생성형 AI 보안과 관련된 중요한 4가지 질문에 대한 답변

생성형 AI를 빠르게 도입하면서
개인정보 보호 및 규정 준수 보장

이 eBook은 비즈니스 리더, 특히 생성형 AI를 조직에 안전하게 통합하려고 계획 중이거나 그 방법을 고민하는 IT 의사 결정권자와 보안 팀 책임자를 위해 작성되었습니다.

목차

서문

이제는 생성의 시대: 생성형 AI를 빠르고 안전하게 도입

생성형 AI를 위한 경쟁이 시작되었습니다. 기업은 생산성과 경험을 크게 개선할 수 있다는 희망으로 고객 경험과 애플리케이션을 혁신하는 작업에 앞다투어 뛰어들고 있습니다.

생성형 인공 지능(AI)은 아직 초기 단계이지만 이미 조직의 거의 모든 사업부에 실질적인 이점을 제공하고 있습니다. 그러나 보안 전문가는 주의를 기울일 것을 권고합니다. 보안 전문가들은 데이터 프라이버시, 모델 편향성, 유해한 콘텐츠 생성(딥페이크 등), 모델에 악의적인 데이터가 입력될 수 있는 위험 등을 이유로 생성형 AI 도입에 신중할 것을 주문합니다.

조직은 신속한 도입을 실현하고 고객 경험을 개선하는 한편 데이터, 사용자 및 평판을 보호하는 방법에 대한 명확한 전략을 바탕으로 생성형 AI에 접근해야 합니다.

이렇게 이 과제는 다각적 측면이 있는 것도 사실이지만, 조직이 기억해야 할 사실은 AI, 기계 학습(ML), 데이터 보호 및 클라우드 워크로드 보안에 대한 표준 모범 사례가 여전히 유효하다는 것입니다. 실제로 조직은 생각보다 생성형 AI 보안을 위한 준비가 더 잘 되어 있을 수 있습니다.

이제 생성형 AI 워크로드를 보호할 수 있는 적절한 조치를 마련하면 조직 전반에 혁신을 가져와 팀이 중요한 아이디어를 추진하는 데 필요한 확신과 비즈니스 성장에 집중할 여유를 확보할 수 있습니다.

이 eBook에서는 보다 안전한 생성형 AI 워크로드를 향한 여정에 도움이 될 4가지 주요 질문을 살펴보겠습니다.

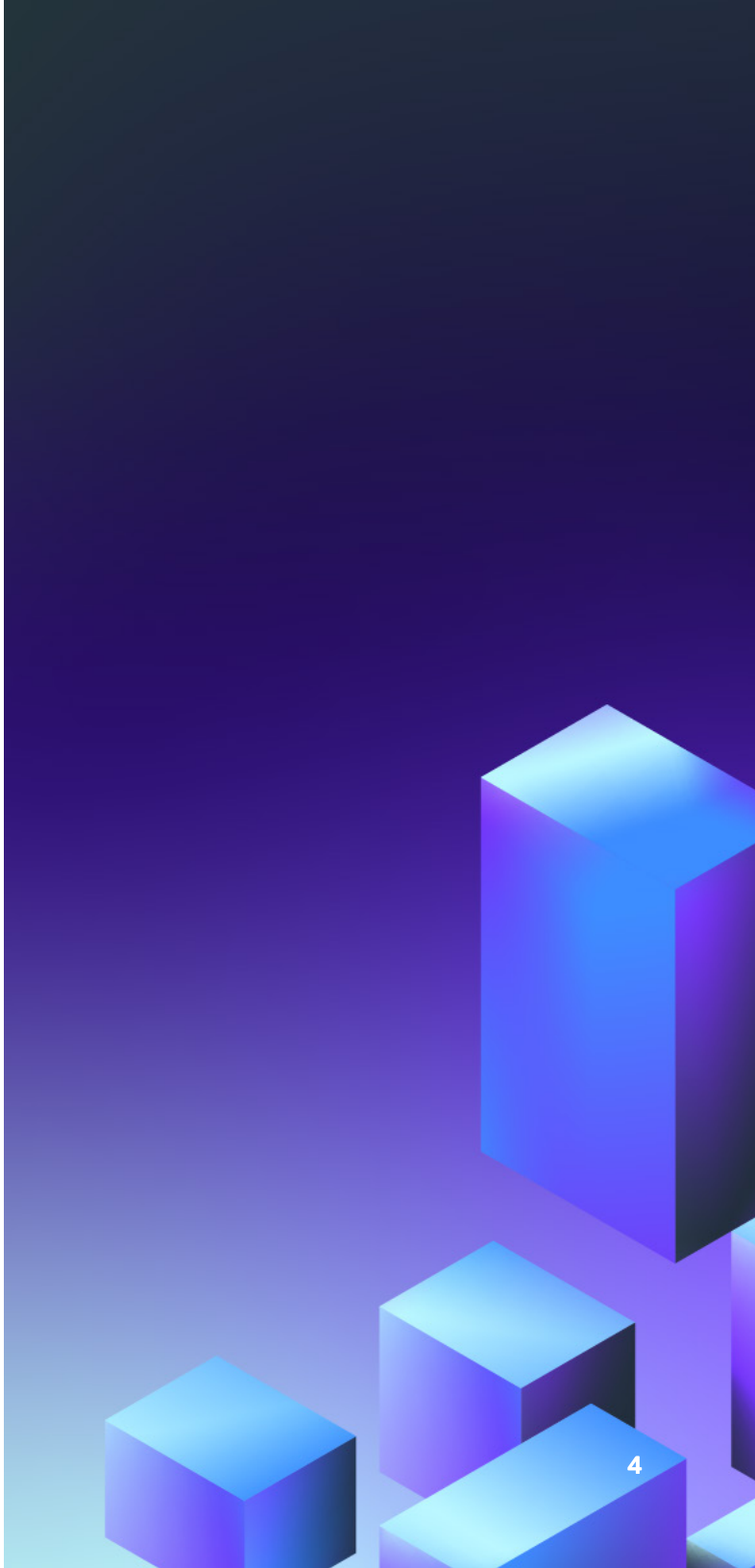
1. **무엇을 보호해야 할까요?**
2. **규정 준수 문제를 어떻게 해결할 수 있을까요?**
3. **어떻게 해야 모델이 의도한 대로 동작할 수 있을까요?**
4. **어디서부터 시작해야 할까요?**

