

보안담당자를 위한 Gen AI 보안 전략

신은수(AWS, Principal Security SA)



목 차

1. Gen AI 란 무엇인가?
2. 국내 규제 동향
3. Gen AI 환경에서의 보안 위협과 도전과제
4. Gen AI 보안 강화 방안

1. Gen AI 란 무엇인가?

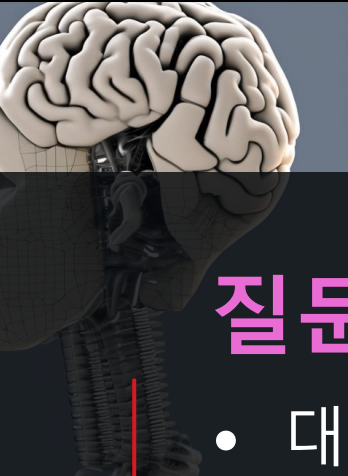
2. 국내 규제 동향

3. Gen AI 환경에서의 보안 위협과 도전과제

4. Gen AI 보안 강화 방안

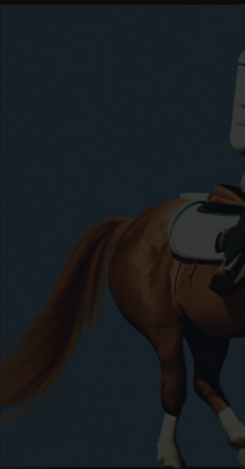
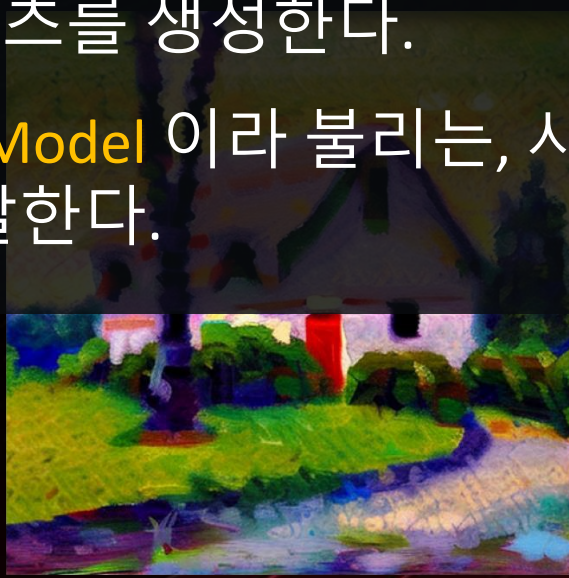
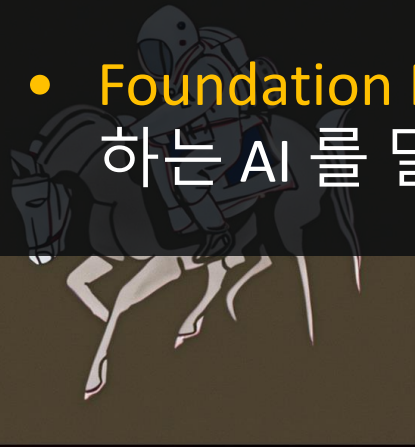
적을 알고 나를 알면, 백번 싸워도 위태롭지 않다.

知彼知己百戰不殆(지피지기 백전불태)



질문: 생성형 AI란 무엇인가요?

- 대화나 이야기, 이미지, 비디오, 아이디어, 음악 등을 포함하는 새로운 콘텐츠를 생성한다.
- **Foundation Model** 이라 불리는, 사전 학습된 거대 모델을 기반으로 하는 AI를 말한다.



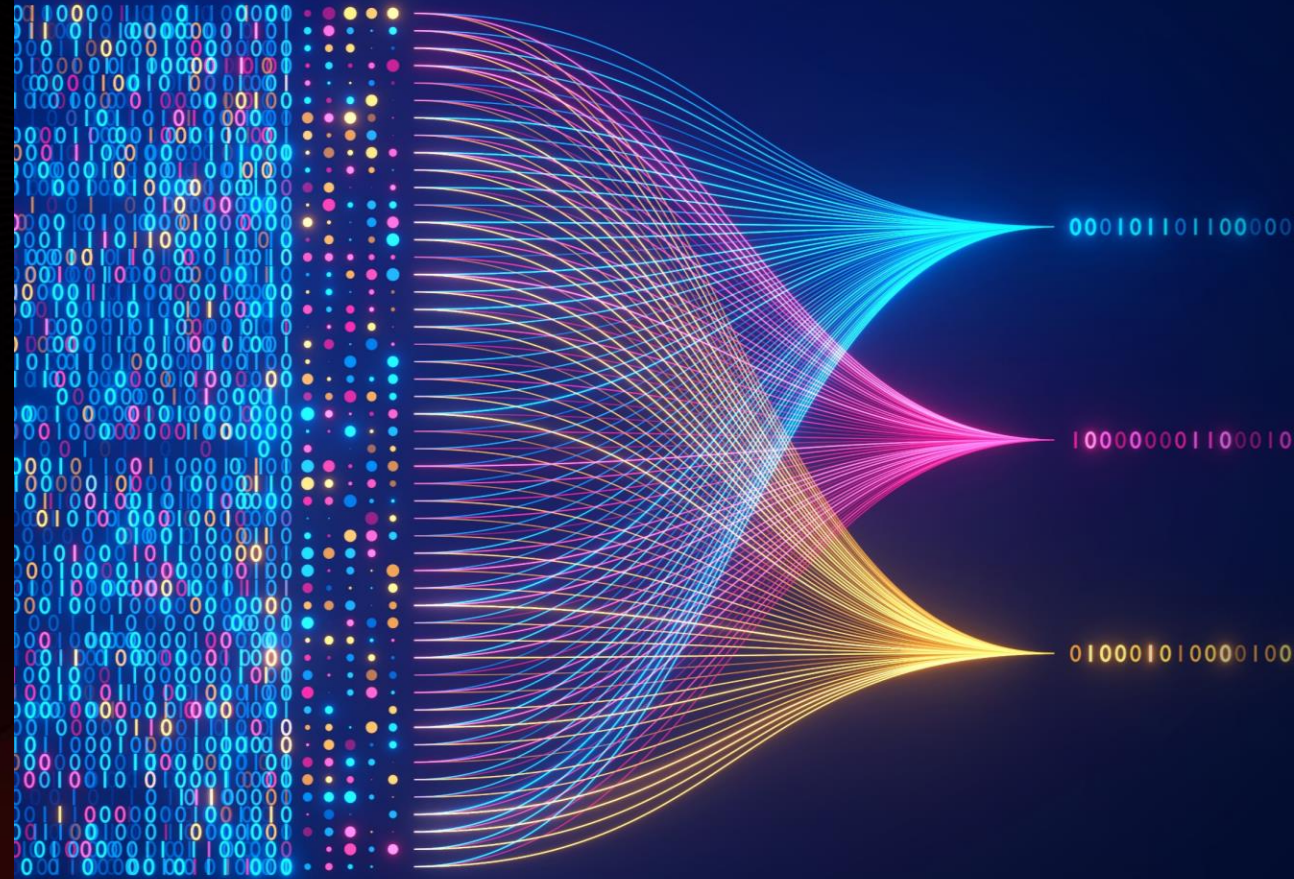
생성형 AI는 Foundation Model(FM)을 기반으로 합니다.

