

AI 보안 격차: 보안팀을 위한 제로 트러스트 구현 가이드

섀도 AI 탐지부터 에이전틱 런타임 보호에 이르기까지, 제로 트러스트 원칙을 적용하여 AI 전체 라이프사이클을 안전하게 보호하기 위한 실무자 가이드로, 30/60/90일 구현 로드맵 및 경영진 보고용 KPI를 담았습니다.

EBOOK

©2026 Zscaler, Inc. 모든 권리 보유.

소개

보안팀은 AI 문제가 있다는 것을 알고 있습니다. 대부분의 사람들이 아직 답하지 못하는 것은 그 규모가 얼마나 되는지입니다.

현재 여러분의 환경에서 실행 중인 인공지능(AI) 앱은 무엇인가요? 개발자들은 통합 개발 환경(IDE) 플러그인과 명령줄 인터페이스(CLI) 도구를 통해 어떤 모델을 조회하고 있나요? 프롬프트를 통해 어떤 데이터가 어떻게 전송되며, 그 중에 데이터의 소스 코드, 고객 개인 식별 정보(PII) 또는 규제상 공개 규칙의 적용을 받는 콘텐츠가 있나요?사용자를 대신하여 자율적으로 작동하는 에이전트는 무엇이며, 어떤 시스템에 접근할 수 있나요?

AI가 환경에서 실제로 수행하는 작업과 사용자가 확인하고 관리할 수 있는 작업 사이의 간극이 바로 진정한 위험이 존재하는 지점입니다. 이는 이론적인 공격 시나리오가 아니라, 보안 검토 없이 무분별하게 사용되는 AI 도구에서 비롯됩니다. 섀도 AI나 이미 승인된 서비스형 소프트웨어(SaaS) 플랫폼 내의 임베디드 AI 기능들이 바로 이러한 간극의 주요 원인입니다. IT 부서에서 특정 목적으로 승인한 도구에 AI 비서 기능이 기본적으로 탑재되어 있고, 이제 이 도구가 보안 검토 없이 내부 데이터를 처리하고 있는 경우가 그 예입니다.

이 가이드의 데이터와 통계는 *9,893*억 건의 분석을 기반으로 *Zscaler*가 *2026*년에 발표한 *ThreatLabz 2026 AI* 보안 보고서에서 발췌한 것으로, *2025*년 *1*월부터 *12*월까지 *Zscaler Zero Trust Exchange*에서 관찰된 *9,893*억 건의 *AI/ML* 트랜잭션 분석을 기반으로 합니다. 별도로 명시되지 않는 한 모든 수치는 해당 연구 기간을 기준으로 합니다.

©2026 Zscaler, Inc. 모든 권리 보유.

제로 트러스트는 원칙에 입각한 대응책을 제시합니다.

앱, 사용자, 메시지 또는 커넥터를 본질적으로 신뢰하지 않습니다. 모든 계층에서 신원 및 콘텍스트 기반 접근 통제를 시행합니다. 지속적으로 검증합니다. 문제가 발생했을 때 피해 범위를 최소화합니다. 관건은 순서입니다. 정책 시행에 앞서 가시성이 확보되어야 합니다. 자동화 기능을 추가하기 전에 접근 통제가 안정적이어야 합니다. 순서대로 구축하지 않으면 정상적인 작업을 차단하거나 실제 위험을 막지 못하는 제어 시스템이 만들어집니다.

83%

기업 부문의 전년 대비 성장률 AI/ML 트랜잭션 건수 — 이제 도입은 인프라 규모로 이루어지고 있습니다

18,033 테라바이트

2025년에 AI 앱을 통해 전송된 데이터 — 대부분의 프로그램이 추적할 수 있는 수준을 넘어섭니다

39%

능동적으로 차단된 AI 트랜잭션 — 거버넌스 압박이 이미 제기되고 있습니다

93%

AI 데이터 전송량의 전년 대비 증가율 — 보안 프로그램의 역량을 앞지르고 있습니다

출처: ThreatLabz 2026 AI 보안 보고서, Zscaler(pp. 5—6)

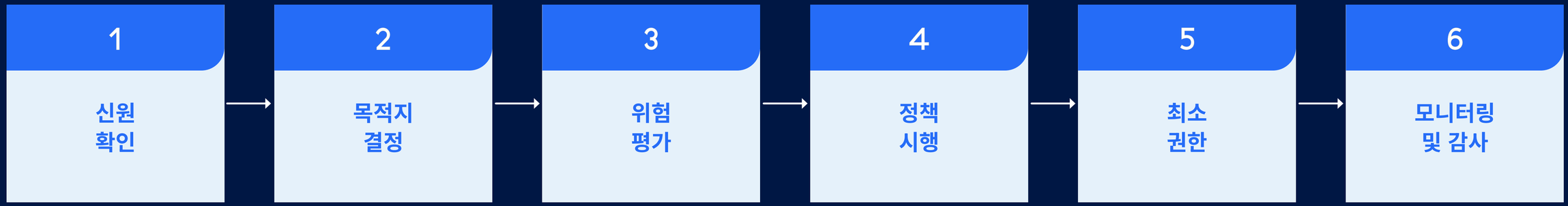
본 가이드에서 다루는 내용

1. AI 시스템, 에이전트 및 데이터 흐름에 특별히 적용되는 7가지 제로 트러스트 원칙
2. 의존성 순서에 따른 단계별 30/60/90일 구현 로드맵
3. 새롭게 부상하는 AI 위험 영역 분석: 섀도 AI, 프롬프트 인젝션, 에이전트 시스템, 모델 콘텍스트 프로토콜(MCP)/에이전트 간(A2A) 트래픽, 임베디드 SaaS AI 및 AI 인프라 보안 태세
4. AI 보안 프로그램 진행 상황을 경영진에게 보고하기 위한 KPI 모델

©2026 Zscaler, Inc. 모든 권리 보유.

AI에 제로 트러스트 원칙 적용

아래의 7가지 원칙은 구현 로드맵의 모든 사항에 대한 의사 결정 프레임워크를 구성합니다. 각각은 통제를 평가하기 위한 의사결정 도구입니다.이 순서는 보안팀이 실제로 적용하는 방식을 그대로 반영합니다.



순서대로 적용됩니다. 각 원칙은 이전 원칙을 기반으로 합니다.

©2026 Zscaler, Inc. 모든 권리 보유.