질문 1:

무엇을 보호해야 할까요?

생성형 AI 애플리케이션을 안전하게 개발하고 배포하기 전에 보호가 필요한 대상이 정확히 무엇인지 이해하는 것이 중요합니다. 이는 다음과 같은 세 가지 범주로 나누어 생각해 볼 수 있습니다.

- **클라우드 워크로드 보호**
- **데이터 보호**
- **생성형 AI 애플리케이션 보호**



클라우드 워크로드 보호

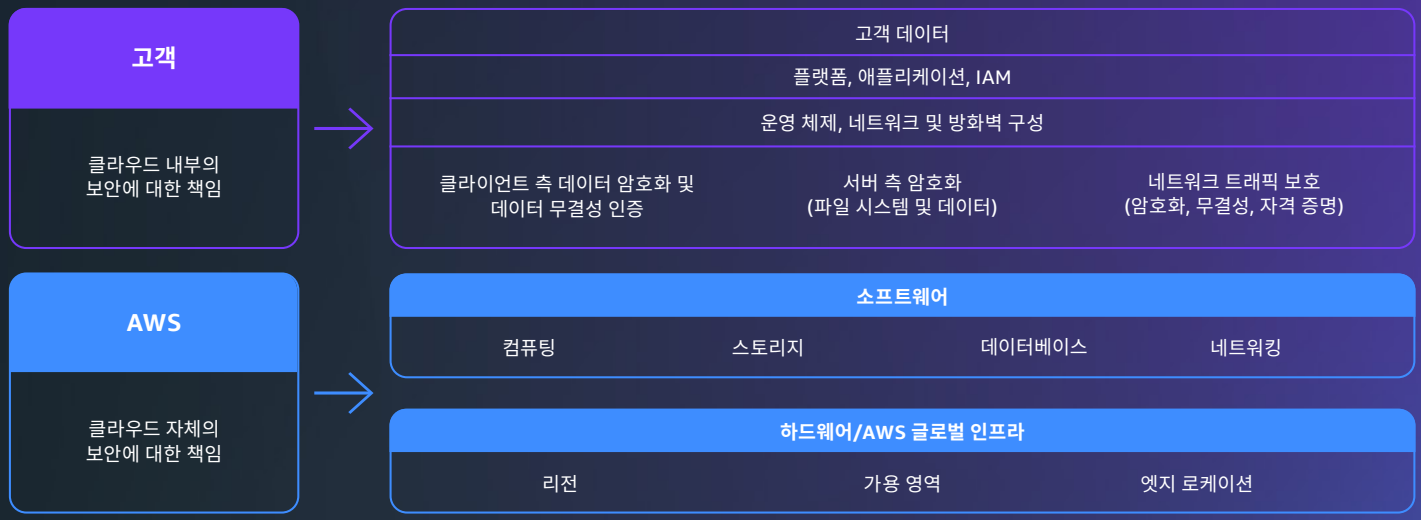
생성형 AI를 사용하면서도 보안 및 개인정보 보호 목표를 달성하려면 전체 클라우드 인프라, 서비스 및 구성을 보호하는 것부터 시작해야 합니다. 이를 위해 먼저 고객의 보안 책임과 클라우드 공급자의 보안 책임을 구별해야 합니다.