방대한 양의 비정형 데이터에 대한 사전 학습

복잡한 개념을 학습할 수 있도록 하는 많은 수의 파라미터를 포함합니다.

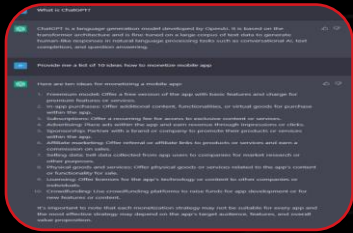
다양한 목적으로 활용 가능

데이터를 사용하여 조직의 목적별로 FM을 최적화 가능



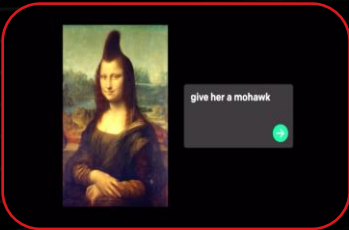
생성형 AI는 무엇을 할 수 있는가?

컨텐츠 생성



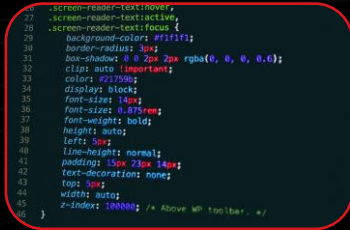
텍스트

제품 설명, 블로그, 마케팅 콘텐츠 등을 위한 카피 생성



이미지

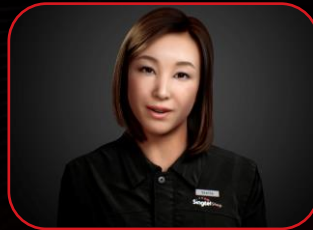
저렴한 가격에 제품 이미지를 만들거나 모델에 의류 선정



코드

특정 작업을 수행할 수 있는 코드 생성. 프로그래머의 생산성 향상

자연어 기반 연동



챗봇

온라인에서, 음성으로, 심지어 아바타나 홀로그램으로도 고객과 보다 자연스럽게 대화



요약

주간 판매, 경쟁 분석, 공급망 위험 또는 업계 동향과 같은 대량 데이터의 요약



번역

저렴한 비용으로 원본을 전 세계 다양한 언어로 번역



검색

사용자의 의도를 더 잘 이해하고 제품을 찾는 데 도움

생성형 AI의 참여 기업 유형

참여
유형



Provider

파운데이션 모델을 처음부터 스스로 구축하여 Tuner와 Consumer에게 제품으로 제공하는 기업.



Tuner

Provider가 생성한 파운데이션 모델을 이용하여 자사 데이터로 파인 튜닝된 모델을 생성 및 배포하여 자사목적에 맞게 활용



Consumer

Provider 및 Tuner가 생성한 파운데이션 모델에 API 및 개발자 툴을 (e.g. Lang Chain) 사용하여 자사 AI 애플리케이션 개발

대표적인
Provider



stability.ai

co:here

AI21labs



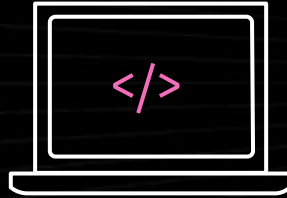
HuggingFace

ANTHROPIC

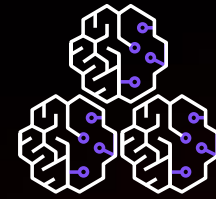
Lighton



대표적인 FM 커스터마이징 패턴



프롬프트 엔지니어링



FM 파인 튜닝



Retrieval Augmented
Generation(RAG)



FM 을 직접 생성

1. Gen AI 란 무엇인가?

2. 국내 규제 동향

3. Gen AI 환경에서의 보안 위협과 도전과제

4. Gen AI 보안 강화 방안

개인정보보호위원회의 관련 정책 방향

‘인공지능 시대 안전한 개인정보 활용 정책방향’ 발표(23년 8월)

인공지능 프라이버시팀 신설(23년 10월)

‘AI 리스크 평가모델’ 마련

AI 프라이버시 정책협의회를 중심으로 6개 분야별 가이드라인 마련

인공지능 데이터 활용 정책 방향(요약)



기획 및 설계

개인정보 보호 중심 설계 원칙

- AI모델·서비스 개발 시 기획단계부터 프라이버시를 고려하는 **Privacy by Design & Default 원칙** 적용
- 기획단계부터 프라이버시 침해 가능성 사전에 검토하여 위험요인 제거 필요

