DELL Technologies

# GenAI 및 LLM 관련 사이버 보안 문제 상위 10개

시작하기

# 소개

AI(인공 지능)는 조직의 운영 방식을 혁신하고, GenAI(Generative AI) 및 LLM(대규모 언어 모델)은 현대 엔터프라이즈 환경에서 중요한 워크로드가 되어가고 있습니다.

다른 워크로드와 마찬가지로 이러한 애플리케이션은 자체적인 복잡성과 취약성을 해결해야 합니다. 기업이 혁신, 효율성 및 경쟁 우위 확보를 위해 계속해서 AI를 도입함에 따라 이러한 애플리케이션의 보안을 확보하는 것이 근본적인 필수 요소가 되었습니다. 좋은 사이버 위생은 모든 워크로드를 보호하는 토대이며, 모든 워크로드에서 보안을 우선시하는 것처럼 AI에 있어 올바른 사이버 위생을 실시하는 것이 중요합니다. 여기에는 적절한 시스템 패치 적용, 다단계 인증, 역할 기반 액세스 및 네트워크 세분화와 같은 관행 구현이 포함됩니다. 이러한 조치는 기본적이지만 이 기능이 워크로드의 특정 아키텍처와 그 사용에 어떻게 부합하는지 이해하는 것이 핵심입니다.

Dell은 AI 워크로드와 여기에서 발생하는 고유한 보안 문제를 심층적으로 이해하고 있습니다. Dell은 위협 요인이 이러한 워크로드를 표적으로 삼는 방식을 파악하여 강력한 보안 전략을 수립하도록 지원합니다. 여기에는 훈련 데이터 오염, 모델 도난 또는 조작, 데이터 세트 재구성 등과 같은 위험 해결이 포함됩니다.

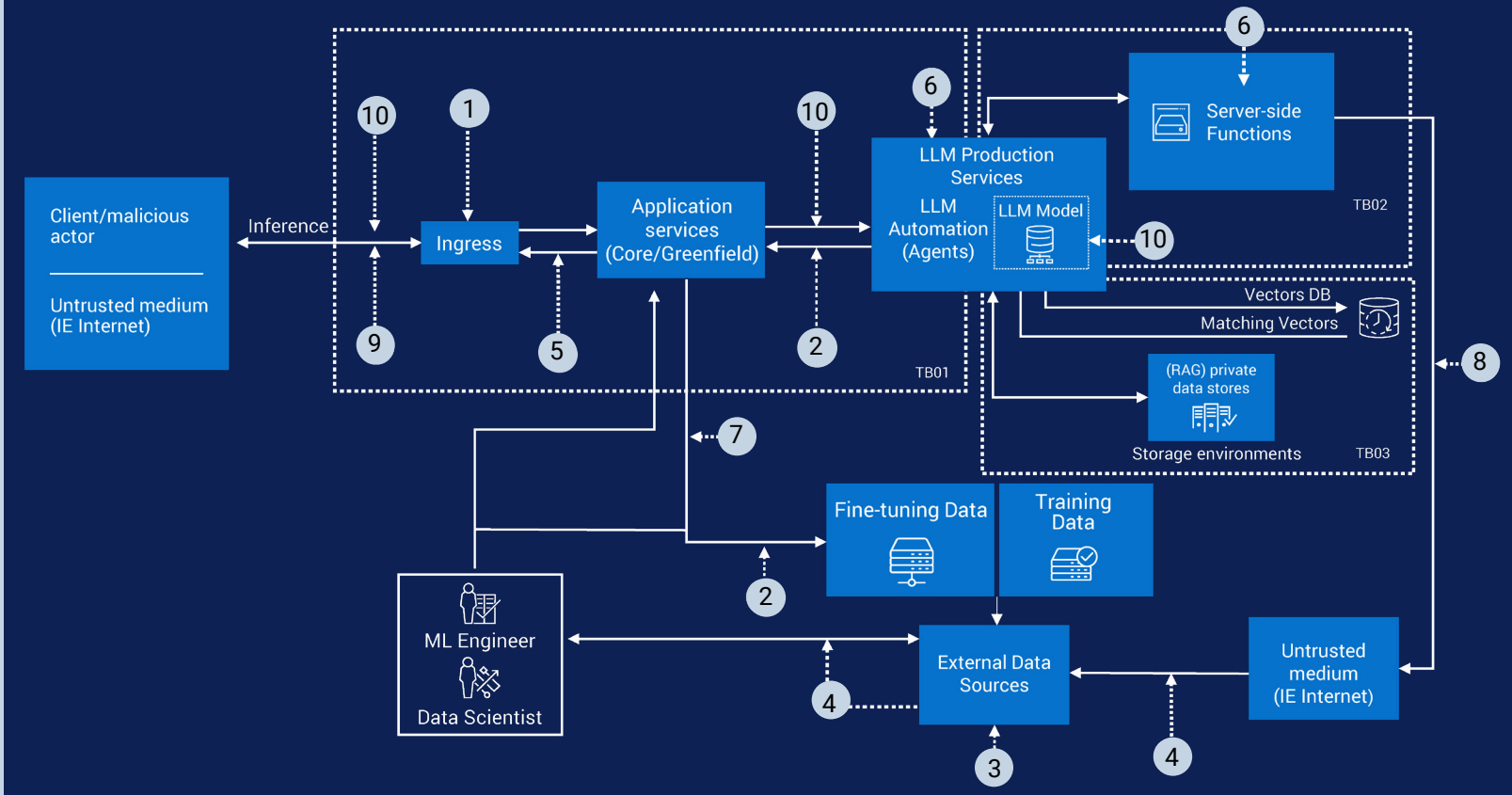
또한 기밀 정보 공개 방지, 안전하지 않은 주제 또는 편향성의 완화, 규정 준수 보장 등 AI 모델 입력과 관련된 문제를 관리하는 데 중점을 둡니다. 출력 측면에서는 모델에 대한 과도한 의존성 및 규정 준수 관련 위험 등의 문제 해결을 지원합니다.

Dell은 기업이 기존 사이버 보안 솔루션을 활용하거나 시스템을 보호하기 위한 새로운 툴과 관행을 파악함으로써 이러한 위험을 완화할 수 있도록 지원합니다. 보안이 혁신에 방해가 되지 않도록 하는 것을 목표로 합니다. AI 워크로드의 작동 방식과 직면한 보안 위협을 이해함으로써, 강화된 보안 태세를 구축하여 환경의 회복탄력성을 높이는 동시에 안심하고 혁신할 수 있도록 합니다. Dell은 전문 지식을 바탕으로 강력한 보안을 유지하면서 AI의 잠재력을 확실하게 활용할 수 있도록 지원합니다.