Amazon Web Services(AWS) 고객은 **공동 책임 모델**을 참조하여 보안 책임에 대한 지침을 얻을 수 있습니다. 지침에 따르면 AWS 클라우드에서 제공되는 모든 서비스를 실행하는 인프라를 운영, 관리 및 제어할 책임은 대체로 AWS에 있습니다. 이를 '클라우드 자체의 보안'이라고 합니다.

반면 AWS 고객은 게스트 운영 체제(업데이트 및 보안 패치 포함) 및 기타 관련 애플리케이션 소프트웨어, 그리고 AWS가 제공하는 보안 그룹 방화벽의 구성을 관리할 책임이 있습니다. 고객에게 요구되는 구체적인 책임과 그 범위는 고객이 사용하는 AWS 서비스에 따라 다릅니다. 이를 '클라우드 내부의 보안'이라고 합니다.

생성형 AI의 인기는 새로운 것이더라도, 여전히 기존 보안 모범 사례를 출발점으로 삼을 수 있습니다. 모범 사례 중에는 다음에 대한 기본적인 보안 수칙이 포함됩니다.

- 자격 증명 및 액세스 관리(IAM)
- 탐지 및 대응
- 인프라 보호
- 데이터 보호
- 애플리케이션 보안



데이터 보호

다음으로, 생성형 AI 애플리케이션에 사용되는 데이터의 보안과 개인정보 보호를 보장해야 합니다. 여기에는 독점 정보, 중요한 지적 재산권(IP) 및 개인 식별 정보(PII)가 포함될 수 있습니다.

생성형 AI 애플리케이션은 방대한 양의 데이터로 학습된 파운데이션 모델(FM)을 기반으로 합니다. FM은 이 데이터를 분석하여 패턴을 식별하고 유사한 새 콘텐츠를 생성하는 방법을 학습합니다. 특정 비즈니스 요구 사항을 충족하는 생성형 AI 애플리케이션을 구축하려면, 일반적으로 조직의 데이터로 기존 FM을 학습시켜 맞춤화해야 합니다.

이 데이터를 보호하기 위해서는 데이터 프라이버시 제어 및 IAM 정책 모범 사례를 고려해야 합니다.

팀은 FM을 맞춤화할 때 FM 개선에 사용되지 않으며 안전하게 저장되어 있는 모델 버전을 사용해야 합니다. **Amazon Bedrock**에 단일 테넌트 전용 용량을 설정하면 서비스에서 추론 인스턴스를 **Amazon Virtual Private Cloud**(VPC)에 연결하여 **Amazon Simple Storage Service**(S3)에서 읽고 쓸 수 있습니다.

IAM을 효과적으로 사용하면 적합한 관리자와 시스템이 적절한 조건에서 적절한 리소스에 접근할 수 있는지 검증할 수 있습니다. **AWS Well-Architected Framework**는 자격 증명 관리에 도움이 되는 설계 원칙과 아키텍처 모범 사례를 설명합니다. 이 리소스는 IAM 정책을 개발하고 위협 탐지 및 네트워크 보안과 같은 기타 보안 문제를 해결하는 데 유용한 도구입니다.

생성형 AI 애플리케이션 보호

애플리케이션 수준에서 생성형 AI를 보호하려면 위험을 지속적으로 식별, 분류, 해결 및 완화해야 합니다. 첫 번째 단계는 환경과 데이터를 안전하게 유지하기 위한 기존 모범 사례를 따르는 것입니다.

이에 기반하여 개발 프로세스의 초기 단계에서 보안을 적용할 방법을 고민해야 합니다. 그러면 작업이 간소화되고 개발 팀이 보안 병목 현상 없이 더 빠르고 자유롭게 혁신할 수 있습니다.

다음으로 모든 AI 애플리케이션의 세 가지 핵심 구성 요소인 입력, 결과물, 모델 자체를 보호하는 방법을 고려해야 합니다.

입력 보호

먼저 AI 시스템에 입력되는 데이터를 검토하세요. 사용자는 조작, 스푸핑 또는 프롬프트 주입과 같은 무결성 공격의 위험을 줄이기 위한 입력 필터링 없이 FM에 직접 접근해서는 안 됩니다. 이러한 공격 기법은 제어를 우회하거나 모델을 악용합니다. 입력을 보호하기 위해 고려해야 할 다른 전략으로는 데이터 품질 자동화, 지속적인 모니터링 및 위협 모델링이 있습니다.