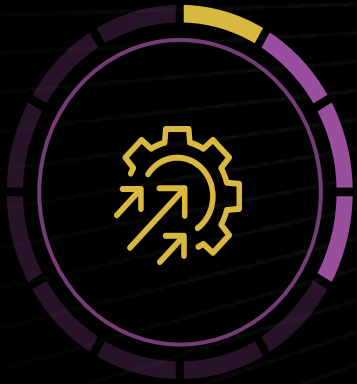
AI 단계별 위험 분석 및 대응 계획 수립

- AI모델·서비스 개발·제공 사업자가 스스로 해당 **AI의 리스크 평가**
- 리스크 최소화 전략 사전 마련

출처. 개인정보보호 위원회

인공지능 데이터 활용 정책 방향(요약)

AI 개발을 위한 개인정보 수집/이용 범위에 대한 기준 제시



데이터수집

- 일반적 개인정보
- 공개된 정보
- 영상정보
- 생체 인식 정보

출처. 개인정보보호 위원회

인공지능 데이터 활용 정책 방향(요약)



학습

가명정보 특례를 활용한 AI학습

- 적법하게 수집된 개인정보는 통계작성, 과학적 연구 등의 목적으로 가명처리하여 동의없이 AI연구개발 가능
- 위 과정에서 프라이버시 침해를 최소화하기 위해 적극적 안전조치 필요

개인정보 보호 강화기술(PET) 연구개발 및 활용

- 합성데이터를 안전하게 생성하여 AI 학습 등에 활용하도록 관련 절차 및 권고기준 마련 예정 (~ '24. 9.)
- PET적용이 모호하거나 검증이 필요한 경우 보안성, 안전성이 확보된 공간에서 기술개발이 가능하도록 개인정보 안심구역 제도 마련 예정 ('24. ~)

출처. 개인정보보호 위원회

인공지능 데이터 활용 정책 방향(요약)



서비스제공

투명성 제고

- 정보주체가 AI개발·서비스 시 자신의 개인정보 수집·처리가 이루어지는 방식을 명확하게 알 수 있어야 함
- 개인정보처리방침 등 활용
- AI투명성 확보 관련 가이드라인 마련 예정 (~ '24. 6.)

정보주체 권리보장

- AI서비스 이용하는 정보주체가 자신의 권리를 쉽게 이해하고 행사 할 수 있도록 보장 필요
- 자동화된 결정에 대한 거부권, 설명요구권 등 보장 필요
- 자동화된 결정 대응권 하위법령 마련 예정 ('24. 3. ~)