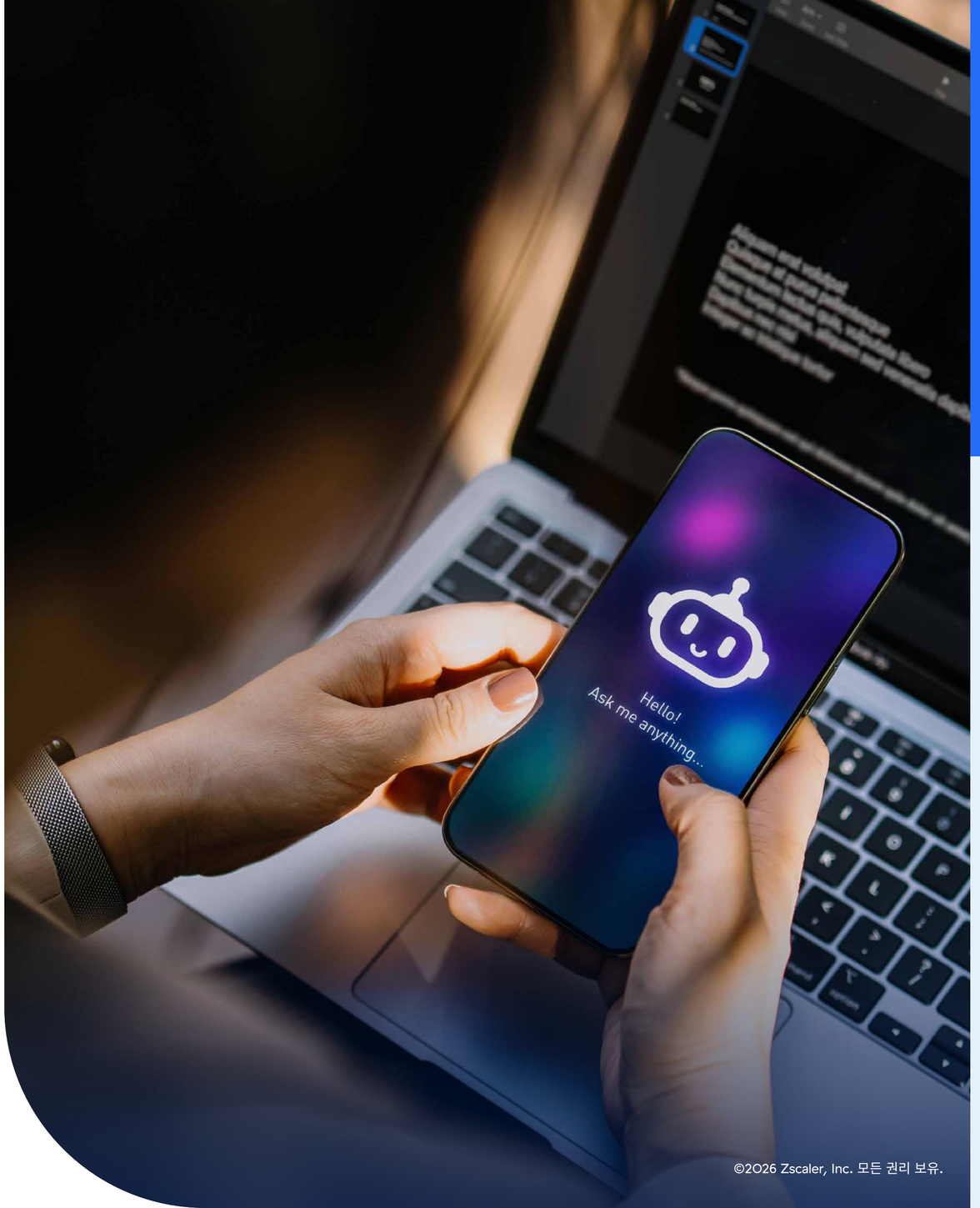
원칙 1:
사용자, 개발자 및 서비스의 신원 확인

AI 접근은 신원 확인에서 시작됩니다. 생성형 AI(GenAI) SaaS 도구에 액세스하는 모든 사용자, IDE 플러그인에서 모델 엔드포인트를 쿼리하는 모든 개발자, 그리고 AI 인프라 구성 요소에 연결하는 모든 서비스 계정은 일관된 ID 제어 평면을 통해 인증해야 합니다. 싱글 사인온(SSO)과 다단계 인증(MFA)은 기본 사항이지만, 세부 사항이 중요합니다.

임직원과 개발자 집단은 서로 다른 위험 프로필을 가지고 있으므로 각기 다른 정책적 접근 방식이 필요합니다. 콘텐츠 초안 작성을 위해 ChatGPT를 사용하는 마케팅 분석가는 내부 저장소에 접근 권한이 있는 MCP 서버에 연결된 GitHub Copilot 플러그인을 통해 코드 자동 완성 기능을 사용하는 소프트웨어 엔지니어와는 다른 데이터 노출 위험을 안고 있습니다. AI 에이전트를 위한 서비스 간 인증은 대부분의 조직이 아직 준비되지 않은 새로운 특수한 상황입니다.

주요 통제

- AI SaaS 및 AI 개발자 도구에 대한 SSO 및 MFA 시행
- 업무용 GenAI 사용과 개발자용 AI 도구(IDE, CLI, MCP 통합) 사용에 대해 별도의 ID 정책 적용
- **AI 에이전트를 위한 서비스 계정 제어:** 최소 권한 범위 지정, 단기 자격 증명 및 감사 로깅
- **ID 신호로서의 기기 보안 태세** 관리형 엔드포인트는 사용자 소유 기기(BYOD)와 다른 접근 권한을 갖습니다.



©2026 Zscaler, Inc. 모든 권리 보유.

원칙 2:
승인된 AI와 미승인 AI의 목적지 결정

모든 AI 앱이 동일한 것은 아니며, 모든 앱을 일괄적으로 차단하는 접근 방식은 운영상 현실적이지도 않고 보안상 효과적이지도 않습니다. **목표는 단계별 접근 통제 정책을 수립하는 것입니다.** 승인된 도구는 인라인 검사를 통해 완전한 접근 권한을 부여받고, 허용된 도구는 사용 가드레일을 준수하며 모니터링되는 접근 권한을 부여받고, 승인되지 않았거나 위험도가 높은 도구는 차단되거나 격리됩니다. 이러한 AI 접근 통제 대상 목록은 대부분의 조직이 예상하는 것보다 훨씬 많으며, 기존 플랫폼에 새로운 AI 기능이 추가되면서 단 1년 만에 4배나 증가했습니다.