결과물 보호

생성형 AI 애플리케이션 결과물에 대한 위험에는 정보 공개, IP 인시던트, 조직의 평판을 손상시킬 수 있는 모델의 오용이나 남용 등이 포함됩니다. 위협 모델을 개발할 때는 정보의 범위와 사용 맥락을 고려하고 복잡한 행동 탐지 및 모니터링을 포함해야 합니다.

모델 자체 보호

마지막으로, 공격자가 어떻게 모델 자체 또는 관련 구성 요소에서 데이터를 제거하려고 시도할 수 있을지 생각해 보세요. 위험에는 실제 세계나 데이터에 대한 모델의 잘못된 표현, 모델의 무결성 또는 가용성 손상이 포함됩니다. 비즈니스 목표에 맞게 위협을 모델링하고 이러한 위협 시나리오에 대해 모니터링하세요.

질문 2:

규정 준수 문제를 어떻게 해결할 수 있을까요?

조직은 생성형 AI 애플리케이션을 설계하고 개발하는 데 따르는 위험을 완화함으로써 파트너 및 고객과의 신뢰를 구축하고 브랜드 평판을 유지하며 규정 준수 요구 사항을 지속적으로 따를 수 있습니다.

생성형 AI 애플리케이션에 대한 법적 규제는 아직 초기 단계이며, 모범 사례에 대해 합의가 이루어지지 않았습니다. 따라서 여러 관할 구역에서 서로 상충하여 복잡하게 얽힌 기준과 감독 문제를 헤쳐 나가는 것은 복잡하고 지속적인 과제입니다.

법률 고문 및 개인정보 보호 전문가와 상담하여 생성형 AI 애플리케이션 구축의 요구 사항 및 영향을 평가하세요. 여기에는 특정 데이터 및 모델을 사용하기 위한 법적 권리를 검토하고 개인정보 보호, 생체 인식, 차별 금지 및 기타 사용 사례별 규정에 관한 법률의 적용 여부를 결정하는 것이 포함될 수 있습니다.

주, 도, 국가별로 서로 다른 법적 요구 사항과 전 세계적으로 제기되고 있는 새로운 AI 규정에 주의를 기울여야 합니다. 향후 배포 및 운영 단계에서 이러한 사항을 다시 고려해 보세요.

동료, AI 전문가 및 정부 기관과 협업하면 고객에게 법률 및 윤리적 AI 표준 준수에 대한 진정성을 보여주는 동시에 지속적으로 규정을 준수하는 데 도움이 될 수도 있습니다. 최근 Amazon은 백악관 및 6개의 주요 AI 기업과 함께 **책임 있고 안전한 AI 개발에 대한 자발적 약속**을 발표함으로써 그러한 참여의 가치를 입증하고 향후 협업의 토대를 마련했습니다.

인공 지능에 내재된 위험

ML을 사용하는 모든 솔루션과 마찬가지로 생성형 AI 애플리케이션은 기존 소프트웨어를 넘어서는 위험을 내재하고 있습니다. 생성형 AI로 애플리케이션을 안전하게 구축하고 배포하려면 다음과 같은 위험 완화 전략을 수립하고 실행해야 합니다.

- 편향되거나, 사실이 아니거나, 오해의 소지가 있거나, 유해하거나, 불쾌감을 주는 결과물
- 대규모 복잡성 및 비용
- 너무 커지거나 오래되거나 의도한 맥락에서 벗어난 데이터 세트
- 불투명성 증가 및 재현성에 대한 우려
- 완전히 정립되지 않은 테스트 표준 및 절차

다음 섹션에서는 이러한 위험 중 일부를 줄이기 위한 광범위한 전략과 생성형 AI 애플리케이션이 직업과 조직, 그리고 사회에 미치는 영향을 정의하는 모범 사례를 살펴보겠습니다.