# GenAI 및 LLM 관련 사이버 보안 문제 상위 10개

OWASP에서 설명하는 바와 같이 GenAI/LLM 모델 보호에 있어 주된 문제는 다음과 같습니다.
**각 문제를 클릭하여 자세히 알아보십시오.**

1. 프롬프트 주입
2. 기밀 정보 공개
3. 공급망
4. 모델 데이터 오염
5. 부적절한 출력 처리
6. 과도한 에이전시
7. 시스템 프롬프트 유출
8. 벡터 및 임베딩 약점
9. 잘못된 정보
10. 무제한 소비

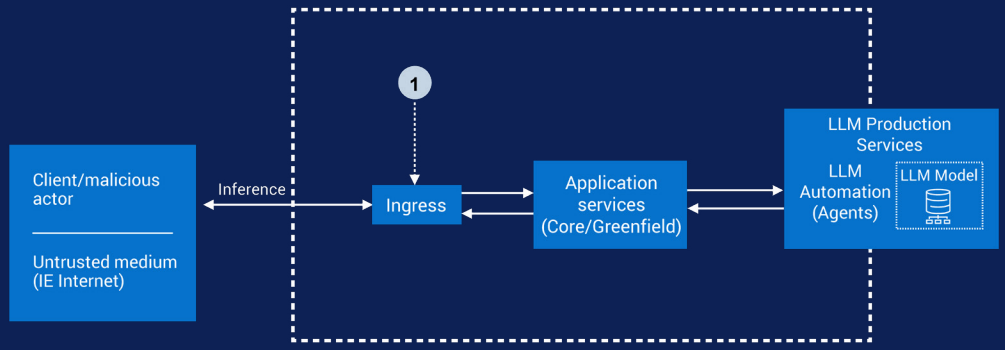


# 문제 1: 프롬프트 주입

**프롬프트 주입 완화 전략:**

- **데이터 완전 삭제 및 입력 검증:** 사용자 입력을 철저하게 차단하여 유해한 콘텐츠를 제거합니다. 정규화 및 인코딩을 사용하여 오용을 방지합니다.
- **NLP(Natural Language Processing) 및 머신 러닝 기반 접근 방식:** NLP 및 머신 러닝을 사용하여 조작되거나 악의적인 프롬프트를 탐지하고 차단합니다.
- **명확한 출력 형식 및 응답 제어:** 출력이 의도한 형식을 따르도록 하고 승인되지 않은 작업을 방지하기 위해 엄격한 응답 경계를 설정합니다. 프롬프트 필터링 및 응답 검증을 사용하여 무결성을 유지합니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **보안 프롬프트 엔지니어링:** 입력 처리를 보호하기 위한 전체 소프트웨어 보안의 일환으로 보안 프롬프트 설계 및 분석을 활용합니다.
- **모델 검증:** ML 모델을 정기적으로 검증하여 배포 전에 변조되지 않았는지 확인하고, 이를 통해 정확성과 무결성을 보장합니다.
- **프롬프트 필터링, 순위 지정 및 응답 검증:** 안전한 입력만 처리되도록 프롬프트를 분석하고 순위를 지정합니다. 응답을 검증하여 오용을 방지합니다.
- **견고성 검사:** 정기적으로 평가를 실시하여 취약성을 식별 및 수정하여 AI의 안전성과 신뢰성을 유지합니다.

아키텍처 다이어그램으로 돌아가기



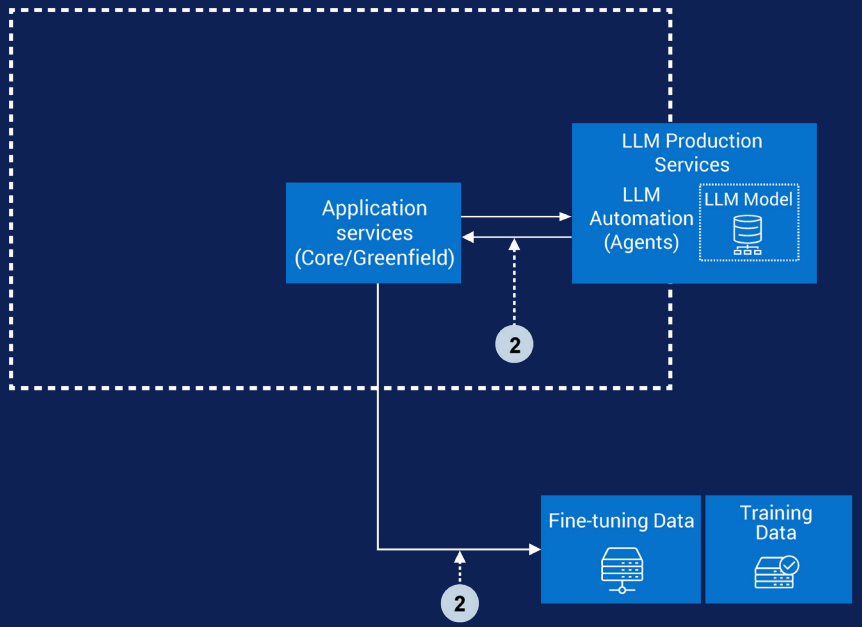

프롬프트 주입은 GenAI(Generative AI) 분야의 새로운 과제로, 악의적인 입력을 제작하여 모델의 동작을 조작하거나 무결성을 손상시키려는 것을 목적으로 합니다. 이러한 공격은 AI 시스템이 사용자 입력을 처리하고 대응하는 방식의 취약성을 악용하여 승인되지 않은 작업, 잘못된 정보 또는 기밀 데이터의 노출을 초래할 수 있습니다. GenAI가 중요 비즈니스 워크플로에 점점 더 많이 통합되는 가운데 이러한 위험을 해결하는 것이 신뢰와 보안을 유지하는 데 필수적입니다.