주요 통제

- 앱 및 사용자 그룹별로 허용, 경고, 격리 또는 차단을 설정할 수 있는 계층형 목적지 정책
- 트래픽에 새롭게 등장하는 AI 도구를 실시간으로 찾아내는 동적 앱 검색 기능
- GenAI SaaS, 기업용 SaaS 내 임베디드 AI, 그리고 AI 개발자 도구에 대한 별도의 정책 트랙
- 새로운 AI 도구 요청에 대한 승인 워크플로를 제공하여 사용자가 제어 기능을 우회하는 대신 진행할 수 있는 경로 제시

3,400 이상

기업 트래픽 전반에 걸쳐 추적된 다양한 AI 애플리케이션이 단 1년 만에 4배 증가했으며, 매달 증가하고 있습니다

ThreatLabz 2026 AI 보안 보고서, *Zscaler*(5페이지)



©2026 Zscaler, Inc. 모든 권리 보유.

원칙 3:
상황에 기반한 의사결정을 통해 위험 평가

AI 접근 권한 결정은 절대로 이분법적이어야 하지 않습니다. 동일한 사용자가 동일한 AI 도구에 접근하더라도 기기, 데이터 민감도 및 작업 내용 등 상황에 따라 서로 다른 위험이 발생할 수 있습니다. 효과적인 위험 평가를 위해서는 소스 코드, PII, 지불 카드 산업(PCI) 데이터, 보호 대상 건강 정보(PHI), 미공개 재무 실적, 기밀 인수합병(M&A) 세부 정보 등 레드 데이터를 사전에 정의해야 합니다. 기업 트래픽에서 가장 많이 발견된 위반 유형을 보면 직원들이 제출하는 내용이 무엇인지 정확히 알 수 있습니다. 바로 이름 및 국가 식별 정보, 소스 코드, 그리고 의료 기록입니다.

주요 통제

- 소스 코드, PII/PCI/PHI, 시크릿, 미공개 계획을 위한 명시적인 '레드 데이터' 분류
- 사용자 그룹, 부서, 기기 보안 태세 및 데이터 민감도에 따른 위험도 점수 산정
- 정적인 허용/차단 규칙에 그치지 않고 콘텍스트 신호를 기반으로 한 동적 정책 조정
- 파일 업로드를 포함하는 AI 상호작용과 텍스트만 제공하는 상호작용에 대해 별도의 위험 처리 구분

4억 1천만 이상

CHATGPT와 관련된 데이터 손실 방지(DLP) 위반 사례 다수 발생: 주요 유형은 이름 유출, SSN, 소스 코드, 의료 데이터입니다 *ThreatLabz 2026 AI 보안 보고서, Zscaler (6, 15쪽)*

ThreatLabz 2026 AI 보안 보고서, Zscaler (pp. 6, 15)



©2026 Zscaler, Inc. 모든 권리 보유.

원칙 4:
세션별로 정책을 인라인으로 적용

로그를 보면 무슨 일이 일어났는지 알 수 있습니다. 개발자가 승인되지 않은 모델에 독점 소스 코드를 제출했다는 기록을 찾을 때쯤이면 데이터는 이미 사라진 후입니다. 대부분의 조직은 AI 도구에 파일을 업로드할 때 DLP를 적용하면서도 프롬프트 텍스트는 검사하지 않아 이러한 문제를 더욱 악화시킵니다. 개발자가 명령 프롬프트 창에 소스 코드를 직접 입력하면 경고 메시지가 전혀 표시되지 않고 해당 컨트롤이 모두 지워집니다. 데이터 전송량 증가는 보안이 허술한 채널을 통해 얼마나 많은 정보가 오가고 있는지를 보여줍니다.

주요 통제

- URL/카테고리 수준에 그치지 않는, 프롬프트, 응답, 파일 업로드에 대한 인라인 DLP
- 정책 시행 지점에서 허용, 경고, 수정, 차단 및 격리 조치 수행 가능
- 콘텍스트 인식 감지 기능을 갖춘 주요 GenAI 앱을 위한 신속한 분류
- 주제와 무관하거나, 제한적이거나, 경쟁에 민감한 상호작용에 대한 콘텐츠 관리
- 허용되지만 위험성이 있는 AI 도구에 대한 브라우저 격리: 완전한 신뢰 없이 접근 가능

93%

기업 AI 데이터 전송량의 전년 대비 증가율이 대부분의 보안 프로그램이 추적하거나 관리할 수 있는 수준을 앞지르고 있습니다(*ThreatLabz 2026 AI* 보안 보고서, *Zscaler, 5, 14*페이지)



©2026 Zscaler, Inc. 모든 권리 보유.

원칙 5:
접근 및 노출 최소화

AI 시스템은 기존의 신원 및 네트워크 모델에서 벗어난 접근 경로를 생성하는 경우가 많습니다. 검색 증강 생성(RAG) 파이프라인은 모델이 문서 코퍼스에 접근할 수 있도록 해줍니다. 에이전트는 내부 도구, 데이터 세트, 벡터 저장소 또는 MCP 서버에 연결할 수 있습니다. 그러한 권한은 표준 접근 권한 검토에 나타나지 않더라도 사용자의 접근 권한만큼이나 실재합니다.

접근 및 노출을 최소화한다는 것은 AI 시스템이 도달할 수 있는 범위를 줄이고 AI 구성 요소 간의 통신 방식을 제한하는 것을 의미합니다. 에이전트, 데이터 파이프라인, RAG 구성 요소 및 백엔드 데이터 소스는 명시적으로 승인된 경로를 통해서만 연결되어야 합니다. 최소 권한 범위 지정 및 세분화가 없으면 과도한 권한을 부여받은 에이전트 또는 손상된 AI 구성 요소가 측면 이동 경로가 될 수 있습니다.

주요 통제

- AI 앱, 에이전트, 모델 커넥터, 데이터 세트, 벡터 저장소 및 AI 인프라 구성 요소에 대한 역할 기반 액세스 범위 지정
- 민감한 데이터 세트 및 AI 연결 시스템에 대한 과도한 접근 권한에 대한 정기적인 검토 및 축소
- AI 에이전트를 위한 최소 권한 설정: 접근 가능한 항목을 정의하고, 그 외의 것은 모두 제한
- AI 서비스, 에이전트 플랫폼, 데이터 파이프라인, RAG 구성 요소 및 백엔드 데이터 소스 세분화
- AI 구성 요소 간의 동서 방향 접근을 명시적으로 정의된 경로로 제한
- MCP 및 A2A 트래픽에 대한 프로토콜 인식 검사 및 로깅

모든 AI 접근 경로에서 피해 범위 최소화

모든 AI 커넥터, 에이전트 권한, RAG 파이프라인, 벡터 저장소, MCP 서버 및 구성 요소 간 연결을 노출 지점으로 간주합니다. 접근 가능한 영역을 범위로 설정하고, 통신 가능한 영역을 구분하고, 트래픽을 검사하고, 활동을 기록합니다.