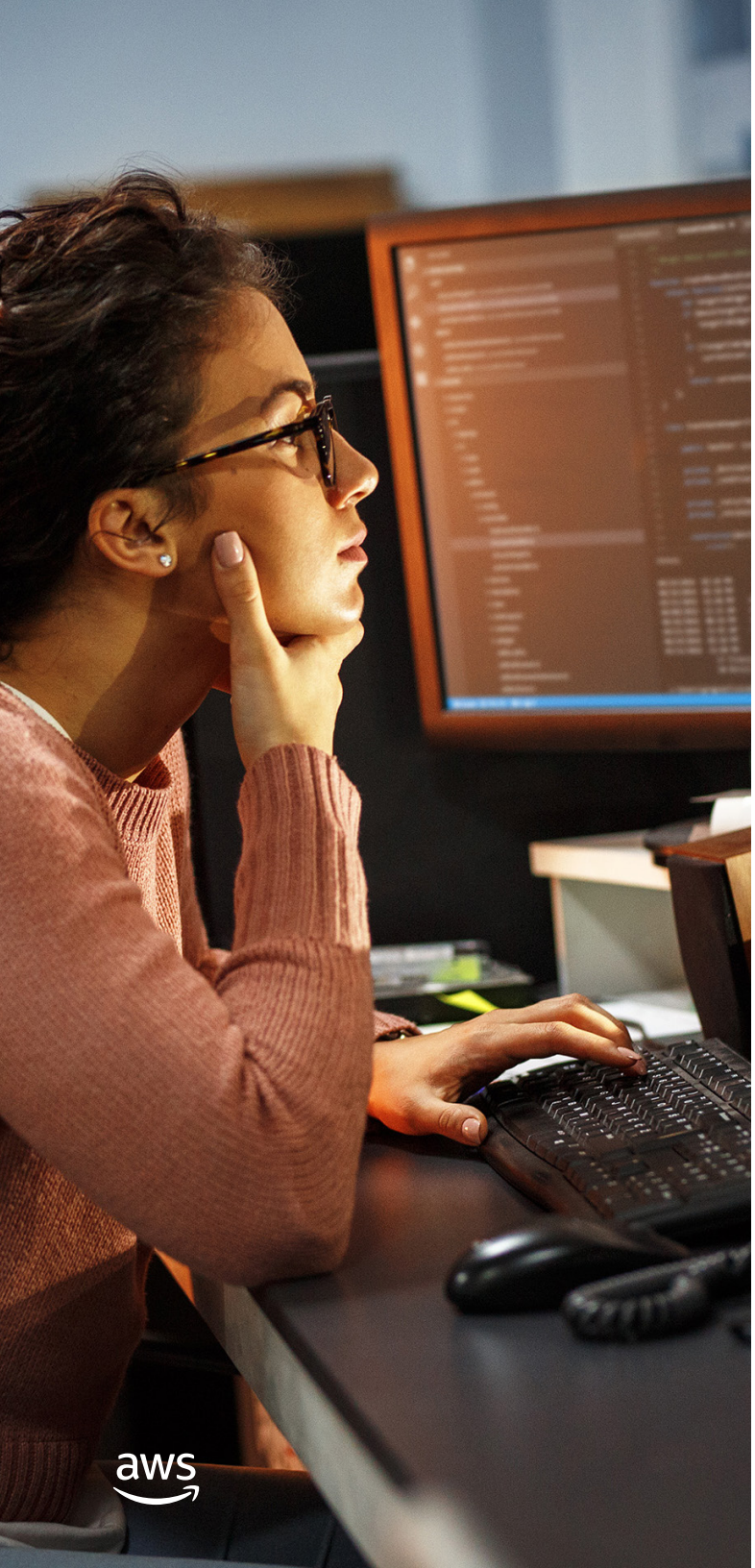
질문 3:

어떻게 해야 모델이 의도한 대로 동작할 수 있을까요?

생성형 AI의 책임 있는 사용은 비즈니스의 중요한 과제이자 지속적인 혁신을 위한 필수 조건이 되었습니다.

FM은 대규모 데이터 세트를 기반으로 학습하며 유사한 콘텐츠를 생성하는 방법을 이해하기 위한 복잡한 분석을 수행합니다. 많은 FM이 놀라운 성과를 내고 있지만, "콩 심은 데 콩 난다"는 오래된 격언은 여전히 유효합니다. FM에 부정확하거나 불완전하거나 편향된 데이터가 공급되는 경우 결과물에 유사한 결함이 나타날 수 있습니다.

결함이 있는 데이터는 오용, 악의적인 행동 및 기타 위험의 가능성을 높입니다. 생성형 AI 애플리케이션의 사용자, 범위, 기능이 확장됨에 따라 이러한 문제의 잠재적 영향도 커집니다.



책임 있는 AI 육성

책임 있는 AI 전략을 수립하면 이러한 위험을 해결하는 데 도움이 됩니다. 책임 있는 AI의 요소에는 설명 가능성, 공정성, 거버넌스, 개인정보 보호, 보안, 견고성 및 투명성이 포함됩니다. 여기에는 또한 애플리케이션이 다양한 문화와 인구 집단을 보고, 다루고, 영향을 주는 방식을 이해하는 것이 포함됩니다.