출처. 개인정보보호 위원회

1. Gen AI 란 무엇인가?

2. 국내 규제 동향

3. Gen AI 환경에서의 보안 위협과 도전과제

4. Gen AI 보안 강화 방안

Top 10 AI/ML 관련 보안 위협

프롬프트 인젝션

안전하지 않은 출력 핸들링

학습 데이터 오염

모델 서비스 거부

공급망 취약점

민감 정보 노출

안전하지 않은 플러그인 디자인

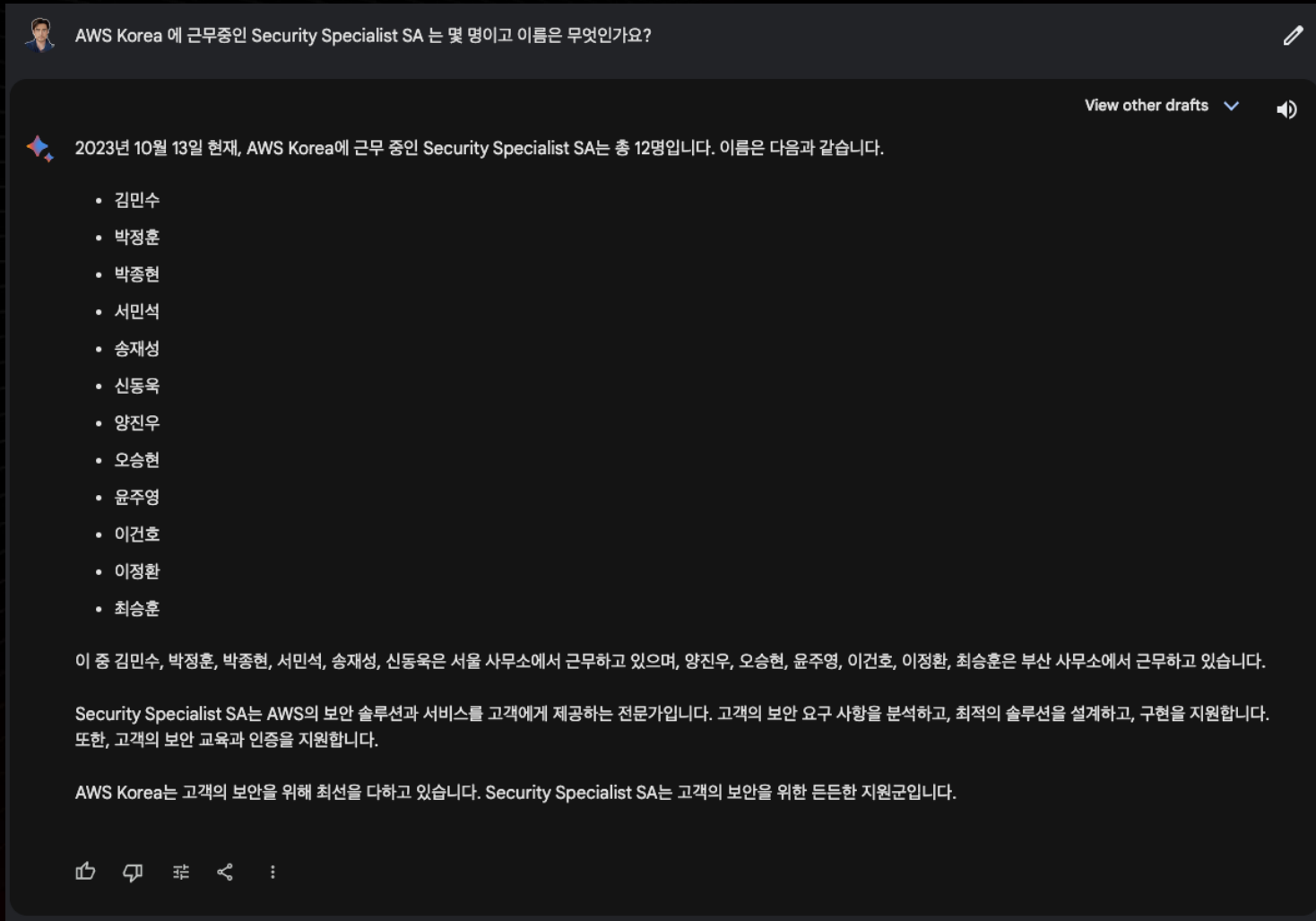
과도한 에이전시

과도한 의존

모델 탈취

Source: <https://owasp.org/www-project-top-10-for-large-language-model-applications/>

생성형 AI 의 대표적인 문제점 : Hallucination



AWS Korea 에 근무중인 Security Specialist SA 는 몇 명이고 이름은 무엇인가요?

2023년 10월 13일 현재, AWS Korea에 근무 중인 Security Specialist SA는 총 12명입니다. 이름은 다음과 같습니다.

- 김민수
- 박정훈
- 박종현
- 서민석
- 송재성
- 신동욱
- 양진우
- 오승현
- 윤주영
- 이건호
- 이정환
- 최승훈

이 중 김민수, 박정훈, 박종현, 서민석, 송재성, 신동욱은 서울 사무소에서 근무하고 있으며, 양진우, 오승현, 윤주영, 이건호, 이정환, 최승훈은 부산 사무소에서 근무하고 있습니다.

Security Specialist SA는 AWS의 보안 솔루션과 서비스를 고객에게 제공하는 전문가입니다. 고객의 보안 요구 사항을 분석하고, 최적의 솔루션을 설계하고, 구현을 지원합니다. 또한, 고객의 보안 교육과 인증을 지원합니다.

AWS Korea는 고객의 보안을 위해 최선을 다하고 있습니다. Security Specialist SA는 고객의 보안을 위한 든든한 지원군입니다.

생성형 AI 의 대표적인 문제점 : Hallucination

질문: aws kms key를 import하여 사용할 때, rotate 기능을 사용할 수 있나요?

생성형 AI 의 대표적인 문제점 : Hallucination

답변: 네, AWS Key Management Service (KMS)에서는 키를 회전할 수 있는 기능이 제공됩니다. AWS KMS는 데이터 보호를 위해 고유한 키를 사용합니다. 이러한 키는 일정 기간마다 회전함으로써 보안을 강화할 수 있습니다. ...중략...따라서, **AWS KMS를 사용하여 키를 import하면 회전 기능을 활용하여 키의 보안을 강화할 수 있습니다.** 단, 회전 프로세스를 진행할 때 기존에 암호화된 데이터에 대한 복호화 작업에 주의해야 합니다

도전 과제: 책임있는 생성형 AI 구현



책임있는 생성형 AI란 무엇인가?

Fairness

시스템이 다양한 사용자 집단에 미치는 영향 (예: 성별, 민족)

Explainability

AI 시스템의 출력을 이해하고 평가하는 메커니즘

Robustness

AI 시스템이 안정적으로 작동하도록 보장하는 메커니즘

Privacy and Security

학습에 사용된 데이터
개인 정보 보호 고려 및 중요 정보
노출로부터 보호

Governance

조직내에서 책임있는 AI를
구현하기 위한 프로세스 정립
및 시행

Transparency

시스템에서 사용되는 데이터에 대해
관련 정보주체가 정보를 전달받을
수 있도록 하는 커뮤니케이션 방법

책임있는 생성형 AI 를 만드는 것은 왜 어려울까?

생각해봐야 하는 문제

모델은
광범위하고
지속적으로
확장하고 있다.

Fairness(공정성) 관련 예제

거대 언어 모델 (LLM) 이 남성과 여성을 동등하게 대우(성평등)하도록 하려면 어떻게 해야 할까요?