# 문제 2: 기밀 정보 공개

**기밀 정보 공개 완화 전략:**

- **데이터 완전 삭제 및 입력 검증:** 사용자 입력을 철저하게 차단하여 유해한 콘텐츠를 제거합니다. 정규화 및 인코딩을 사용하여 오용을 방지합니다.
- **동형 암호화 활용:** 콘텐츠 노출 없이 기밀 데이터를 안전하게 처리합니다. 이를 통해 사용 중인 동안에도 데이터가 암호화된 상태로 유지되어 침해로부터 보호됩니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- AI 데이터 상호작용에 **안전한 API 및 시스템 인터페이스를 활용**하고, 구성을 정기적으로 검토하여 노출 및 공격 노출 지점을 최소화합니다.
- **데이터 수집, 스토리지 및 정책을 보호**하고 포괄적인 데이터 보호 및 거버넌스 정책을 적용하여 규정 준수를 보장하고 데이터 위험을 최소화합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM 과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **안전한 개발, 구성 및 감사:** 보안 코딩 관행을 적용하고, 자동화된 구성 관리 툴을 사용하고, 정기적인 검토, 감사 및 업데이트를 실시하여 AI 시스템 구성을 안전하면서도 최신 상태로 유지합니다.
- **사용자 교육 및 보안 인식:** 사용자와 관리자에게 지속적인 AI 관련 보안 인식 교육을 제공하여 안전하지 않은 사용과 우발적인 데이터 유출을 줄입니다.

아키텍처 다이어그램으로 돌아가기



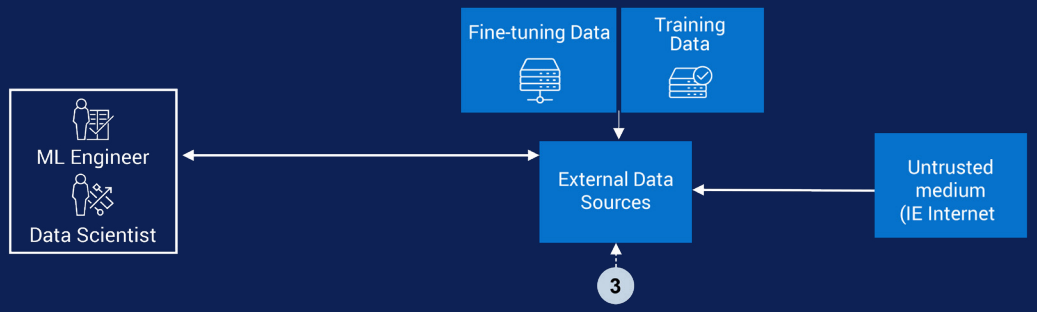

GenAI는 놀라운 발전을 가져왔지만, 특히 기밀 정보의 의도하지 않은 노출과 같은 심각한 위험도 발생합니다. PII(개인 식별 정보) 또는 독점 비즈니스 데이터 등 GenAI 툴을 잘못 사용하거나 취급하면 데이터 유출, 규정 위반 또는 평판 손상으로 이어질 수 있습니다. 따라서 조직은 이러한 위험을 이해하고 사전 예방적으로 해결하여 AI 시스템의 안전한 구현 및 사용을 보장하는 것이 중요합니다.

# 문제 3: 공급망 취약성

**공급망 취약성 완화 전략:**

- **공급업체 조사 및 안전한 공급망 관행으로 규정 준수:** 공급업체를 평가하고 공급망 보안을 우선시하는 계약을 수립합니다.
- **소프트웨어 BOM(Bill of Materials) 구현:** 소프트웨어 구성 요소의 출처를 추적 및 확인하여 투명성을 보장하고 코드 손상의 위험을 줄입니다.
- **모델 검증:** ML 모델을 정기적으로 검증하여 배포 전에 변조되지 않았는지 확인하고, 이를 통해 정확성과 무결성을 보장합니다.
- **최소 권한으로 컨테이너와 포드 실행:** 침해가 발생할 경우 잠재적 영향을 줄이고 무단 액세스를 제한합니다.
- **방화벽 구축:** 불필요한 네트워크 연결을 차단하여 잠재적인 위협에 대한 노출을 줄이고 공격자의 경로를 제한합니다.
- **데이터 및 주석 보호:** 데이터 및 관련 주석을 보호하여 중요 정보의 변조, 무단 액세스 및 손상을 방지합니다.
- **하드웨어 보안:** 보안 검증된 하드웨어를 사용하여 하드웨어 기반 공격으로 인한 취약성을 방지하고, 이는 인프라스트럭처의 강력한 기반이 됩니다.
- **ML 소프트웨어 구성 요소 보안:** 신뢰할 수 있고 심사를 받은 ML 소프트웨어 구성 요소를 사용하여 취약성을 줄이고 머신 러닝 워크플로의 전반적인 보안을 강화합니다.
- **안전한 개발, 구성 및 감사:** 보안 코딩 관행을 적용하고, 자동화된 구성 관리 툴을 사용하고, 정기적인 검토, 감사 및 업데이트를 실시하여 AI 시스템 구성을 안전하면서도 최신 상태로 유지합니다.

아키텍처 다이어그램으로 돌아가기



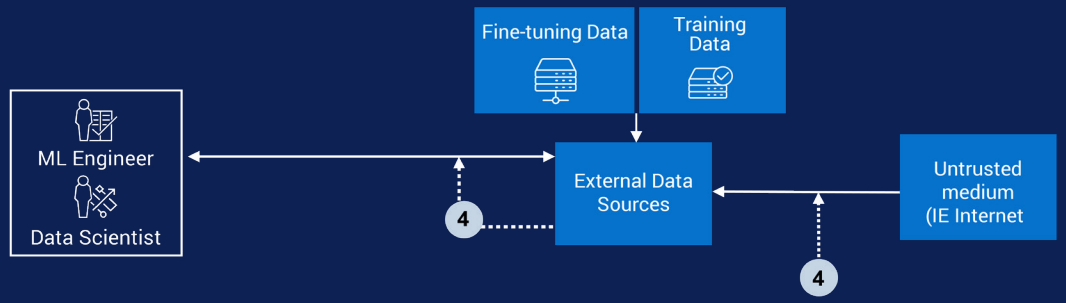

LLM 공급망의 취약성을 탐색하여 사전 훈련된 모델 무결성 및 타사 어댑터와 같은 중요한 구성 요소에 영향을 미칠 수 있는 문제를 확인합니다. AI 시스템은 배포 이전에 장시간 동안 해킹될 가능성이 있는 하드웨어와 소프트웨어에 의존합니다. 공격자는 머신 러닝 공급망의 다양한 단계에서 GPU 하드웨어, 데이터 및 해당 주석, ML 소프트웨어 스택의 요소 또는 모델 자체를 약점으로 악용할 수 있습니다. 이렇게 고유한 부분을 침해함으로써 공격자는 조기에 시스템 액세스 권한을 얻을 수 있으며, 이는 보안과 무결성에 심각한 위험을 초래할 수 있습니다. 이러한 취약성을 이해하고 완화하는 것은 강력하고 안전한 AI 솔루션 구축에 있어 매우 중요합니다.