생성형 AI 개발 초기에 책임 있는 AI를 고려하는 것부터 시작하여 애플리케이션 수명 주기 전반에 걸쳐 이를 핵심 비전으로 삼는 것이 좋습니다. 비교적 작고 간단한 작업부터 시작하세요. 그런 다음 시간이 지남에 따라 책임 있는 AI가 설계, 개발, 운영에 영향을 미치는 방식의 규모를 조정하세요.

책임 있는 AI 및 거버넌스 정책을 수립할 때는 생성형 AI 애플리케이션이 사용자, 고객, 직원 및 사회에 어떤 영향을 미칠지 고려하세요. 알고리즘의 공정성, 다양하고 포괄적인 대표성, 편향 탐지를 반드시 고려해야 합니다.

유해성 해결

대규모 언어 모델(LLM)의 유해성은 무례하거나 존중이 결여되어 있거나 불합리한 텍스트의 생성을 의미합니다. 생성형 AI 애플리케이션의 유해성을 방지하고 공정성을 보장하는 데 도움이 되는 많은 전략이 있습니다. 예를 들어, 학습 데이터에서 불쾌감을 주는 언어나 편향된 문구를 식별하여 제거할 수 있습니다. 또한 애플리케이션의 특정 사용 사례, 대상 고객 또는 애플리케이션이 수신할 가능성이 가장 높은 프롬프트와 쿼리에 초점을 맞춰 보다 엄격한 공정성 테스트를 실시할 수도 있습니다.

또한 다양한 유형과 유해성 정도를 식별하는 주석이 달린 데이터 세트를 기반으로 가드레일 모델을 학습시킬 수 있습니다. 이를 통해 FM은 학습 데이터, 입력 프롬프트 및 생성된 결과물에서 원치 않는 콘텐츠를 자동으로 감지하고 필터링하는 방법을 학습할 수 있습니다.

프라이버시 보호

생성형 AI 애플리케이션을 사용할 때 민감한 정보, 영업 비밀 및 IP가 원치 않게 노출되는 것을 방지하기 위해 몇 가지 조치를 취할 수 있습니다.

모델 삭제는 개인정보 보호 문제를 해결하기 위한 한 가지 방법입니다. 여기에는 잘못 사용된 데이터를 식별하는 즉시 제거하여 해당 데이터가 FM의 구성 요소에 미치는 영향을 제거하는 것이 포함됩니다.

또 다른 접근 방식은 샤딩입니다. 샤딩은 훈련 데이터를 작은 부분으로 나누어 개별 하위 모델을 학습시키고, 최종적으로는 그러한 하위 모델을 결합하여 전체 FM을 만드는 방식입니다. 이 방법을 사용하면 개인 정보를 포함하고 있거나 개인 정보가 노출될 위험이 있는 FM을 훨씬 간단하게 교정할 수 있습니다. 전체 모델을 재학습시키는 대신, 샤드에서 원하지 않거나 잘못 사용된 데이터를 제거한 다음 해당 하위 모델을 재학습시키기만 하면 됩니다.

필터링 및 차단도 효과적인 접근 방식일 수 있습니다. 이 방식은 사용자에게 표시하기 전에 보호된 정보를 생성된 콘텐츠와 명시적으로 비교합니다. 두 항목이 너무 유사하면 노출되지 않도록 콘텐츠를 표시하지 않거나 교체합니다. 학습 데이터에 특정 콘텐츠가 표시되는 횟수를 제한하는 것도 도움이 될 수 있습니다.

설명 가능성 및 감사 가능성 향상

더 책임 있는 AI를 만들기 위해 애플리케이션의 결과물에 영향을 미치는 방법론과 주요 요소에 대한 설명이 필요할 수 있습니다. 감사 가능성은 책임 있는 AI의 또 다른 중요한 구성 요소입니다. 생성형 AI 애플리케이션의 개발 및 운영을 추적하고 검토할 수 있는 메커니즘을 구현하세요. 이는 문제의 근본 원인을 추적하고 거버넌스 요구 사항을 충족하는 데 도움이 됩니다.

개발의 전 단계에 걸쳐 관련 설계 결정사항 및 입력을 문서화하세요. 추적 가능한 기록을 남기면 내부 또는 외부 팀이 생성형 AI 애플리케이션의 개발 및 기능을 평가하는 데 도움이 됩니다.

