어, AI 모델의 계산 비용은 얼마인가요? 지연 시간은 어떻게 되나요? 이것이 실제로 재정적으로 가장 좋은 모델인가요? 배송 시 고려해야 할 지역적 뉘앙스가 있나요?

개발자는 Azure AI Foundry의 벤치마킹 기능을 사용하여 요구 사항에 가장 적합한 모델을 식별하는 프로세스를 간소화할 수 있습니다. 모델 벤치마크는 사용자가 다양한 메트릭을 종합적으로 비교할 수 있는 기능을 제공하여, 작업을 시작하기 전에 모델과 데이터 세트의 지속 가능성에 대해 스스로 교육하고 정보에 입각한 결정을 내릴 수 있게 해줍니다. 정확성, 일관성, 유창성, GPT 유사성, 근거성, 관련성 등의 메트릭을 비교하여 각 모델의 강점과 약점을 빠르게 파악할 수 있습니다. 벤치마크는 기존 모델에 새로운 지표와 데이터 세트가 추가되고 카탈로그에 새로운 모델이 추가됨에 따라 정기적으로 업데이트됩니다.



오픈 소스 데이터 세트를 사용하여 작업별로 모델을 비교하고 모델 벤치마크를 활용하여 다양한 성능 메트릭을 평가한 후에는 초기 모델 선택을 재고해야 하는 경우가 종종 있습니다. 다행히 Azure AI Foundry는 통합 환경과 유연한 도구 덕분에 개발자가 처음부터 다시 시작하지 않고도 AI 모델을 전환할 수 있는 원활한 환경을 제공합니다. 데이터나 구성을 내보내고 가져올 필요 없이 동일한 워크플레이스 내에서 모델을 교체할 수 있습니다. 또한 유연한 배포 옵션을 통해 최소한의 변경만으로 동일한 엔드포인트 또는 애플리케이션에 새 모델을 배포할 수 있습니다. 이렇게 하면 새 모델을 테스트하고 검증하는 데 필요한 시간과 노력을 줄일 수 있습니다.

심층 분석 – 미세 조정
추론 속도 및 예산과 같은 요소를 평가합니다

가장 중요한 것에 집중하세요. 모델에 필요한 다른 작업을 결정하는 데 도움이 되는 보다 세분화된 질문을 하세요. Azure AI Foundry의 모델 벤치마크를 활용하여 이 심층 조사를 지원하므로 가장 중요한 성능 메트릭을 더 쉽게 평가하고 비교할 수 있습니다.

추론 속도 요구 사항 식별

추론 속도는 애플리케이션의 성능과 사용자 경험에 직접적인 영향을 미치기 때문에 중요한 요소입니다. 따라서 고려할 가치가 있습니다. 실시간 챗봇이나 대화형 시스템과 같이 초고속 응답이 필요한 애플리케이션에서는 Microsoft Phi3 또는 Llama 8b와 같은 소형 모델을 선택하는 것이 필수적입니다. 이러한 모델은 속도에 최적화되어 있으며 소비자 하드웨어에서 효율적으로 실행할 수 있어 빠른 상호 작용을 보장합니다.

사람의 읽기 속도와 같이 빠른 응답이 필요한 애플리케이션의 경우 GPT-4 Turbo 또는 Llama3 70b와 같은 모델이 속도와 성능 간의 균형을 잘 맞출 수 있습니다. 시기적절하면서도 정확한 출력이 필요한 작업에 적합합니다.

응답 시간이 덜 중요한 시나리오에서는 클라우드 배포에서 실행하는 가장 크고 강력한 모델을 사용할 수 있습니다. 이 접근 방식은 속도 요구 사항의 제약을 받지 않고 모델의 기능을 극대화합니다.