# 문제 4: 모델 데이터 오염

**모델 데이터 오염 완화 전략:**

- **훈련 중 이상 징후 탐지 및 데이터 검증 활용:** 데이터의 불일치를 식별 및 해결하고, 모델 훈련에 깨끗한 고품질의 데이터만 사용될 수 있도록 합니다.
- **미세 조정 단계에서 환경 격리:** 중요한 개발 단계에서 모델의 무단 액세스 또는 오염을 방지합니다.
- **모델 검증:** ML 모델을 정기적으로 검증하여 배포 전에 변조되지 않았는지 확인하고, 이를 통해 정확성과 무결성을 보장합니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **데이터 완전 삭제 및 입력 검증:** 사용자 입력을 철저하게 차단하여 유해한 콘텐츠를 제거합니다. 정규화 및 인코딩을 사용하여 오용을 방지합니다.
- **안전한 개발, 구성 및 감사:** 보안 코딩 관행을 적용하고, 자동화된 구성 관리 툴을 사용하고, 정기적인 검토, 감사 및 업데이트를 실시하여 AI 시스템 구성을 안전하면서도 최신 상태로 유지합니다.
- **견고성 검사:** 정기적으로 평가를 실시하여 취약성을 식별 및 수정하여 AI의 안전성과 신뢰성을 유지합니다.
- **네트워크 세분화 구현:** 안전하지 않은 인터페이스 및 중요한 시스템 구성 요소에 대한 액세스를 제한합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.

아키텍처 다이어그램으로 돌아가기




모델 데이터 오염은 AI 수명주기에서 발생하는 보안 위협으로, 공격자가 의도적으로 훈련 데이터에 손상되었거나, 오해의 소지가 있거나, 악의적인 입력을 주입하여 오염시키는 행위입니다. 이 위험은 원시 데이터 수집 및 주석부터 머신 러닝 또는 대규모 언어 모델에 사용되는 데이터 세트의 큐레이션 및 통합에 이르기까지 중요한 구성 요소에 영향을 미칠 수 있습니다. AI 시스템의 신뢰성은 데이터 소스의 무결성에 달려있는데, 데이터 소스는 훈련 전, 전처리 중 또는 외부 데이터 파이프라인을 통해 조작될 수 있습니다.

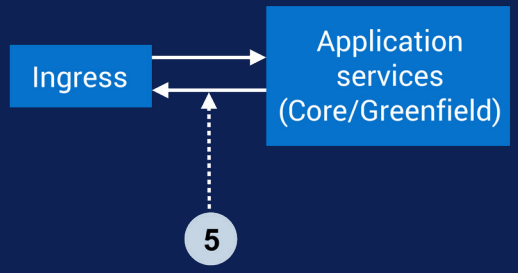
공격자는 데이터 오염을 활용하여 모델 정확도를 저하시키거나 취약성을 유발하거나 유해한 출력을 트리거합니다. 공격자는 데이터 출처, 주석 품질 또는 데이터 세트 수집 프로세스의 약점을 표적으로 삼아 보안, 신뢰성 및 회복탄력성을 저해할 수 있습니다. 강력하고 신뢰할 수 있는 AI 솔루션을 구축하려면 이러한 데이터 기반 위협을 인식하고 완화해야 합니다.

# 문제 5: 부적절한 출력 처리

**부적절한 출력 처리 완화 전략:**

- **컨텍스트 인식 출력 인코딩:** 주입 공격과 같은 취약성을 방지하기 위해 출력이 사용될 특정 컨텍스트(예: HTML, SQL 또는 API 환경)에 맞춰진 인코딩 및 이스케이프 기술을 항상 적용합니다.
- **출력 완전 삭제:** 안전한 다운스트림 사용을 보장하고 보안 위험을 완화하기 위해 OWASP(Open Web Application Security Project) ASVS(Application Security Verification Standard) 가이드라인에 따라 모델 출력에 대한 엄격한 유효성 검사 및 완전 삭제 관행을 준수합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **자동 출력 보안 테스트:** 자동화 툴을 사용하여 정기적인 보안 테스트를 실시하여 XSS(Cross-Site Scripting) 또는 주입 취약성과 같은 출력의 위험 요소를 식별하고 사전에 해결합니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **휴먼 인 더 루프 검토:** 금융 또는 의료 등 고위험 분야의 경우 정확성, 보안 및 안전을 보장하기 위해 모델 출력에 대한 인적 감독 및 검토가 필요합니다.
- **개인 정보 보호 및 규정 준수:** 개인 정보 보호 기술을 출력 프로세스에 내장하고 기밀 정보의 안전한 사용과 관련된 규정 및 표준 준수를 확인합니다.

아키텍처 다이어그램으로 돌아가기



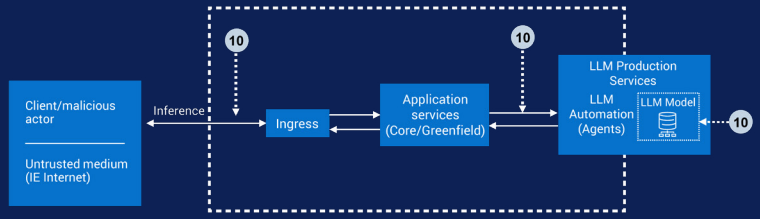

AI 모델 출력에 대한 검증이나 완전 삭제가 충분하지 않으면 권한 에스컬레이션 및 데이터 침해를 비롯한 심각한 보안 위험이 발생할 수 있습니다. AI 모델이 제대로 확인되거나 필터링되지 않은 출력을 생성할 경우, 악의적인 공격자가 이러한 취약성을 악용하여 무단으로 액세스하거나 시스템 내에서 권한을 에스컬레이션할 수 있습니다. 감독이 부족하면 데이터 손상, 무단 조치 및 심각한 보안 침해가 발생할 수 있습니다. 따라서 AI에서 생성된 출력에 대해 강력한 검증 및 완전 삭제 프로세스를 구현하는 것이 중요합니다.