계속 책임져 나가기

마지막으로, 책임 있는 AI 정책을 지속적으로 준수하기 위한 방법에 대해 생각해 보세요. 배운 교훈과 경험을 적용하여 보안 및 개인정보 보호 수칙을 발전시키세요. 모든 직원에게 생성형 AI 보안 수칙을 준수할 의무에 대해 정기적으로 교육하세요. 책임 있는 AI 문화를 조성하고, 적절한 도구를 사용하여 모델 성능을 모니터링하고 위험을 알리며, 팀이 필요할 때 모델과 구성 요소를 검사할 수 있는 환경을 만드세요. 확신이 들 때까지 몇 번이고 테스트하세요.

질문 4:

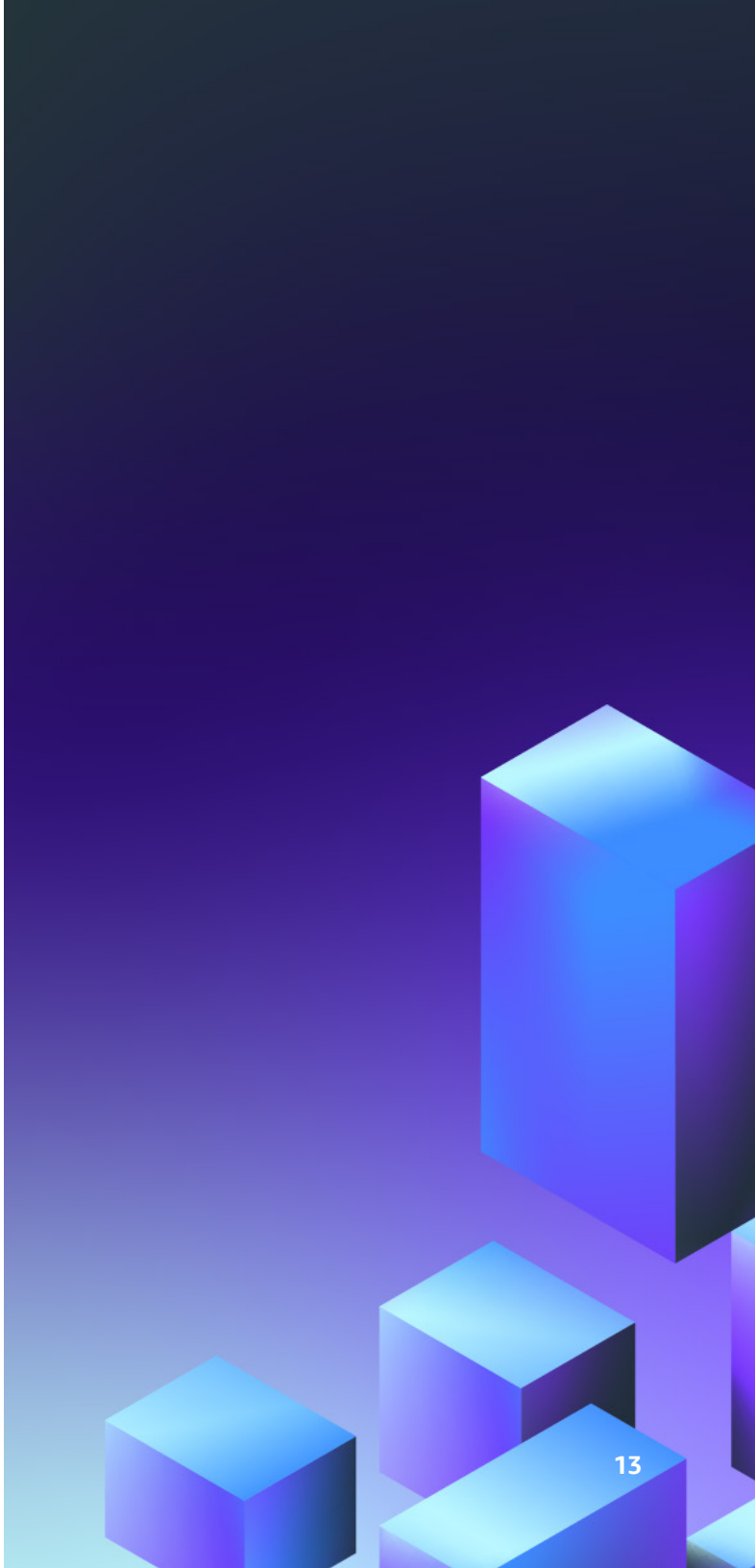
어디서부터 시작해야 할까요?

생성형 AI 애플리케이션을 보호하는 것은 간단한 일이 아니며, 이를 달성하기 위한 보편적인 방법은 없습니다. 그러나 적절한 공급업체와 협력하고 적절한 도구를 배포하면 성공으로 가는 길이 훨씬 더 명확하게 보입니다.