©2026 Zscaler, Inc. 모든 권리 보유.

원칙 6:
지속적인 모니터링 및 감사 가능성

지속적인 검증은 달성하고 넘어가는 이정표가 아니라 운영 모델입니다. AI 시스템의 경우, 이는 즉각적인 대응 로깅, 정책 위반에 대한 실시간 알림, 사용량 대시보드, 그리고 서로 다른 여러 도구에서 로그를 수동으로 상관 분석할 필요가 없는 조사 워크플로를 의미합니다. **이러한 요구 사항을 뒷받침하는 규제 압박은 현실적입니다.** 유럽연합 인공지능법(EU AI Act), 미국 국립표준기술연구소(NIST) 인공지능 위험 관리 프레임워크(NIST AI RMF), 그리고 새롭게 등장하는 AI 거버넌스 프레임워크 모두 입증 가능하고 문서화된 통제를 요구합니다.

주요 통제

- 인라인 AI 상호작용에 대한 프롬프트 및 응답 로깅
- 사용자, 부서 및 AI 앱별 사용량 대시보드와 위험 신호 오버레이
- AI 관련 정책 위반 및 이상 징후에 대한 경고 및 조사 워크플로
- 감사 준비 보고서: 사용 중인 AI, 적용된 통제 및 발생한 위반 사항
- NIST AI RMF, EU AI법 및 기타 관련 프레임워크에 따른 거버넌스 태세 매핑



©2026 Zscaler, Inc. 모든 권리 보유.

1	2	3	4	5
반응형	**가시성**	**거버넌스 적용**	**자동화**	**지속성**
인벤토리 없음 임시 차단	AI 발자국 파악 기본 정책	역할 기반 정책 거버넌스 워크플로	자동화된 코칭, 지속적인 규정 준수	레드티밍 + 런타임 가드레일

프로그램 성숙도 →

1~30일
가시성 확보

보이지 않는 AI는 보호할 수 없습니다. 처음 30일은 모든 것을 확정짓는 것이 아니라, 정보에 입각한 의사결정을 내릴 수 있도록 충분한 인벤토리를 확보하고 기준선을 구축하는 데 집중해야 합니다.

1~30일 가시성 확보	31~60일 정책 성숙도 구축	61~90일 자동화 및 검증
• AI 기반 앱 검색 기능 배포 • 레드 데이터 범주 정의 • 기본 접근 정책 • 최초의 경영진 대시보드	• 역할 기반 접근 정책 • 거버넌스 워크플로 가동 • AI-SPM 도입 • 프라이빗 AI 보호	• 자동화된 사용자 코칭 • 지속적인 규정 준수 • 레드티밍 주기 • 폐쇄 루프 가드레일

30/60/90일 구현 로드맵 개요



©2026 Zscaler, Inc. 모든 권리 보유.

30일 구현 체크리스트

- **정기적인 AI 도구 검토 주기수립:** 트래픽에서 발견된 새로운 도구, 접수된 예외 요청 및 정책 변경 사항 검토
- 경영진이 볼 수 있는 최초 AI 보안 대시보드 구축: AI 앱 발자국(인증된 앱 vs. 섀도 앱), 카테고리별 DLP 이벤트, 정책 조치 및 추세선
- AI 관련 보안 이벤트 중 가장 흔한 사고 플레이북 정의: 프롬프트 내 민감 데이터, 차단된 목적지에 대한 접근, AI 도구에 대용량 파일 업로드
- **AI 보안 로그 중앙 집중화:** 사용자 활동, 앱 사용량 등 프롬프트/응답 이벤트, 정책 조치 및 위반 사항을 단일 대시보드에서 확인

운영 체계화

- **첫 2주간의 텔레메트리를 활용한 DLP 정책 검토 및 조정:** 오탐을 줄이고, 확인된 데이터 노출 패턴에 대한 탐지율 향상
- **허용 가능한 사용 정책 시행을 위한 콘텐츠 모더레이션 활성화:** 주제와 무관한 상호 작용, 제한된 주제, 경쟁에 민감한 콘텐츠 및 부적절한 결과물 등 차단
- **허용되지만 위험도가 높은 계층에 속하는 AI 도구에 대해 브라우저 격리 기능 추가:** 사용자는 도구에 접근할 수 있지만 업로드, 다운로드, 복사/붙여넣기는 브라우저 계층에서 제한됨
- **DLP 적용 범위를 레드 데이터 전체로 확장:** 소스 코드(추가 언어 및 패턴), PII(명확한 형식 이외 포함), PCI 카드 소지자 데이터, PHI, 기밀 비즈니스 콘텐츠

데이터 보호 강화

- **AI 사용 허용 기준에 대한 트래픽 라이트 가이드 게시:** 녹색(허용), 노란색(제한적 사용), 빨간색(사유 명시 후 차단). 예외 요청 절차 공개
- 레드 데이터 범주에 대해 신뢰도가 높은 탐지 규칙을 시작으로 프롬프트에 대한 인라인 DLP 활성화: 소스 코드 패턴, 자격 증명 형식, PII 스키마
- **미승인 도구에서 가장 위험한 작업 제한:** 업로드, 대량 복사/붙여넣기, 관리되지 않는 저장소로의 AI 출력물 다운로드
- **목적지 정책 시행:** 명확히 승인되지 않았거나 위험도가 높은 AI 목적지에 대해서는 사용자 그룹 및 앱별로 경고 또는 차단

기준 통제 적용

- **기준 AI 위험 인벤토리 구축:** 어떤 앱이 어떤 유형의 데이터에 접근하는지, 그리고 어떤 사용자 그룹이 AI를 가장 많이 사용하는지 파악
- **레드 데이터 범주를 명확하게 정의:** 소스 코드, PII, PCI 데이터, PHI, 시크릿 및 자격 증명, 미공개 재무 데이터, 기밀 M&A 또는 전략 계획