# 문제 6: 과도한 에이전시

### 과도한 에이전시 완화 전략

- **최소 권한 적용:** 의도한 작업을 수행하는 데 필요한 최소 권한만 LLM 및 에이전틱 하위 시스템에 부여하고 액세스 제어를 정기적으로 검토합니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **운영 경계 설정:** LLM/에이전트가 액세스하거나 실행할 수 있는 항목을 명확하게 정의합니다.
- **휴먼 인 더 루프 검토:** 금융 또는 의료 등 고위험 분야의 경우 정확성, 보안 및 안전을 보장하기 위해 모델 출력에 대한 인적 감독 및 검토가 필요합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM 과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **자율성 제한:** LLM 기능을 제한하여 무제한 액세스 또는 제어를 방지합니다.
- **안전한 개발, 구성 및 감사:** 보안 코딩 관행을 적용하고, 자동화된 구성 관리 툴을 사용하고, 정기적인 검토, 감사 및 업데이트를 실시하여 AI 시스템 구성을 안전하면서도 최신 상태로 유지합니다.
- **방화벽 구축:** 불필요한 네트워크 연결을 차단하여 잠재적인 위협에 대한 노출을 줄이고 공격자의 경로를 제한합니다.
- **견고성 검사:** 정기적으로 평가를 실시하여 취약성을 식별 및 수정하여 AI의 안전성과 신뢰성을 유지합니다.

아키텍처 다이어그램으로 돌아가기



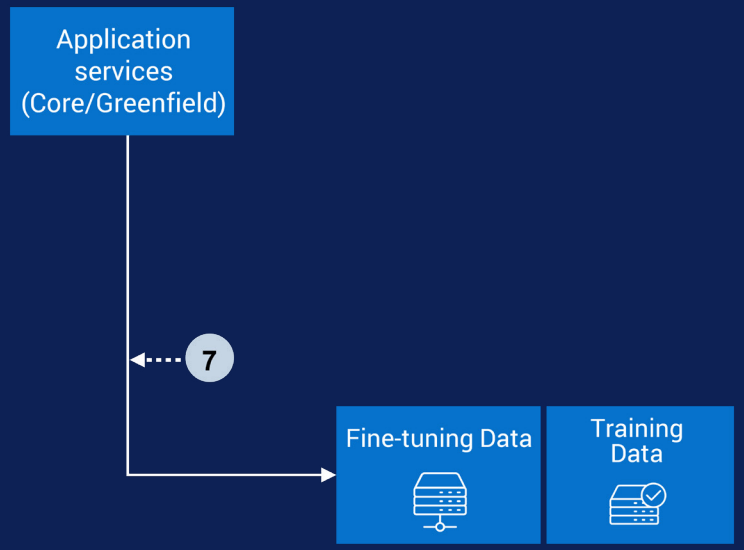

워크플로 내에서 AI 에이전트 또는 플러그인에 과도한 자율성이나 불필요한 기능을 부여하면 상당한 위험이 발생할 수 있습니다. AI 시스템에 필요 이상의 권한이나 기능이 부여되면 의도하지 않은 결과가 발생할 가능성이 높아집니다. 이는 LLM(대규모 언어 모델) 기반 시스템이 과도한 사용 권한으로 설계되어, 해서는 안 될 작업을 수행하거나 액세스해서는 안 될 정보에 액세스할 수 있도록 설계될 경우 발생할 수 있습니다. 이러한 권한 남용은 오류, 데이터 오용 또는 보안 취약성으로 이어질 수 있어, 안전하고 책임감 있는 사용이 보장되려면 AI 기능을 신중하게 제한하고 모니터링하는 것이 중요합니다.

# 문제 7: 프롬프트 유출

### 프롬프트 유출 완화 전략

- **프롬프트에 민감한 정보 제외:** 프롬프트에 자격 증명, API 키 또는 독점 로직을 포함하지 않고, 시스템 외부에서 안전하게 관리합니다.
- **보안 제어와 프롬프트 분리:** 인증, 권한 부여 및 세션 관리는 프롬프트가 아닌 애플리케이션 로직에서 처리합니다.
- **입력 및 출력 검증:** 강력한 검증을 통해 프롬프트 및 응답을 완전 삭제하여 의심스러운 패턴이나 조작을 차단합니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **암호화 및 보안 프롬프트:** 프롬프트와 구성을 암호화된 보안 스토리지에 저장하여 무단 액세스를 방지합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM 과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **프롬프트 정기 검토:** 주기적으로 프롬프트를 검토 및 완전 삭제하여 민감한 데이터를 제거하고 보안 규정 준수를 보장합니다.
- **약점에 대한 레드 팀 테스트 실시:** 가상의 적을 구성하고 테스트를 실시하여 프롬프트 관리 또는 결과의 취약성을 식별 및 수정합니다.
- **사용자 입력에서 프롬프트 격리:** 사용자 쿼리가 프롬프트를 조작하거나 노출시키는 것을 방지하는 시스템을 설계합니다.
- **사용량 제한 적용:** API 사용량을 제한하고, 의심스러운 활동을 제한하며, 자동화된 프롬프트 공격을 차단합니다.

아키텍처 다이어그램으로 돌아가기



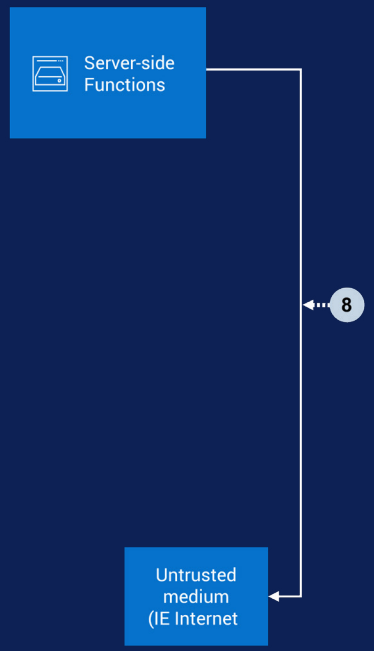

LLM(대규모 언어 모델) 또는 AI 시스템을 대상으로 하는 시스템 프롬프트 유출 공격은 공격자가 모델의 동작을 안내하고 운영 경계를 설정하는 숨겨진 명령(시스템 프롬프트)을 추출 또는 추론할 수 있을 때 발생합니다. 이러한 프롬프트는 핵심 규칙, 제한 사항 및 때로는 민감한 운영 로직을 포함하므로 대체로 최종 사용자에게는 표시되지 않습니다. 공격자는 특수하게 조작된 입력을 사용하거나 취약점을 악용하여 LLM이 시스템 프롬프트의 전체 또는 일부를 공개할 수 있도록 유도합니다. 유출될 경우 제한 사항을 리버스 엔지니어링하거나, 안전 필터를 우회하거나, 새로운 표적 공격을 개발하는 데 사용될 수 있으며, 최종적으로 프롬프트 주입, 권한 에스컬레이션, 그리고 무결성에 의존하는 하위 시스템의 오용 위험을 높입니다.

