

AX 시스템 안정적 운영

AI 도입 이후의 과제: AI Platform Operations로 전환하는 기업 AI 운영 전략

글 | SK AX, Cloud Tech본부 유영채 본부장
SK AX, Cloud Tech본부 김지한 전문위원

AI

01· AI 도입의 초점 변화:
PoC에서 운영으로

02· AI Platform Operations의 필요성

03· RAG Operations:
지식 기반 AI 서비스 운영

04· LLMOps:
생성형 AI 서비스의 품질, 비용, 성능 관리

05· Agentic AIOps:
AI 운영 이벤트 대응의 고도화

06· AI MSP의 진화:
AI 운영체계를 서비스화하는 모델

07· AI 확산을 위한 운영체계 전환과 실행 과제

1. AI 도입의 초점 변화: PoC에서 운영으로

- AI 도입의 초점은 기능 검증에서 안정적인 운영체계 확보로 이동하고 있다.
- 품질, 비용, 보안, 성능을 함께 관리해야 PoC 성과를 확산할 수 있다.

2. AI Platform Operations의 필요성

- AI Platform Operations 운영 대상과 책임 지표를 연결하는 프레임이다.
- 핵심 지표를 경고와 조치 기준에 연결해야 일관된 운영이 가능하다.

3. RAG Operations: 지식 기반 AI 서비스 운영

- RAG 운영은 문서 품질, 최신성, Index, 검색 성능 변화를 관리한다.
- 데이터와 검색 기준을 정립해야 잘못된 지식의 활용을 줄일 수 있다.

4. LLMOps: 생성형 AI 서비스의 품질, 비용, 성능 관리

- LLMOps는 모델과 프롬프트의 품질, 비용, 성능 변화를 관리한다.
- 동일한 평가 기준으로 변경 영향을 검증하고 배포와 롤백을 결정한다.

5. Agentic AIOps: AI 운영 이벤트 대응의 고도화

- Agentic AIOps는 운영 이벤트를 해석하고 원인과 조치 후보를 제시한다.
- 신뢰 가능한 데이터와 검증 절차를 전제로 사람의 승인 아래 적용해야 한다.

6. AI MSP의 진화: AI 운영체계를 서비스화하는 모델

- AI MSP는 AI 운영 기준과 절차를 표준화해 서비스로 제공하는 모델이다.
- 기업은 통제권을 유지하며 전문 역량으로 운영체계를 보완할 수 있다.

7. AI 확산을 위한 운영체계 전환과 실행 과제

- AI 확산 전 운영 대상, 역할, 지표, 변경 관리 기준을 정립해야 한다.
- AI 도입의 성패는 모델보다 지속 가능한 운영 체계 확보에서 결정된다.

01 AI 도입의 초점 변화: PoC에서 운영으로

생성형 AI와 AI Agent는 일부 혁신 조직의 실험을 넘어 고객 응대, 내부 지식 검색, 문서 작성, 개발 업무, IT 운영 등 실제 업무로 빠르게 확산되고 있다. McKinsey는 기업의 AI 활용이 늘고 있지만, 상당수 조직은 여전히 파일럿 단계에 머물러 있으며 전사 확산과 가치 실현에는 제약이 남아 있다고 분석한다.^[1]

● AI 도입의 논점

“왜 써야 하는가”보다 “도입 이후 어떻게 안정적으로 운영할 것인가”에 가까워지고 있다.

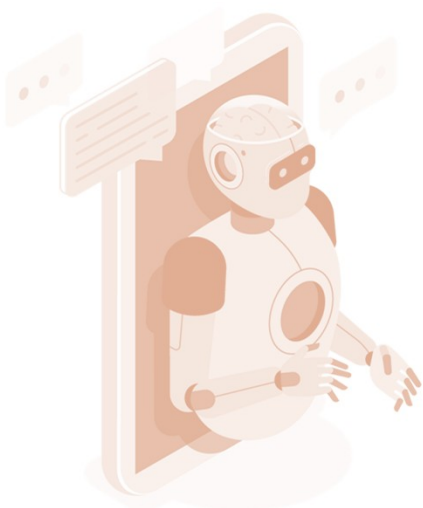


PoC 단계의 관심은 기능 시연과 사용 가능성 검증에 가깝다. 특정 업무에 AI를 적용했을 때 답변이 생성되는지, 검색이 가능한지, 사용자가 체감할 만한 효용이 있는지를 확인하는 데 초점이 맞춰진다. 그러나 운영 전환 단계에서는 판단 기준이 달라진다. PoC가 성공했다는 사실만으로 전사 확산을 결정하기 어렵다. AI 응답 오류가 발생했을 때의 책임 범위, 사용량 증가에 따른 비용 통제, 민감 정보 검색과 외부 모델 호출 과정의 보안 기준까지 함께 검토해야 한다. 이 지점부터 AI는 기술 검증 과제를 넘어 운영 의사결정의 영역으로 들어온다.