보호해야 할 대상 정의

- 가장 많은 트래픽이 발생하는 AI 앱에 대한 가시성을 즉시 확보하여 프롬프트와 응답을 통해 어떤 데이터가 이동하는지 파악
- 어떤 AI 도구가 내부 시스템과 통합되어 있는지, 그리고 어떤 데이터 유형에 접근하는지 파악
- **섀도 AI 파악:** 공식 승인 없이 실제 사용 중인 도구
- AI 앱 탐지 체계를 배포하여 환경 전반에서 활용되는 GenAI SaaS, 엔터프라이즈 플랫폼 임베디드 AI, AI 개발자 도구에 대한 인벤토리 구축

고려해야 할 트레이드오프

엄격한 차단보다는 경고 정책부터 시작하면 마찰을 줄이고 DLP 규칙이 정상적인 작업을 차단하기 전에 오탐을 발견할 수 있습니다. 단점은 경고 단계에서도 데이터 유출 사고가 발생할 수 있다는 것입니다. 소스 코드 유출이 심각한 위험이라면, 코드 패턴 탐지를 엄격하게 차단하는 것이 초기 오탐으로 인한 손실을 감수할 만한 가치가 있는지 여부를 위험 허용 범위와 비교하여 균형을 맞춰야 합니다.

30일 내 주요 산출물

- AI 앱 인벤토리 및 위험도 등급 분류(허용, 조건부 허용, 금지)
- 위험도가 가장 높은 AI 앱 및 사용자 그룹에 대한 초기 접근 정책 적용
- 레드 데이터 범주에 대한 프롬프트 DLP 활성화 및 초기 조정 완료
- AI 활용도 및 위험 지표를 포함한 최초의 경영진 대시보드
- AI 사용 허용 가이드 및 예외 요청 절차가 게시 완료

31~60일
정책 성숙도 구축

31~60일 차까지는 가시성과 초기 제어 기능을 기반으로 성숙한 정책 모델, 효율적인 거버넌스 워크플로 및 AI 인프라 인벤토리를 구축합니다.



©2026 Zscaler, Inc. 모든 권리 보유.

60일 구현 체크리스트

접근 정책 모델 고도화

- **부서별 AI 접근 정책 수립:** HR, 재무, 엔지니어링, 영업 부서별 AI 툴 수요 및 데이터 노출 위험 프로파일 반영
- **계층별 조건부 정책 제어:** 관리형 장치 대 BYOD 액세스 계층, 지리적 액세스 신호, 장치 상태 위험 점수
- **AI 앱의 세 가지 액세스 등급을 공식화:** 승인(인라인 검사를 통한 전체 액세스), 허용(사용 제한이 있는 모니터링 액세스), 금지(예외 경로가 있는 차단)
- 처음 30일 동안 접수된 예외 요청 재검토 및 고도화된 정책 모델 기준 만료 처리

거버넌스 프로세스 수립

- **AI 거버넌스 요구사항을 내부 정책 및 적용 가능한 프레임워크에 매핑:** NIST AI RMF, EU AI법, PHI가 해당되는 경우 HIPAA
- **AI 도구 승인 워크플로 공식화:** 제출, 보안 검토, 위험 분류, 조건부 승인 및 주기적 재검증
- AI 활용과 관련된 데이터 유형에 대한 표준 민감도 레이블 설정 및 각 레이블에 대한 명확한 처리 규칙 마련
- **AI 보안 사고 발생 시 대응 절차 정의:** 누가 대응을 담당하는지, 경영진 보고가 필요한 사고는 무엇인지, 그리고 어떤 증거를 보존해야 하는지 명시

인벤토리 AI 인프라

- AI 모델, 에이전트, 서비스, 데이터 세트, 벡터 저장소, 파이프라인의 인벤토리 구축 및 데이터 세트에서 런타임 사용까지의 계보를 포함한 AI-BOM 구축
- **AI-SPM 평가 실행:** AI 인프라 구성 요소 전반에 걸쳐 잘못된 구성, 과도한 권한, 민감 데이터 노출 및 취약점 식별
- **고위험 AI-SPM 조치 우선 처리:** 과도한 권한, 노출된 민감 데이터, 제약 없는 AI 구성 요소 접근 경로
- **주요 AI 플랫폼 전반의 적용 범위 검증:** 클라우드 호스팅 모델, 전사 구축형 에이전트 플랫폼, MCP 서버 연동 및 관리되지 않는 AI 인프라

프라이빗 AI 애플리케이션 보호

- **프라이빗 AI 앱 통제에 프롬프트 인젝션 탐지 기능 추가:** 모델 지침을 무시하거나 민감 정보를 추출하도록 악의적으로 조작된 입력값
- **RAG 파이프라인에 대한 데이터 포이즈닝 모니터링 구현:** 모델 출력에 영향을 줄 수 있는 문서 코퍼스에 대한 무단 수정 탐지
- **프라이빗 AI 시스템에 대한 이상 징후 탐지와 함께 프롬프트/응답 로깅 활성화:** 비정상적인 쿼리 패턴, 예상치 못한 출력 내용 및 범위를 벗어난 상호 작용
- 민감도가 높은 프라이빗 AI 애플리케이션에 대한 모델 상호작용 검토 워크플로 구축