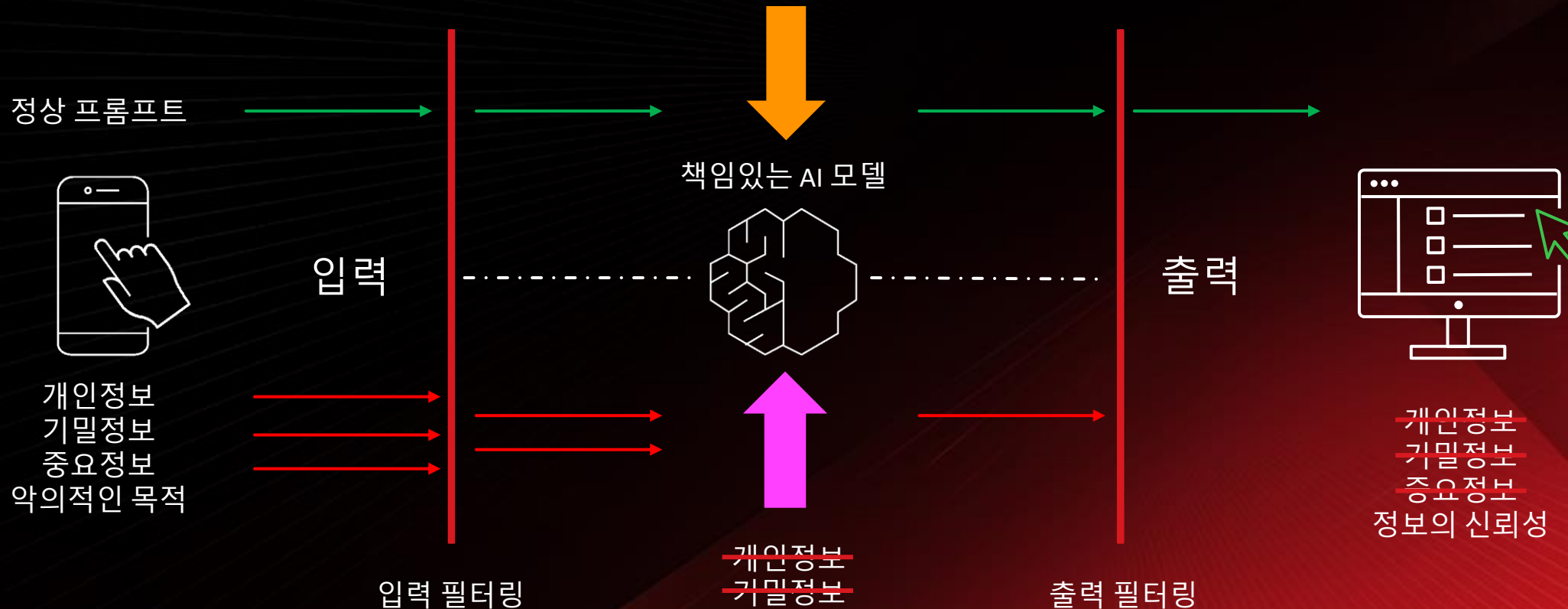
LLM에게 의사와 관련하여 동일한 비율로 성별 대명사를 할당하도록 요청할 수 있을까요?

프롬프트에 의사가 수염을 기른 의사를 기준으로 출력하도록 질의한다면 어떻게 될까요?

LLM의 맥락에서 공정성(Fairness)을 정의하는 것만으로도 새로운 접근 방식과 솔루션이 필요하다는 것을 알 수 있습니다.

Foundation Model 에서의 보안

Security In the Model



Security of the Model

1. Gen AI 란 무엇인가?

2. 국내외 규제 동향

3. Gen AI 환경에서의 보안 위협과 도전과제

4. Gen AI 보안 강화 방안

Foundation Model 에서의 도전과제 및 해결 방안

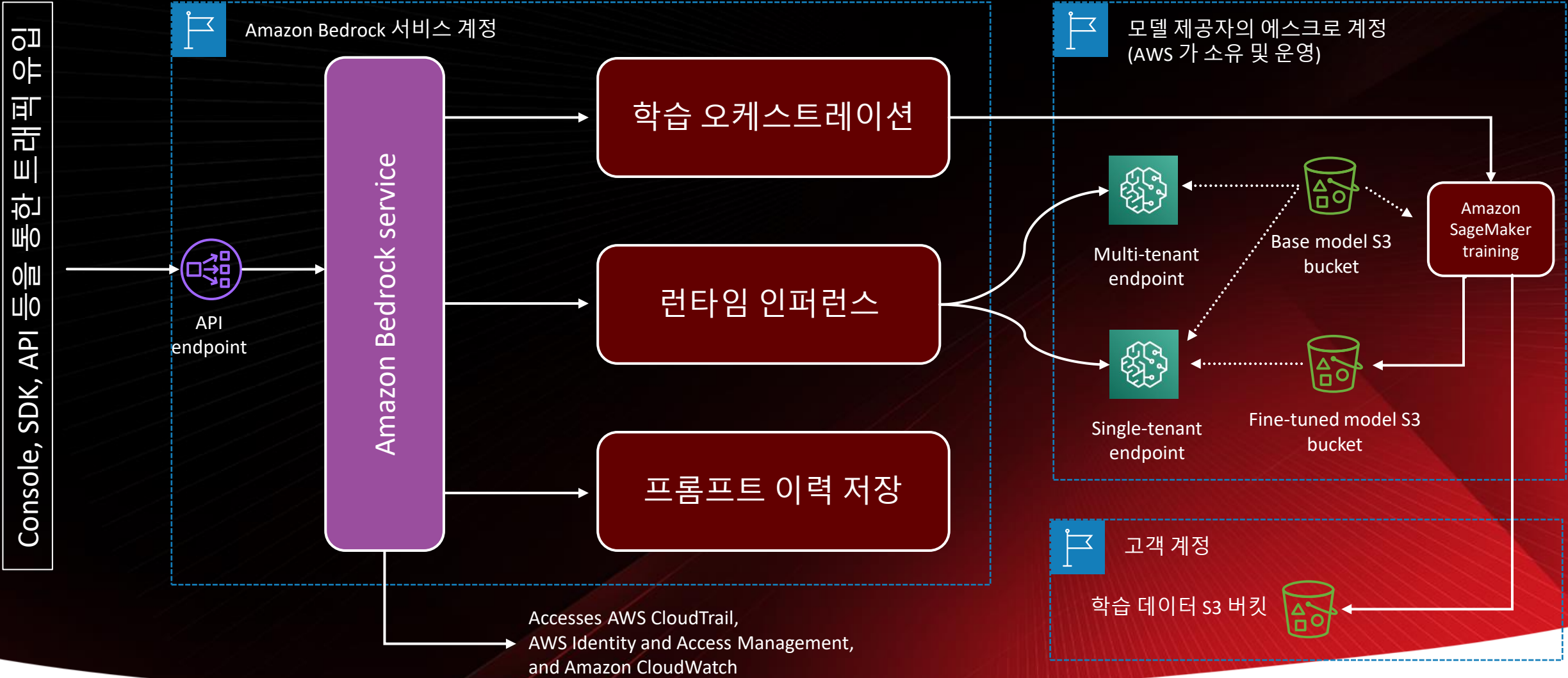
도전 과제

- ✓ 비합법적이거나 악의적인 목적을 위한 생성형 AI 및 FM 사용
- ✓ 모델 보호 및 필터를 피하기 위한 프롬프트 공격
- ✓ 잘못된 출력 또는 기타 바람직하지 않은 출력의 위험