# 문제 8: 벡터 및 임베딩 약점

### 벡터 및 임베딩 약점 완화 전략

- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **암호화:**AES와 같은 강력한 암호화 표준을 사용하여 전송 중인 벡터 데이터와 저장된 벡터 데이터를 보호합니다.
- **구성 및 모니터링 보안:** 시스템을 강화하고, 안전하게 구성하고, 잘못된 구성, 무단 액세스 또는 이상 징후를 지속적으로 모니터링합니다.
- **취약성 관리:** 보안 위험 해결을 위해 모든 소프트웨어, 종속성 및 벡터 저장소 엔진을 정기적으로 업데이트하고 패치를 적용합니다.
- **데이터 완전 삭제 및 입력 검증:** 사용자 입력을 철저하게 차단하여 유해한 콘텐츠를 제거합니다. 정규화 및 인코딩을 사용하여 오용을 방지합니다.
- AI 데이터 상호작용에 **안전한 API 및 시스템 인터페이스를 활용**하고, 구성을 정기적으로 검토하여 노출 및 공격 노출 지점을 최소화합니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM 과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **하드웨어 보안:** 보안 검증된 하드웨어를 사용하여 하드웨어 기반 공격으로 인한 취약성을 방지하고, 이는 인프라스트럭처의 강력한 기반이 됩니다.
- **안전한 개발, 구성 및 감사:** 보안 코딩 관행을 적용하고, 자동화된 구성 관리 툴을 사용하고, 정기적인 검토, 감사 및 업데이트를 실시하여 AI 시스템 구성을 안전하면서도 최신 상태로 유지합니다.

아키텍처 다이어그램으로 돌아가기



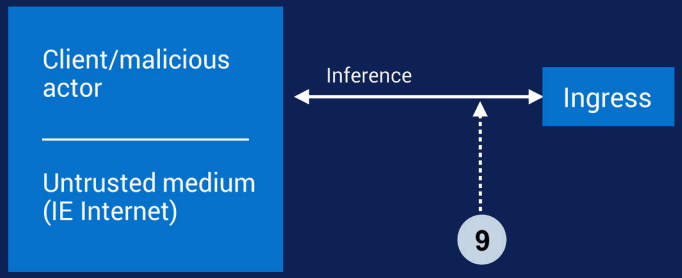

LLM(대규모 언어 모델) 또는 AI 시스템, 특히 RAG(Retrieval Augmented Generation)를 사용하는 시스템에 대한 벡터 및 임베딩 약점 공격은 정보가 숫자 벡터 및 임베딩으로 인코딩, 저장 및 검색되는 방식의 취약성을 표적으로 합니다. 이러한 메커니즘의 약점은 임베딩 인버전(임베딩으로부터 민감한 데이터 재구성), 데이터 오염(모델 동작을 조작하기 위해 유해하거나 편향된 콘텐츠 주입), 벡터 데이터베이스에 대한 무단 액세스(데이터 유출 초래) 또는 검색 결과 조작과 같은 악의적인 행위를 통해 악용될 수 있습니다. 공격자가 기밀 정보를 공개하거나 출력을 변경하거나 AI 기반 애플리케이션에서 사용자의 신뢰를 손상시키면서 개인 정보 보호, 무결성 및 신뢰성을 위협합니다. 진화하는 위협을 막으려면 적절한 액세스 제어, 데이터 검증, 암호화 및 지속적인 모니터링이 중요합니다.

# 문제 9: 잘못된 정보

### 잘못된 정보 완화 전략

- **소스를 신뢰할 수 있는 RAG(Retrieval-Augmented Generation):** RAG를 사용하여 검증되고 신뢰할 수 있는 데이터베이스 및 지식 저장소에서 정보를 검색하고 통합하여 환각 현상을 줄입니다.
- **모델 조정 및 출력 보정:** 다양한 데이터 세트로 모델을 정밀 조정하고 편향과 잘못된 정보를 최소화하기 위한 기술을 적용합니다.
- **자동 팩트 체크:** 신뢰할 수 있는 소스를 사용하여 출력을 상호 검증하고 잘못된 정보에 자동으로 플래그를 지정합니다.
- **불확실성 모니터링:** 중요한 사례의 인적 검토를 위해 신뢰도가 낮은 응답에 플래그를 지정합니다.
- **휴먼 인 더 루프 검토:** 금융 또는 의료 등 고위험 분야의 경우 정확성, 보안 및 안전을 보장하기 위해 모델 출력에 대한 인적 감독 및 검토가 필요합니다.
- **사용자 피드백:** 사용자의 오류 보고를 지원하여 지속적으로 모델을 개선하고 잘못된 정보 경로를 신속하게 수정합니다.
- **액세스 제한 및 인적 감독:** RBAC(역할 기반 액세스 제어), MFA(다단계 인증) 및 ID 관리를 적용하여 액세스를 제한합니다. 중요한 의사 결정에 인적 검토를 활용합니다.
- **안전한 개발, 구성 및 감사:** 보안 코딩 관행을 적용하고, 자동화된 구성 관리 툴을 사용하고, 정기적인 검토, 감사 및 업데이트를 실시하여 AI 시스템 구성을 안전하면서도 최신 상태로 유지합니다.
- **위험 커뮤니케이션:** 사용자에게 AI 제한 사항에 대해 교육하고 독립적인 검증을 장려합니다.
- **의도적 UI 및 API 설계:** AI 생성 콘텐츠임을 강조하고 사용자에게 책임 있는 사용을 안내합니다.