예산과 성과 사이의 적절한 균형 찾기

예산은 AI 모델이 고객의 요구에 적합한지 여부를 재평가하는 데 중요한 역할을 합니다. 예산이 없는 경우 오픈 소스 모델을 사용하는 것이 좋습니다. 유연성이 뛰어나며 재정적 간접 투자 없이 시작할 수 있습니다. 초기 단계에 있고 예산이 확실하지 않은 경우, 특히 최소 실행 가능 제품(MVP)을 테스트할 때는 오픈 소스 모델로 시작하는 것이 현명한 방법입니다. 이러한 모델을 사용하면 큰 비용 없이 MVP를 실험하고 검증할 수 있습니다. 프로젝트가 발전하고 요구 사항이 명확해지면 더 성능이 뛰어난 모델로 전환할 수 있습니다.

가성비가 좋은 제품을 찾는 분들에게는 OpenAI ChatGPT-3.5가 제격입니다. 빠르고 저렴하며 신뢰할 수 있는 결과를 제공하므로 다양한 애플리케이션에 탁월한 선택입니다.

반면에 돈이 문제가 되지 않는다면 가장 크고 강력한 모델을 선택할 수 있습니다. 호스팅된 하드웨어에서 실행하거나 최상위 API 서비스를 사용하면 최상의 성능과 기능을 얻을 수 있습니다.

필요한 정확도 및 전문 지식의 수준을 결정합니다

맞춤형 솔루션을 위한 AI 모델을 선택하고 평가할 때는 사용 사례에 필요한 정밀도와 도메인 지식을 고려하는 것이 중요합니다. 전문 분야에서 높은 정밀도가 요구되는 프로젝트의 경우, 필요한 특정 지식에 맞게 미세 조정할 수 있는 Llama 또는 Phi와 같은 모델이 이상적입니다.

반면에 애플리케이션에 광범위한 범용 지식에 걸쳐 고성능이 필요한 경우, 대형 독점 LLM이 최선의 선택일 수 있습니다. 이 모델은 다양한 분야를 처리하고 강력한 결과를 제공하는 데 탁월합니다.

극단적인 전문화가 필요하지 않은 시나리오의 경우 독점 모델과 오픈 소스 모델 중에서 유연하게 선택할 수 있습니다. 이러한 경우 검색 증강 생성(RAG)과 같은 기술을 통해 필요한 근거 지식을 얻을 수 있습니다.

운영화

시간 경과에 따른 모델 변경 사항 배포 및 관리

질문 3

실제 워크로드에 맞게 확장할 수 있나요?

최적화 단계가 끝날 무렵에는 사용 사례에 더 적합한 AI 모델을 찾았을 것입니다. Azure AI Foundry에는 모델 개발의 시간, 비용 및 복잡성을 줄이기 위해 자체 학습에서 모델을 미세 조정할 수 있는 기본 제공 최적화 및 도구도 제공됩니다. 모델을 충분히 테스트하고 필요에 맞게 조정했다면 이제 배포할 차례입니다.

하지만 먼저 워크로드가 증가함에 따라 모델을 안전하고 원활하게 확장하기 위해 고려해야 할 사항이 무엇인지 자문해 보아야 합니다. AI 여정에서 엔터프라이즈급 보안, 데이터 개인 정보 보호, LLM 공격 벡터에 대한 보호, 콘텐츠 안전 및 콘텐츠 필터링 측면과 같은 요소를 다른 구성 요소 중에서 멈추고 고려할 수 있는 완벽한 시점입니다.

선택한 AI 모델 확장 준비

다음 체크리스트를 실행하여 선택한 AI 모델을 대규모로 배포할 준비가 되었는지 확인해 보세요.

부하 시 성능: AI 모델이 실제 시나리오에서 접하게 될 데이터의 양과 다양성을 처리할 수 있도록 하는 것은 필수적입니다. 여기에는 워크로드가 증가해도 성능과 정확성을 유지하는 것도 포함됩니다. Azure AI Foundry를 사용하면 실제 워크로드를 시뮬레이션하고 모델을 스트레스 테스트하여 다양한 수준의 수요에서 모델이 성능을 발휘할 수 있는지 확인할 수 있습니다.