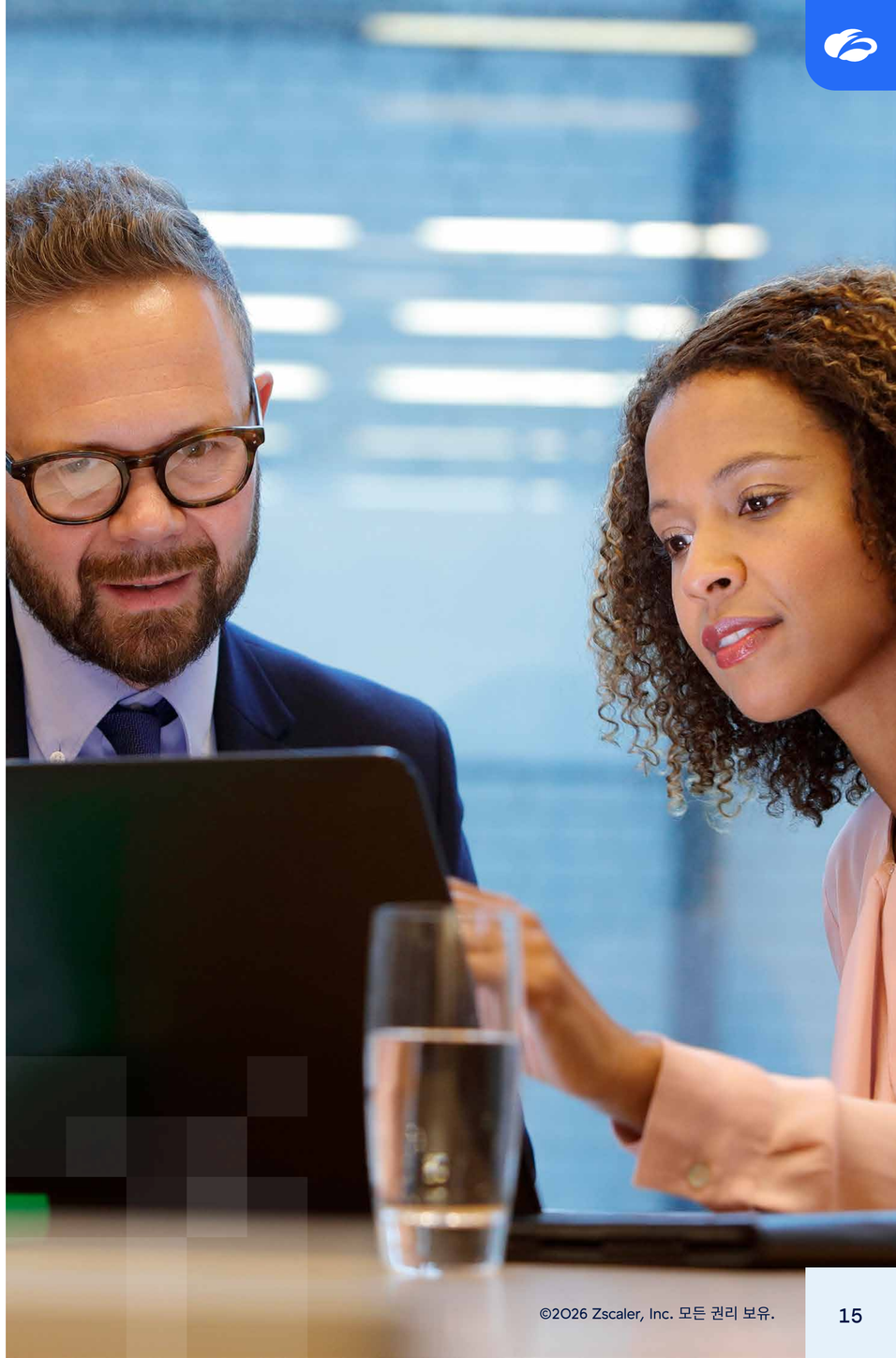
©2026 Zscaler, Inc. 모든 권리 보유.

60일 내 주요 산출물

- 조건부 적용 기능을 갖춘 성숙한 역할 기반 접근 통제 정책 모델
- AI 거버넌스 워크플로 운영 완료: 도구 승인, 민감도 라벨링, 에스컬레이션 경로 정의
- AI-SPM 가시성 기준선 구축 및 우선순위가 지정된 개선 백로그 완료
- AI 인프라의 주요 구성 요소를 다루는 AI-BOM
- 최우선 순위 내부 AI 앱에 대한 프라이빗 AI 보호 조치 마련
- 규정 준수 보고 주기 정의 및 첫 보고서 작성

61~90일
자동화 및 검증

61~90일, 프로그램 자립화: 자동화된 시행, 지속적인 규정 준수 모니터링, 레드티밍을 통한 적대적 검증.



©2026 Zscaler, Inc. 모든 권리 보유.

90일 구현 체크리스트

법 시행 자동화

- **정책 위반에 대한 자동 사용자 코칭 구현:** 차단으로 이어지기 전에 상황에 맞는 안내 제공 및 경고
- **반복 위반에 대한 ITSM 또는 SOAR 티켓팅 연동:** 동일한 통제를 반복적으로 트리거하는 사용자에 대해 조사 티켓을 자동으로 생성
- **고위험 행동에 대한 자동 격리 및 분리 트리거 구성:** 대량의 민감 데이터 제출, 금지된 AI 대상 접근, 비정상적인 프롬프트 패턴
- **피드백 루프 구축:** 정책 위반 추세를 시간 경과에 따라 추적하여 정책 시행 실패가 아닌 정책 조정이 필요한 행동 유형 파악

규정 준수 보고 확정

- **AI 규정 준수 보안 태세를 위한 상시 모니터링 구성:** 어떤 통제 기능이 활성화되어 있는지, 어떤 자산이 모니터링 대상에 포함되는지, 그리고 모니터링 범위에 공백이 있는 곳은 어디인지 확인
- **적용 가능한 프레임워크에 대한 제어 사항 매핑:** NIST AI RMF(생성형 AI 위험에 대한 NIST AI 600-1 포함), EU AI법 의무 사항, GDPR 및 PHI가 적용되는 HIPAA
- **감사 대응 보고 패키지 구축:** AI 앱 사용 요약, 신속한 DLP 조치 및 결과, 승인 및 만료된 정책 예외, AI-SPM 분석 결과 조치 현황
- **분기별 AI 보안 프로그램 검토 주기 수립:** 정책 성과, 적용 범위 격차, 새로운 위험 영역 및 로드맵 조정 사항 검토

레드팀 및 보안 강화

- **자체 구축 AI 애플리케이션에 대한 AI 레드티밍 주기 수립:** 사전 구축된 공격 프로브 라이브러리와 특정 애플리케이션 콘텍스트를 대상으로 하는 맞춤형 테스트 케이스
- **AI 공격 표면 전체를 포괄하는 테스트 실행:** 프롬프트 인젝션, 탈옥, 콘텍스트 포이즈닝, PII 추출, 소스 코드 유출, 주제 이탈 행동 및 모델 벤치마크 회귀
- **운영 중인 AI 시스템에 런타임 보호 정책 배포:** 프롬프트 인젝션, 데이터 유출(PII 및 소스 코드), 탈옥 시도 및 콘텐츠 모더레이션 위반 탐지
- **레드티밍 분석 결과를 런타임 가드레일로 자동 변환:** 대부분의 AI 애플리케이션 보안 프로그램은 테스트와 실제 운영 환경 적용 사이의 간극에서 실패

16분	72%	100%
AI 레드팀 테스트에서 첫 번째 심각한 오류 발생까지의 중간값 25개 이상의 환경에서 222,000건 이상의 적대적 공격 실행	첫 번째 테스트에서 치명적인 취약점을 노출한 기업 성숙한 환경에서도 처음부터 존재하는 심각한 위험	테스트된 AI 시스템에는 최소 하나 이상의 심각한 취약점이 있었습니다. 어떤 AI 시스템도 기본적으로 안전하지 않습니다. 지속적인 테스트는 필수적입니다.