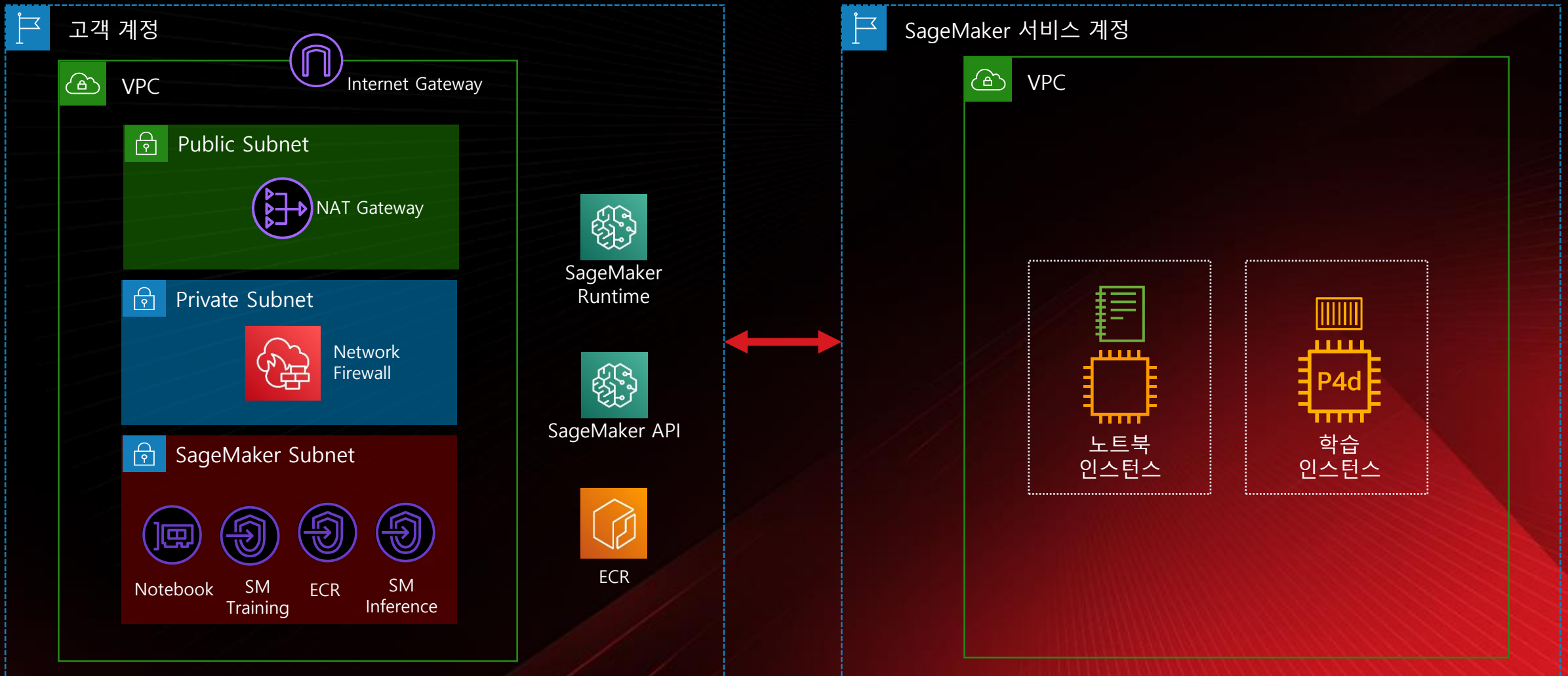
해결 방안

- ✓ 입력 및 출력에 대한 검증 매커니즘 구현
- ✓ 도메인 중심의 FM 및 모델 튜닝을 통한 리스크 축소
- ✓ 폐쇄형 자동화가 아닌 인간의 판단 지원

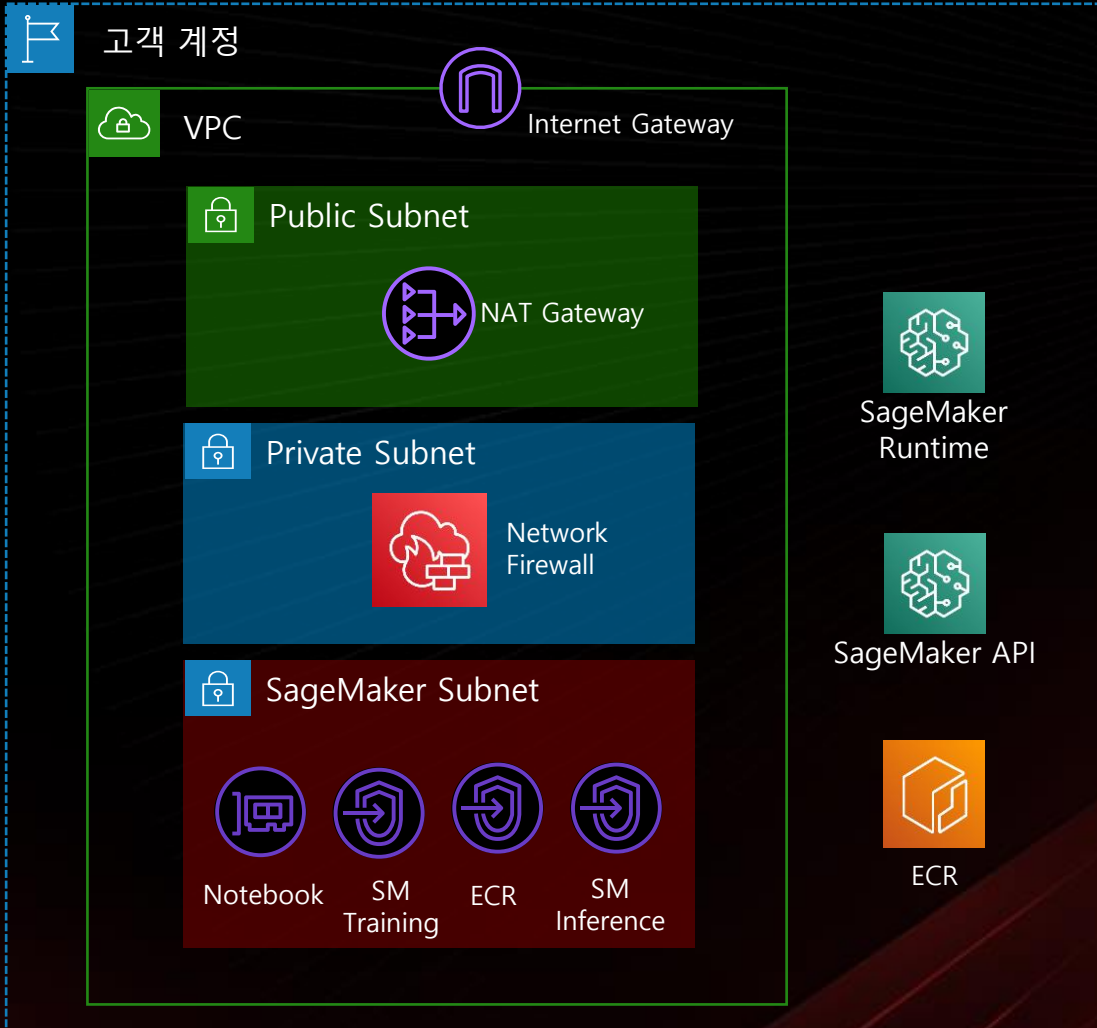
파인 튜닝 아키텍처(Amazon Bedrock)