아키텍처 다이어그램으로 돌아가기



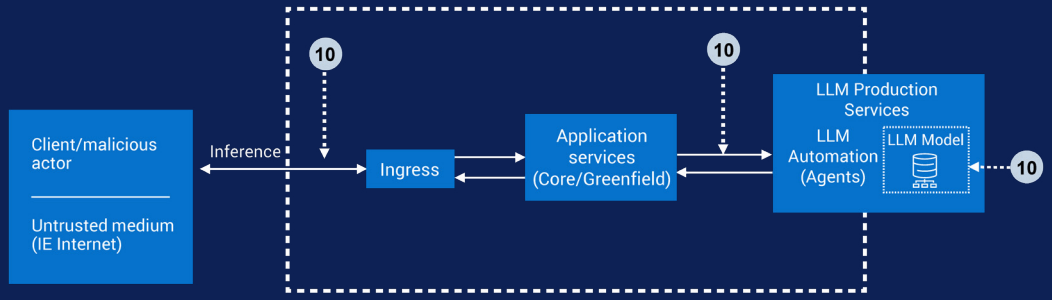

LLM 또는 AI 시스템에 대한 잘못된 정보 공격은, 모델이 출력을 통해 잘못된 정보, 오해의 소지가 있는 정보 또는 겉보기에는 신뢰할 수 있지만 잘못된 정보를 생성하거나 확산시키도록 의도적으로 노력하는 행위입니다. 이 취약성은 모델의 '환각(hallucination)' 경향(조작되었지만 그럴듯하게 들리는 콘텐츠 생성), 훈련 데이터에 존재하는 편향성 또는 격차, 적대적 프롬프트의 영향 등 여러 요인으로 인해 발생합니다. 환각은 LLM이 사실을 진정으로 이해하기보다는 패턴에 맞는 텍스트를 통계적으로 생성하여, 신뢰할 수 있는 것처럼 보이지만 사실은 근거가 없는 답변으로 이어지기 때문에 발생합니다. 이러한 공격을 인한 위험으로는 보안 침해, 평판 감소, 법적 책임 등이 있습니다. 특히 사용자가 정확성이나 타당성을 검증하지 않고 LLM 응답에 과도하게 의존하여 중요한 의사 결정 및 프로세스에 오류 또는 잘못된 정보를 포함시킬 수 있는 환경에서는 더욱 심각합니다.

# 문제 10: 무제한 소비

### 무제한 소비 완화 전략

- **사용량 제한 및 사용자 할당량 적용:** 사용자, API 키 또는 앱당 요청, 토큰 또는 데이터에 엄격한 제한을 설정하여 남용을 방지합니다.
- **인증 및 사용자 세분화 요구:** 강력한 인증(예: API 키, OAuth)을 사용하고 역할 또는 계층을 할당하여 승인된 요청만 처리합니다.
- **입력 유효성 검사 및 크기 제한:** 프롬프트 크기와 구조를 검증하여 너무 크거나 잘못된 형식의 쿼리를 차단 또는 트리밍합니다.
- **처리 시간 초과 및 리소스 제한 적용:** 각 요청에 대한 시간 초과 및 리소스 한도를 설정하여 작업의 장시간 실행 또는 리소스 소모를 방지합니다.
- **스마트 캐싱 배포 및 중복 제거:** 중복 쿼리 또는 유사한 쿼리에 대한 캐시 응답의 중복 제거로 불필요한 처리를 줄입니다.
- **모니터링, 로깅 및 이상 징후 탐지:** MDR/XDR/SIEM과 같은 솔루션을 사용하여 AI 시스템 활동을 지속적으로 모니터링하고 기록하여 무단 액세스, 이상 징후 또는 데이터 유출을 신속하게 탐지, 조사 및 대응합니다.
- **예산 추적 및 지출 제어:** 대시보드와 알림을 사용하여 예산 임계값을 두고 비용을 모니터링한 다음 사용량을 차단합니다.
- **샌드박싱과 격리 기술:** 권한이 제한된 격리된 환경에서 워크로드를 실행하여 위험을 줄입니다.
- **호출 깊이 및 대화 전환 제한:** 재귀 호출 또는 대화 단계에 제한을 적용하여 악용을 방지합니다.
- **계층형 모델 또는 리소스 할당 적용:** 우선순위가 높은 요청은 프리미엄 모델로, 우선순위가 낮은 트래픽은 비용 효율적인 모델로 라우팅합니다.

아키텍처 다이어그램으로 돌아가기




LLM 또는 AI 시스템에 대한 무제한 소비 위협은 애플리케이션이 효과적인 속도 제한, 인증, 사용 제한 없이, (악의적인지 여부와 무관하게) 사용자가 과도하고 통제되지 않은 추론 요청 또는 프롬프트를 제출할 수 있도록 허용하는 보안 취약성을 의미합니다. LLM 추론은 컴퓨팅 비용이 많이 들기 때문에 이러한 통제 불능 상태는 여러 가지 방법으로 악용될 수 있습니다. 공격자는 시스템 리소스를 과부하시켜 DoS(Denial of Service)를 유발하거나, 사용량 기반 지불 또는 클라우드 호스팅 배포에서 예상치 못한 경제적 손실을 발생시키거나, 모델을 체계적으로 쿼리하여 동작을 복제하고 지적 재산을 탈취할 수 있습니다. 이로 인해 서비스 중단, 다른 사용자의 성능 저하, 재정적 부담, 민감한 모델 유출 위험 증가가 발생할 수 있습니다. 기본적으로 리소스 사용량이 제대로 관리되지 않을 때 무제한의 소비가 발생하므로 LLM 기반 애플리케이션은 우발적이거나 의도적인 악용에 노출됩니다.

# Dell의 AI 보안 솔루션을 선택해야 하는 이유

Dell은 하드웨어, 소프트웨어 및 매니지드 서비스를 포괄하는 접근 방식을 통해 조직이 AI 모델과 LLM을 보호할 수 있도록 지원합니다. 보안은 공급망에서 디바이스, 인프라스트럭처, 데이터, 애플리케이션에 이르는 전 영역에서 제로 트러스트 원칙에 따라 구축되어 있습니다. 포트폴리오 전반에서 Dell의 솔루션은 MFA, RBAC, 최소 권한, 지속적 검증과 같은 기능을 통해 사이버 보안을 강화하도록 설계되었습니다. 이 포괄적인 '보안을 고려한 설계' 접근 방식을 통해 조직은 AI 및 LLM으로 자신 있게 혁신할 수 있으며, 모델 도난, 데이터 유출, 적대적 공격 및 기타 지능형 사이버 위협으로 인한 위험을 최소화할 수 있습니다.

## 공급망

Dell의 안전한 공급망은 제품 개발, 제조 및 제공의 모든 단계에 보안을 내장하여 AI 모델 및 LLM의 기본적인 보호를 제공합니다. 암호화 방식으로 서명된 BIOS 및 펌웨어 업데이트, 보안 구성 요소 검증, AI 중심 SBOM(Software Bill-of-Materials), 데이터 세트 계보 추적, 통합 보안 소프트웨어 및 구성, 글로벌 표준에 부합하는 엄격한 공급업체 위험 평가를 통해 Dell은 변조, 무단 액세스 및 공급망 공격으로 인한 위험을 최소화합니다. 이를 통해 조직은 투명성, 무결성 및 규정 준수를 확보한 상태에서 신뢰할 수 있고 탄력적인 AI 워크로드를 배포할 수 있습니다.