리소스 관리: 확장에는 컴퓨팅 리소스를 효율적으로 관리하는 것이 포함됩니다. 즉, 인프라가 과도한 비용이나 리소스 낭비 없이 증가하는 수요를 지원할 수 있도록 보장해야 합니다. Azure AI Foundry는 모델을 교육하고 배포하는 데 필요한 컴퓨팅 리소스를 모니터링하고 최적화할 수 있도록 인사이트를 제공합니다.

안정성 및 가동 시간: 실제 애플리케이션, 특히 규제가 엄격한 산업에서 신뢰성은 타협할 수 없는 요소입니다. 모델은 심각한 다운타임 없이 지속적으로 운영될 수 있어야 하며, 이를 위해서는 강력한 인프라와 장애 복구 메커니즘이 필요합니다. Azure AI Foundry는 Azure의 강력한 인프라와 통합되어 고가용성 및 재해 복구 옵션을 제공하여 불리한 조건에서도 모델이 안정적으로 작동하도록 보장합니다. 이러한 고려 사항은 적대적 공격 및 기타 악의적인 활동으로부터 보호하는 데 특히 중요합니다.

사용자 환경: 사용자 기반이 증가함에 따라 AI 모델은 긍정적인 사용자 경험을 유지하기 위해 일관되고 빠른 응답을 제공해야 합니다. 성능 저하 없이 많은 사용자에게 동시에 서비스를 제공할 수 있어야 합니다. Azure의 관리형 서비스를 통해 모델을 배포하면 사용자 기반이 증가함에 따라 지연 시간이 짧은 응답과 높은 처리량을 보장하여 긍정적인 사용자 환경을 유지할 수 있습니다.

적합한 AI 모델 선정

규정 준수 및 보안: 솔루션을 배송하기 전에 모델이 대규모의 규정 준수 및 보안 표준을 준수하는지 확인해야 합니다. 여기에는 사용자 데이터 보호, 무단 액세스 방지, 업계 규정 준수가 포함됩니다. Azure AI Foundry는 데이터 보호 및 규정 준수를 위한 업계 표준을 충족하는 데 사용할 수 있는 기본 제공 보안 기능 및 규정 준수 인증(예: GDPR 및 CCPA)을 제공합니다.

비용 효율성: 효율적인 확장 전략은 성능과 비용의 균형을 유지하여 리소스 사용을 최적화하고 운영 비용을 제어하는 동시에 성능 요구 사항을 충족하는 데 도움이 됩니다. Azure AI Foundry에는 이러한 비용을 추적하고 제어하는 데 사용할 수 있는 비용 관리 도구가 있어 성능 요구 사항을 충족하면서 예산을 더 잘 보호할 수 있습니다.

변화하는 요구 사항에 대한 적응성: 실제 워크로드는 예측하기 어려울 수 있습니다. 그렇기 때문에 확장 가능한 모델은 사용량이 갑자기 급증하거나 시간이 지남에 따라 더 복잡한 데이터를 처리해야 하는 등 변화하는 수요에 적응할 수 있습니다. Azure AI Foundry는 지속적인 통합 및 지속적인 배포 파이프라인을 지원하므로 요구 사항의 변화에 따라 모델을 원활하게 업데이트하고 개선할 수 있습니다.

실제 조직이 AI 모델을 선택한 이유

Azure OpenAI Service를 선택해야 하는 이유

생산성 및 혁신 향상을 위해 대규모 언어 모델의 힘을 활용하는 EY

Ernst & Young (EY)은 LLM의 힘을 활용하여 문서 비교, 코드 생성, 콘텐츠 제작 등 다양한 전문 업무의 생산성을 높여주는 생성형 AI 플랫폼인 EYQ를 구축했습니다. EY는 Azure AI Foundry 및 Azure OpenAI Service와 Azure AI 모델 카탈로그를 활용하여 GPT-4V/Turbo 및 Llama와 같은 다양한 모델을 지속적으로 평가하고 통합하여 특정 운영 요구 사항에 부합하는 동시에 책임 있는 AI 원칙을 준수하도록 보장할 수 있었습니다. 이러한 사전 예방적 접근 방식은 효율성을 개선했을 뿐만 아니라 글로벌 서비스 전반에서 혁신을 주도했습니다. 이후 275,000명 이상의 직원이 EYQ를 도입하여 다양한 업무를 보다 정확하고 효율적으로 수행할 수 있는 역량을 강화했습니다.