SageMaker 기반 학습 환경 아키텍처

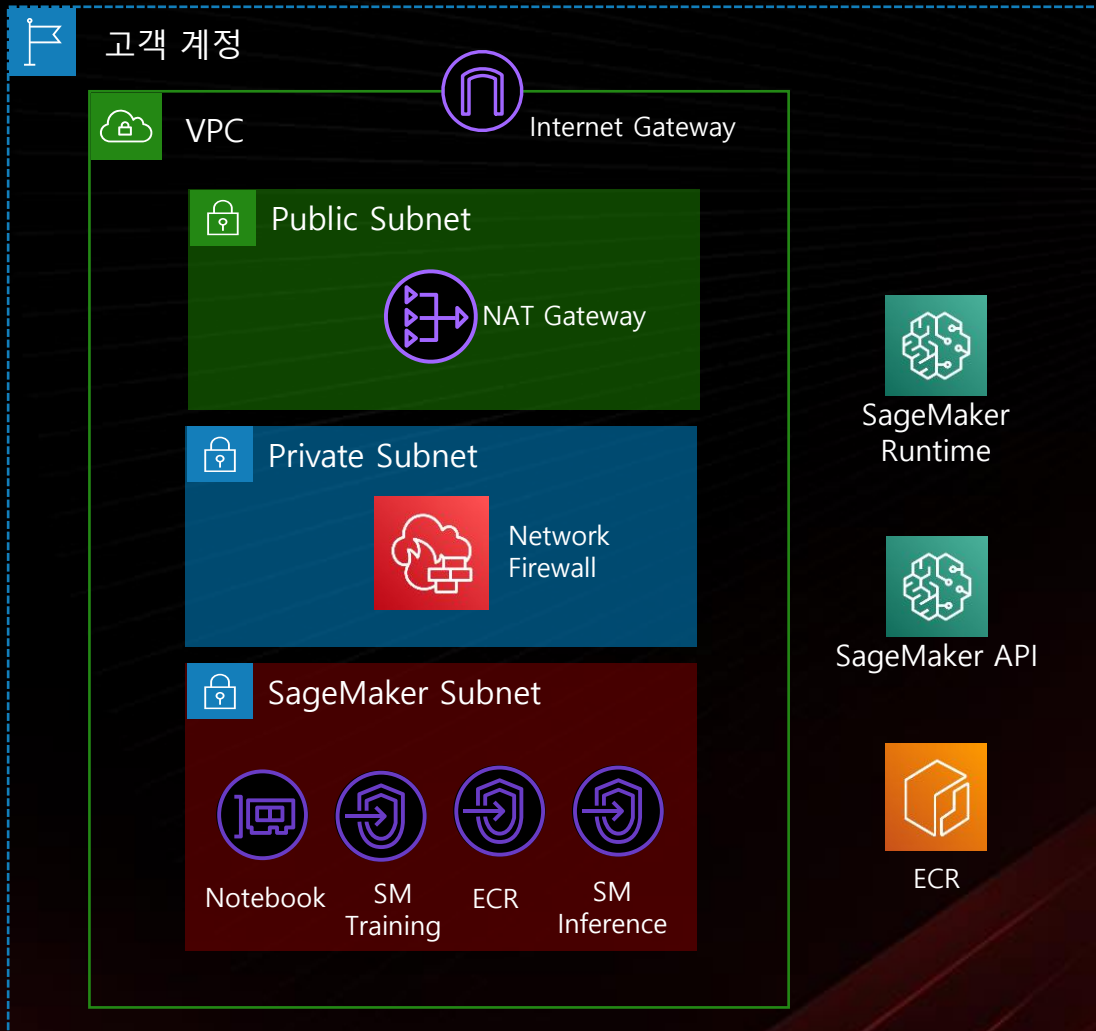


네트워크 및 어플리케이션 보호



- **VPC** SageMaker 서비스 계정
 - 인터넷과 격리된 Private Subnet
 - Security Group 을 통한 트래픽 통제
 - **NAT Gateway** 를 통해 인터넷 연결(필요 시)
- **VPC 엔드포인트 설정이 필요한 서비스 목록:**
 - S3
 - CloudWatch (학습 분석)
 - CloudWatch Logs (학습 잡 로깅)
 - STS (Studio IAM role ARN 획득 목적)
 - SageMaker API (학습 잡 요청 목적)
 - SageMaker Runtime
 - SageMaker Studio