## AI PC

Dell은 온디바이스 AI 워크로드에 기본적인 보안을 제공합니다. 탁월한 수준의 보안을 자랑하는 AI PC*인 Dell Trusted Device는 보안을 염두에 두고 설계되었습니다. 공급망 보안은 제품 취약성 및 변조의 위험을 줄입니다. 하드웨어 및 펌웨어에 직접 내장된 고유한 방어 기능으로 PC와 최종 사용자를 안전하게 보호합니다. Dell SafeBIOS는 심층적인 BIOS 수준 가시성 및 변조 탐지를 제공하며, Dell SafeID는 자격 증명 보안을 강화하고 비밀번호 없는 인증을 지원합니다. 파트너 소프트웨어는 엔드포인트, 네트워크 및 클라우드 환경 전반에 걸쳐 고급 보호 기능을 제공합니다.

## 사이버 회복탄력성

Dell의 PowerProtect 사이버 회복탄력성 솔루션은 암호화되고 변경 불가능한 백업, 신속한 복원 및 격리된 Cyber Recovery 볼트를 통해 AI 데이터를 보호합니다. 이러한 기능은 파괴를 방지하고, 악의적인 업데이트가 미치는 영향을 완화하며, 규정 준수와 공격 후 복구를 지원합니다.

## 서버

PowerEdge 서버는 기밀 컴퓨팅을 활용하여 AI/LLM 프롬프트 및 임베딩을 격리하고 보호하며, 공식적인 소스에 기반한 신뢰할 수 있는 RAG(Retrieval-Augmented Generation) 솔루션, MFA, RBAC, 칩 내장형 RoT(Root of Trust), 서명된 펌웨어 및 지속적인 모니터링을 통해 중요한 AI 워크로드를 보호합니다.

## 스토리지

Dell의 스토리지 포트폴리오는 민감한 AI 데이터를 안전하게 암호화하여 저장하고, 데이터가 저장 또는 전송 중일 때 강력한 AES-256 암호화를 제공합니다. 일부 제품에는 미래의 양자 위협에 대비하여 견고하게 설계된 고급 암호화가 제공됩니다. 이 포트폴리오에는 고속 NVMe 성능, 데이터(AI 워크로드에 활용되는 데이터 포함)를 보호하는 FIPS 준수 암호화 모듈, 불변 스냅샷, 랜섬웨어 공격에 대응하는 에어 갭 Cyber Recovery 볼트가 포함되어 있습니다. 제로 트러스트 아키텍처, 공급망 보안 및 변조 방지 감사 기능은 거버넌스를 강화합니다. 내장된 이상 징후 탐지 및 AIOps ML 모델은 훈련에 고객 데이터를 사용하지 않고도 워크로드를 보호하여, 입력 기반 공격 위험을 최소화할 수 있습니다.

## AIOps

Dell AIOps는 잘못된 구성, 취약성(CVE 포함)을 탐지하기 위한 자동화된 지속 모니터링을 제공하며 AI/LLM 워크로드에 영향을 미치는 공급망 위험 인식을 지원합니다. 실시간 CVE 스캔, 스마트 알림 및 AI 기반 대시보드는 이상 징후에 플래그를 지정하고 해결 워크플로를 추적함으로써 신속한 개입을 지원합니다. 내장된 규정 준수 기능, 역할 기반 액세스 제어 및 자동화 보고 기능은 워크로드 전반에서 안전한 운영을 유지하는 데 도움이 되고, 지원되는 솔루션의 Generative 기능을 비롯한 원활한 EDR/XDR 통합 및 AI 기반 운영 통찰력으로 IT 효율성을 높입니다.

## 네트워킹

Dell Networking 솔루션은 강력한 네트워크 세분화를 통해 AI/LLM 환경을 보호하여 내부 이동을 최소화합니다. 암호화된 네트워크 경로 및 통합 방화벽 제어는 AI 데이터에 대한 무단 액세스를 차단합니다.

## AI 보안 및 회복탄력성 서비스

Dell의 AI 보안 및 회복탄력성 서비스는 조직에 AI를 도입하는 것과 관련된 새로운 위험을 해결하도록 개발되었습니다. 팀과 협업하여 최대한 빠르게 AI를 온보딩하도록 설계된 Dell Technologies의 서비스는 전략 기획, 솔루션 구현, 매니지드 보안 서비스와 관련한 전문 지식을 제공함으로써 운영 부담을 줄이고 AI를 통해 안전하게 혁신하도록 지원합니다. 각각은 조직이 진화하는 AI 위험을 해결하고 안전한 AI 배포를 최적화할 수 있도록 맞춤형으로 제공됩니다.

## Dell AI Factory

Dell AI Factory는 Dell의 안전한 공급망, 최소 권한을 적용하는 제로 트러스트 기능, 모델을 안전하게 보호하도록 설계된 AI MDR 솔루션 등, 특별히 설계된 보안 통합 포트폴리오입니다.

*Dell 내부 분석 기준, 2024년 10월(인텔) 및 2025년 3월(AMD). 인텔 및 AMD 프로세서를 탑재한 PC에 해당됩니다. 모든 PC에서 모든 기능을 사용할 수 있는 것은 아닙니다. 일부 기능을 사용하려면 추가로 구매해야 합니다. Principled Technologies의 검증을 받은 인텔 기반 PC, 2025년 7월.

# 결론

회복탄력성이 뛰어난 AI 프레임워크를 구축하려면 조직과 보안 전문가 간의 협업 접근 방식이 반드시 필요합니다. AI와 LLM이 산업이 계속해서 재편함에 따라 데이터 보안, 모델 무결성, 규정 준수 문제 등 이로 인한 위험을 해결하는 것이 중요합니다. 조직은 AI 여정의 모든 단계에 보안을 내장하는 사전 예방적 전략을 우선시해야 합니다.

Dell Technologies는 이 임무에 있어 신뢰할 수 있는 파트너로서 고객의 고유한 요구 사항에 맞춘 포괄적인 GenAI 맞춤 구성, 보안 컨설팅 및 통합 솔루션을 제공합니다. 기업은 Dell의 강력한 사이버 보안 솔루션을 활용하여 AI 및 LLM 위험을 효과적으로 완화하는 동시에 기존 보안 투자의 잠재력을 극대화할 수 있습니다. Dell은 조직이 AI 인프라스트럭처를 보호할 수 있도록 고급 보안을 현재 프레임워크에 원활하게 통합하고, 미래 지향적이면서 안전한 환경을 보장합니다.

Dell의 포괄적인 AI 솔루션이 GenAI 및 LLM 환경을 보호하는 방법 알아보기: Dell.com/CyberSecurityMonth