Mistral Large를 선택해야 하는 이유

고객 지원을 위한 뱅킹 어드바이저 에이전트를 구축하는 유럽의 선도적인 핀테크 조직

중소기업을 위한 서비스에 주력하는 한 온라인 뱅킹 및 소프트웨어 회사는 대기 시간을 줄여 고객 지원을 개선하고자 했습니다. 이들은 Gen AI를 사용하여 간단한 요청에 대한 응답을 자동화하기로 결정하고, Mistral Large를 지원 플랫폼에 통합하여 수신되는 L1 티켓 솔루션을 처리하기로 했습니다. 이를 통해 티켓당 비용을 80% 절감하고 은행 상담원의 조치 없이도 L1 티켓의 50%를 해결하여 직원의 시간을 확보할 수 있었습니다. 또한 Mistral 기반 플랫폼은 요청 컨텍스트에 따라 사기 사례를 분류하여 회사의 사기 탐지 시스템을 강화했습니다. Mistral은 다양한 EU 언어로 교육되고 대응 솔루션 팀의 지원을 받는 최첨단 모델 제품군으로 선정되었습니다.

Meta Llama를 선택해야 하는 이유

ML, NLP 및 딥러닝 모델로 금융 서비스 마케팅을 혁신하는 Persado

Persado의 동기 부여 AI는 Llama 컬렉션의 고급 ML, NLP 및 딥러닝 트랜스포머 모델을 활용하여 96% 더 효과적으로 행동을 유도하는 감정 정보 기반 메시지를 생성합니다. 이 조직은 딥러닝 알고리즘을 기반으로 자연어를 처리하고 생성할 수 있는 높은 신뢰도와 감사를 받은 오픈 소스 모델 모음으로 Meta를 선택했습니다. Persado는 브랜드에 맞는 효과적인 마케팅 커뮤니케이션을 위해 Llama 모델을 세밀하게 조정했습니다. 이 모델은 수십만 개의 다국어 마케팅 콘텐츠 생성 지침으로 구성된 독점 데이터 세트를 통해 더욱 정교해졌습니다.

Cohere Command R+ 및 Rerank를 선택해야 하는 이유

생산성 및 IT 지원을 혁신하기 위해 AI 디지털 어시스턴트를 구축하는 Atomicwork

IT 서비스 관리 회사인 Atomicwork는 기업 생산성 향상과 IT 지원을 목표로 하는 AI 디지털 어시스턴트인 Atom AI를 출시했습니다. 이 도구는 Cohere의 Command R+ 및 Rerank 모델을 RAG와 함께 활용하여 IT 지원 효율성을 크게 향상시킵니다. 평가 벤치마크 결과 Cohere Rerank를 사용하면 정확도가 20% 향상되었으며, 이 솔루션은 GPT-3.5와 같은 모델에 비해 75% 낮은 지연 시간과 168% 높은 정확도로 메트릭을 능가하는 성능을 보였습니다.

Azure AI 모델 카탈로그는 시작점입니다

Azure AI 모델 카탈로그에 포함된 주요 모델 공급자의 최신 AI 모델을 Azure AI Foundry에서 살펴보세요. Azure를 사용하면 최신 오픈 소스 및 파운데이션 모델이 한 지붕 아래에 있으므로 배포하기 전에 쉽게 검색, 비교 및 테스트할 수 있습니다. 다양한 작업과 양식을 아우르는 플래그십 LLM 또는 SLM, 독점 모델 또는 개방형 모델을 모두 살펴보세요. 또한 관리형 컴퓨팅 옵션과 서버리스 API를 모두 포함하는 보다 유연한 배포 옵션의 이점을 Azure OpenAI Service 및 서비스형 모델로 활용하세요.

지금 바로 다음 AI 솔루션 구축 시작

Azure AI 모델 카탈로그에 대해 자세히 알아보세요.

첫 번째

Microsoft Azure