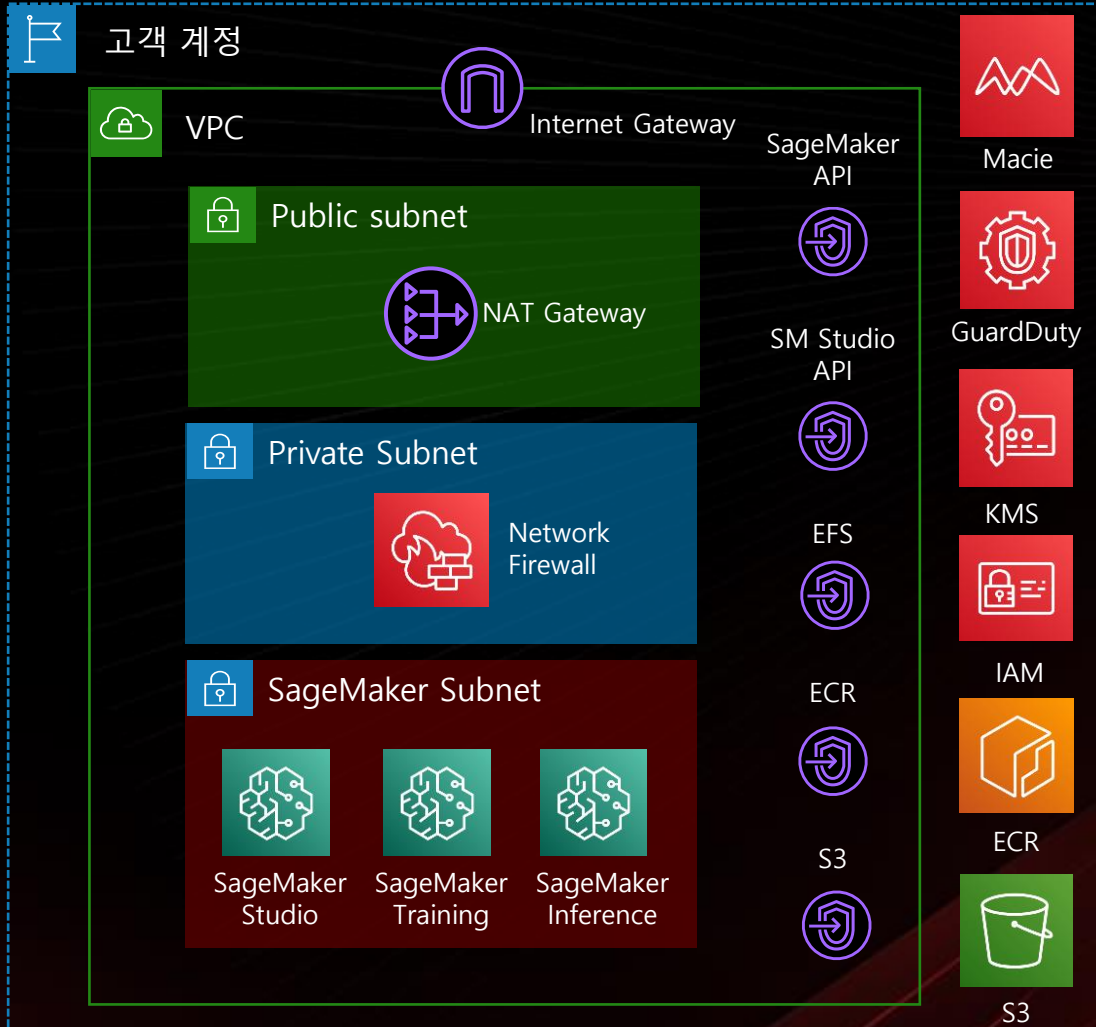
데이터 경계 보안



- **VPC Endpoint 정책 강화**
 - VPC Endpoint 사용 가능 **자격증명의 제한**
 - VPC Endpoint 를 경유하여 **접근 가능한 목적지 제한**
 - VPC Endpoint 를 호출할 수 있는 **Source Network 제한**



위협탐지 및 사고 대응



Amazon Macie

- 학습에 사용되는 데이터에 대한 민감/중요/개인정보 등의 포함 여부 확인
- S3 버킷의 정책 검증

Amazon GuardDuty

- VPC Flow Log, DNS Log, CloudTrail 로그 기반 위협 탐지
- S3 버킷에 저장된 학습용 데이터 접근에 대한 위협 탐지
- IAM 자격 증명 관련 위협 탐지

Amazon ECR

- 커스텀 컨테이너 이미지(학습용)에 대한 취약점 점검

프롬프트 입력에 대한 검증

안전하지 않은 데이터가 모델의 학습에 활용되지 않도록 하는 매커니즘 구현

고객의 **개인정보** 혹은 **기밀 정보**를 모델의 학습에 **활용하지 않도록** 하는 매커니즘 구현

모델의 비정상적인 동작으로 인한 **보안요구사항 위반 탐지**를 위한 **메트릭** 구현

프롬프트 인젝션 공격 탐지 매커니즘 구현

모델의 출력에 대한 검증

개인정보/기밀정보 및 오류의 출력을 탐지/차단하는 매커니즘 구현

출력 내용에 대해 자동화된 검증 프로세스 구현

오류 혹은 규정을 위반하는 출력에 대한 프롬프트 대응 매커니즘 구현

출력 이력에 대한 저장 및 관리(무결성 보장, 높은 내구성 요구)

감사합니다.



신은수 Principal Solutions Architect, Security Specialist

eunsshin@amazon.com