예를 들어 **Amazon Bedrock**을 사용하면 훨씬 빠르고 간편하게 생성형 AI 보안 애플리케이션을 개발할 수 있습니다. Amazon Bedrock은 Amazon 및 주요 AI 스타트업의 FM을 API를 통해 제공하는 완전 관리형 서비스입니다.

Amazon Bedrock으로 모델을 맞춤화하면 팀이 대량의 데이터에 주석을 달지 않아도 서비스에서 특정 작업에 맞게 모델을 미세 조정할 수 있습니다. 그런 다음 Amazon Bedrock은 사용자만 액세스할 수 있는 기본 FM의 별도 사본을 만들고 이 비공개 사본을 훈련합니다. 어떤 데이터도 원본 기본 모델을 학습시키는 데 사용되지 않으므로 독점 데이터의 기밀성과 보안을 유지할 수 있습니다.

또한 **Amazon VPC** 설정을 구성하여 Amazon Bedrock API에 접근하고 모델에 미세 조정 데이터를 안전하게 제공할 수 있습니다. 데이터는 서비스 관리형 키를 통해 전송 중과 저장 중에 항상 암호화됩니다. 또한 **AWS PrivateLink**를 사용하면 공공 인터넷을 거치지 않고 오로지 AWS 네트워크를 통해서만 Amazon Bedrock에 AWS 클라우드 데이터를 전달할 수 있습니다.



AWS를 통한 개인정보 보호 강화

생성형 AI 애플리케이션을 구축할 때 Amazon Bedrock, 다른 서비스(예: **Amazon SageMaker**), 자체 도구 등 어떤 도구를 사용하더라도 AWS에서 애플리케이션을 실행하고 관리하면 업계 최고 수준의 개인정보 보호 및 제어의 이점을 누릴 수 있습니다.

AWS는 전 세계 고객의 요구 사항을 충족하기 위해 143개의 보안 표준 및 규정 준수 인증을 지원합니다. 자체 **AWS Key Management Service**(Amazon KMS) 키를 사용하여 모든 데이터를 저장 상태에서 암호화할 수 있으므로 데이터 및 FM이 저장되고 액세스되는 방식을 완벽하게 제어하고 파악할 수 있습니다.

결론

다음 단계

AWS는 보안, 개인정보 보호 및 규정 준수 목표를 달성하는 동시에 비즈니스를 성장시키는 생성형 AI 애플리케이션을 구축하는 데 도움이 되도록 최선을 다하고 있습니다.

AWS는 생성형 AI 애플리케이션을 안전하게 설계, 개발 및 운영할 수 있다는 굳건한 믿음이 있습니다. 또한 이러한 기술에 대한 보안 및 개인정보 보호 우려의 타당성을 인정합니다. **생성형 AI**는 데이터 프라이버시, IP, 입법 감시, 평등, 투명성과 관련된 문제의 정의, 측정 및 해결에 있어 새로운 과제를 만들어냅니다.

신제품 출시, 솔루션의 복잡성 및 규모 조정, 새로운 훈련 매개변수, 계속 증가하는 데이터 세트로 인해 생성형 AI 보안의 중요성은 앞으로 더욱 높아질 것입니다. 지금 생성형 AI 워크로드를 위한 효과적이고 포괄적인 보안 전략을 개발하면 최고의 경쟁 우위를 확보하고 빠르게 다가오는 미래에 대비할 수 있습니다.

다행히도 생성형 AI 애플리케이션을 안전하게 설계하고 개발하고 실행하는 데 필요한 기본 제어 기능이 수년 전부터 있어 왔으며, 이는 **AWS Well-Architected Framework**의 원칙과 같이 신뢰할 수 있고 검증된 클라우드 보안 원칙에 부합합니다.

이 eBook에 설명된 사례를 살펴봄으로써 여러분은 이미 생성형 AI 워크로드의 보안을 위한 첫걸음을 내디딘 것입니다.

이제 AWS와 함께 다음 단계로 나아가세요. AWS는 새로운 주제를 빠르게 파악하고, 고유한 과제를 해결하고, 생성형 AI의 최대한 활용하면서도 데이터, 고객, 비즈니스를 보호하는 데 필요한 심층적인 통찰력과 구체적인 지침을 제공할 수 있습니다.

AWS 기반 생성형 AI에 대해 자세히 알아보기 ›

Amazon Bedrock으로 빠르게 시작하기 ›

Amazon SageMaker에서 FM 구축 및 맞춤화 ›

AWS로 클라우드에서 보안을 더욱 강화 ›

책임 있는 AI를 이론에서 실천으로 전환 ›