©2026 Zscaler, Inc. 모든 권리 보유.

아키텍처 정합성 검증

- **모든 AI 접근 경로를 제로 트러스트 6대 원칙에 매핑:** 각 원칙이 임직원, 개발자 및 AI 인프라 접근 표면 전반에 적용되었는지 확인
- **접점의 정책 공백 식별 및 해소:** 정책 적용 범위가 일관되지 않은 교차점, 즉 한 보안 통제 계층이 끝나고 다른 계층이 시작되는 지점
- 신원 확인, 목적지 통제, 위험 평가, 현장 집행, 접근 및 노출 최소화, 모니터링을 포괄하는 통합 정책 모델을 통해 시행 표준화

90일 내 주요 산출물

- 자동화된 위반 대응 및 사용자 교육 워크플로 운영 중
- 프레임워크 정합성이 문서화된 실시간 규정 준수 모니터링 가동
- 경영진 및 감사용 거버넌스 보고 패키지
- 런타임 보호 정책이 활성화된 AI 레드티밍 주기 수립
- 레드티밍 분석 결과를 운영 가드레일로 전환하는 폐쇄 루프 방식
- 지속적인 관리 운영을 위한 프로그램 성과표 및 분기별 로드맵



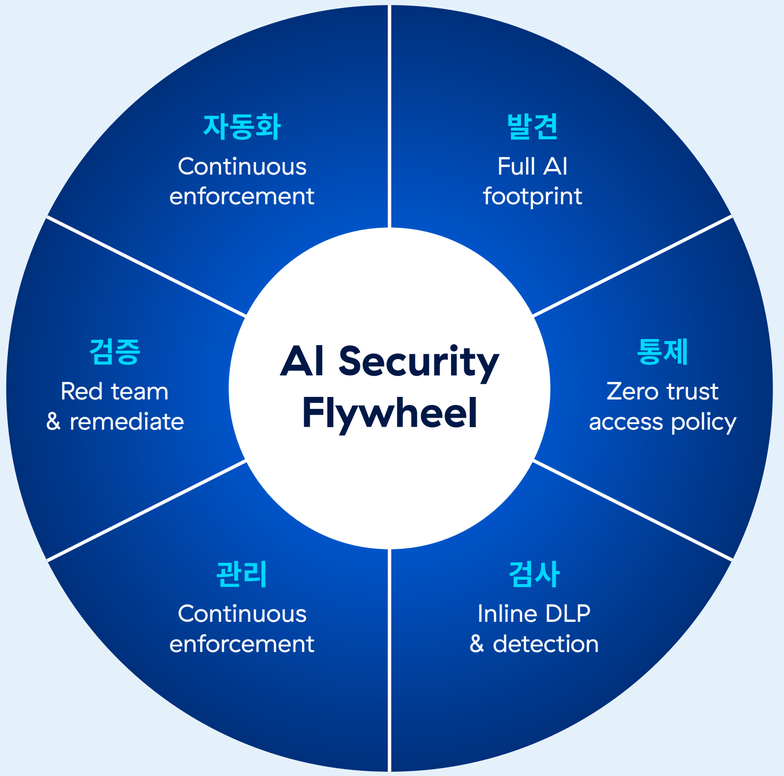
©2026 Zscaler, Inc. 모든 권리 보유.

Zscaler가 안전한 AI 도입을 지원하는 방법

Zero Trust Exchange™ 플랫폼은 기업 규모의 AI를 관리하고 보호하기 위해 구축되었습니다. 이는 전 세계 수천 개의 기업에서 사용자, 워크로드 및 네트워크 액세스를 관리하는 데 이미 사용되고 있는 동일한 플랫폼을 확장한 것이며, 레거시 보안 아키텍처에 추가되는 것이 아닙니다.

플라이휠의 각 단계는 특정 플랫폼 기능 세트와 연결됩니다. 아래 표는 Zscaler가 각 계층에서 제공하는 기능을 보여줍니다.

발견 AI 발자국을 완벽하게 파악	**통제** AI에 대한 제로 트러스트 액세스 시행	**보호** AI 애플리케이션의 보안 강화 및 방어
AI Asset Management Zscaler는 IT 부서에서 승인하지 않은 도구를 포함하여 환경 내의 모든 AI 애플리케이션, 모델, 에이전트, MCP 서버 및 임베디드 SaaS AI 기능에 액세스합니다.	**AI 액세스 보안** AI SaaS, 엔터프라이즈 플랫폼 내 임베디드 AI, AI 개발자 환경에 대해 모든 계층에서 인라인 검사를 통해 제로 트러스트 액세스 제어를 시행합니다.	**AI 레드티밍 및 AI 가드레일** 자체 개발한 AI 애플리케이션을 실제 공격 환경에서 테스트한 후, 결과를 자동으로 런타임 보호 정책으로 변환합니다.
• Shadow AI의 발견 및 모든 사용자 집단에 걸친 사용 현황 파악 • **AI 자재명세서(AI-BOM):** 데이터 세트에서 모델, 에이전트 및 런타임 사용에 이르는 계보 • **AI-SPM:** 잘못된 구성, 과도한 권한 및 민감 데이터 노출에 대한 지속적인 보안 태세 평가 • 기업에 배포된 AI 에이전트와 SaaS 임베디드 에이전트 전반에 걸친 AI 에이전트 탐지	• 사용자 그룹, 앱 및 데이터 유형별로 세부적인 허용, 경고, 격리 및 차단 정책 설정 • 모든 AI 트래픽에 걸쳐 프롬프트, 응답 및 파일 업로드에 대한 인라인 DLP 적용 • 관리되지 않는 환경 및 BYOD 환경에서 AI 도구에 접근하기 위한 제로 트러스트 브라우저 • 주요 GenAI 앱을 위한 신속한 추출 및 분류	• **자동화되고 지속적인 레드티밍:** 프롬프트 인젝션, 탈옥, 콘텍스트 포이즈닝, 데이터 유출, 행동 드리프트 • 프롬프트 보안 강화 및 모델 벤치마킹, NIST AI RMF, EU AI Act, OWASP LLM Top 10에 대한 거버넌스 매핑 제공 • **런타임 가드레일:** 탈옥, PII/소스 코드 유출, 콘텐츠 모더레이션, 주제 이탈 행동에 대한 탐지기 • **폐쇄 루프 정책 생성:** 레드티밍 분석 결과가 자동으로 실제 운영 환경에서의 탐지기로 전환



플랫폼 이점

대부분의 보안 업체는 AI 보안 문제의 한 부분만을 해결합니다. Zscaler는 단일 플랫폼에서 AI 라이프사이클 전반을 지원합니다. 어디에서든 AI를 발견하고, 제로 트러스트 정책을 사용하여 누가 어떤 AI에 접근할 수 있는지 제어하며, 빌드부터 런타임까지 AI 앱과 인프라를 보호합니다.레드티밍 분석 결과와 실제 운영 환경에서의 안전장치 간의 폐쇄 루프는 어떤 단일 솔루션도 따라할 수 없는 핵심 기능입니다.

©2026 Zscaler, Inc. 모든 권리 보유.

자주 묻는 질문

'제로 트러스트 AI'는 실제로 무엇을 의미하나요?
제로 트러스트 AI란 암묵적인 신뢰를 가정하지 않고 지속적으로 검증하는 핵심 원칙을 환경 내 모든 AI 상호작용에 적용하는 것을 의미합니다. 실제로 AI 앱, 사용자, 기기, 프롬프트, 커넥터 등 그 어떤 것도 본질적으로 신뢰되지 않습니다. 접근 권한은 검증된 신원, 기기 상태, 데이터 민감도, 세션 콘텍스트를 기반으로 부여됩니다. 제어는 실시간으로 인라인 방식으로 적용되며, 모든 접근 결정은 기록되어 감사가 가능합니다.

새도 AI 위험을 줄이기 위해 처음 30일 동안 무엇을 해야 할까요?
먼저 AI 위협을 파악하는 것부터 시작하세요. 보이지 않는 AI는 관리할 수 없습니다. ThreatLabz는 기업 트래픽에서 3,400개 이상의 AI 애플리케이션을 추적하고 있는데, 이 수치는 단 1년 만에 4배나 증가했으며, 이는 주로 IT 부서에서 이미 승인한 도구 내 임베디드 AI 기능 때문입니다. AI 위협을 파악한 후에는 위험도가 가장 높은 애플리케이션을 분류하고 기준이 되는 대상 정책을 설정하세요. 위험도가 가장 높은 데이터 범주에 대해서는 프롬프트 DLP를 추가하세요. 투명한 사용 지침과 예외 요청 절차를 게시하여 사용자가 보안 조치를 우회하려는 동기가 아닌 규정을 준수하는 경로를 제시하세요.

AI 접근 통제 및 인라인 DLP는 프롬프트와 업로드를 어떻게 보호하나요?
접근 통제는 누가 어떤 AI 도구에 어떤 기기와 위치에서 어떤 조건으로 접근할 수 있는지를 결정합니다. 인라인 DLP는 콘텐츠 계층에서 작동하며 사용자가 실제로 제출하는 내용, 프롬프트 텍스트, 파일 업로드, 복사/붙여넣기 상호작용 등을 실시간으로 검사합니다. 두 가지 제어 방식은 상호 보완적입니다. 접근 통제는 누가 어떤 AI에 접근할 수 있는지를 제한하고, DLP는 허용된 채널을 통해 어떤 데이터가 이동하는지를 통제합니다. 어느 한쪽이라도 허점이 있으면 노출 위험이 발생하므로, 업로드에는 DLP를 적용하면서도 즉각적인 검사를 생략하는 조직은 여전히 상당한 데이터 유출 위험에 직면하게 됩니다.

AI-SPM이란 무엇이며, AI 도입에 있어 왜 중요한가요?
AI-SPM은 AI 인프라 구성, 권한 및 노출 위험에 대한 지속적인 평가로, 클라우드 인프라의 CSPM에 해당하는 AI 기반 보안 관리 도구입니다. AI 자산을 발견하고, 보안 모범 사례에 따라 구성을 평가하며, 과도한 권한 및 민감 데이터 노출을 식별하고, 시간 경과에 따른 보안 상태 변화를 추적합니다. AI 인프라의 잘못된 구성 가능성은 매우 크고, AI 실험에서 배포 단계로 넘어가는 대부분의 조직이 이를 제대로 이해하지 못하고 있기 때문에 AI-SPM은 매우 중요합니다.

생산성을 저해하지 않으면서 AI 위험이 감소하고 있음을 경영진에게 가장 효과적으로 보여주는 KPI는 무엇일까요?
가장 효과적인 접근 방식은 위험 감소 지표와 도입 지표를 함께 제시하여 경영진이 두 가지 추세를 동시에 파악할 수 있도록 하는 것입니다. 승인된 AI 도입이 증가하는 동시에 비공식 AI 앱의 수가 감소하는 것은 거버넌스가 제대로 작동하고 있음을 보여주는 명확한 신호입니다. 프롬프트 DLP 조치(차단, 경고, 배포 허용)는 규제 집행이 어떻게 조정되는지를 보여줍니다. 예외 요청 처리 시간과 AI 도구 승인 주기 시간은 거버넌스의 대응력을 나타냅니다.

ZSCALER AI SECURITY

- Zero Trust Exchange Platform
- AI Asset Management
- AI 액세스 보안
- AI Red Teaming
- AI 가드레일

zscaler.com/ai-security



©2026 Zscaler, Inc. 모든 권리 보유.

zscaler™ | 빠른 대응. 보안 유지.

Zscaler 소개

Zscaler(NASDAQ: ZS)는 제로 트러스트 보안 분야의 선구자이자 글로벌 리더입니다. 세계 최대 규모의 기업, 중요 인프라 기관 및 정부 기관은 사용자, 지사, 애플리케이션, 데이터 및 기기를 보호하고 디지털 전환 이니셔티브를 가속화하기 위해 Zscaler를 사용합니다. 전 세계 160개 이상의 데이터 센터에 분산된 Zscaler Zero Trust Exchange™ 플랫폼은 첨단 AI와 결합하여 매일 수십억 건의 사이버 위협 및 정책 위반에 대응하고 비용과 복잡성을 줄여 현대 기업의 생산성 향상을 지원합니다.

© 2026 Zscaler, Inc. All rights reserved. Zscaler™ 및 **zscaler.com/legal/trademarks**에 명시된 기타 상표는 (i) 미국 및/또는 기타 국가에서 Zscaler, Inc.의 등록 상표 또는 서비스표이거나 (ii) Zscaler, Inc.의 상표 또는 서비스표입니다. 기타 모든 상표는 해당 소유자의 자산입니다.

+1 408.533.0288 **Zscaler, Inc. (본사) • 120 Holger Way • San Jose, CA 95134** **zscaler.com**