AI 서비스가 핵심 업무 프로세스에 연결되면 운영 리스크도 구체적인 형태로 드러난다. 답변 품질이 일정하지 않으면 AI 서비스에 대한 사용자의 신뢰부터 흔들린다. 응답 지연은 업무 처리 시간을 늘리고, 사용량 증가는 Token과 API 비용으로 이어진다. 권한 통제가 미흡한 환경에서는 부적절한 데이터가 검색 근거로 사용될 가능성도 커진다. 모델, 프롬프트, 데이터 Index의 변경 역시 단순한 기능 개선으로 보기 어렵다. 하나의 변경이 품질과 비용, 보안 수준을 동시에 흔들 수 있기 때문이다.

Deloitte는 AI 확산 단계에서 기업 리더가 ROI, 안전한 활용, 윤리, 인력 준비도, 실행 체계를 주요 과제로 보고 있다고 설명한다.^[2] Stanford HAI가 제시한 AI 투자와 기업 활용 증가 역시 AI가 실험을 넘어 운영 관리 대상으로 전환되고 있음을 보여준다.^[3] 기업의 과제는 도입 규모를 늘리는 데서 끝나지 않는다. 품질, 보안, 비용 기준과 책임 구조를 정립하고, 변경과 장애를 통제할 수 있는 운영 기반을 갖춰야 한다. 그래야 PoC의 성과가 실제 업무 확산으로 이어진다.

이제 필요한 것은 AI 서비스의 운영 대상과 관리 지표, 책임 구조를 정리하는 일이다. AI Platform Operations는 바로 이 기준을 구조화하는 출발점이다.



02 AI Platform Operations의 필요성

AI Platform Operations는 AI 서비스를 실험 결과물에서 지속적으로 운영 가능한 체계로 전환하는 프레임이다. 특정 솔루션이나 모니터링 도구만을 뜻하지 않는다. 모델과 프롬프트, RAG, API Endpoint, 권한, Guardrail 등 AI 서비스를 구성하는 요소를 운영 범위에 포함하고, 각 대상의 책임과 변경 기준, 판단 지표를 연결하는 데 목적이 있다. NIST AI RMF는 AI 리스크 관리를 거버넌스와 책임 체계 수립(Govern), 사용 맥락과 위험 식별(Map), 리스크 측정과 평가(Measure), 리스크 대응과 관리(Manage)로 구조화한다.^[4] AI Platform Operations 역시 AI 운영을 개별 장애 대응이 아니라 조직 차원의 관리 체계로 본다는 점에서 같은 문제의식에 놓여 있다.

기존 IT 운영은 서버, 네트워크, 데이터베이스, 애플리케이션의 가용성과 성능, 장애 대응을 중심으로 발전해 왔다. 시스템 중단, 오류 발생, 응답 지연, 자원 부족이 주요 판단 기준이었다. AI 운영은 여기에 결과의 신뢰성까지 포함한다. 인프라와 API가 정상 이어도 모델 답변이 업무 기준에서 벗어날 수 있고, 검색 기능이 동작하더라도 오래된 문서나 잘못된 권한의 정보가 근거로 사용되면 응답 품질은 흔들린다. 장애가 없어도 특정 조직이나 업무의 Token 사용량이 급증하면 비용 이상으로 판단할 수 있다. 기존 IT 운영이 시스템의 정상 동작을 판단했다면, AI 운영은 적절한 근거와 권한 안에서 신뢰할 수 있는 결과가 생성되는지까지 살펴야 한다. AWS의 Generative AI Lens도 생성형 AI 워크로드를 보안, 신뢰성, 성능 효율, 비용 최적화, 운영 우수성의 관점에서 함께 설계하고 운영할 것을 제시한다.^[5]

AI Platform Operations의 핵심

→ AI Platform Operations의 핵심은 운영 영역별 관리 대상과 판단 지표를 연결하는 데 있다.

서비스 실행 영역

Endpoint의 가용성, 오류율, P95 Latency를 기준으로 상태를 판단한다.

모델 호출 영역

성공률과 Token 사용량, 분당 또는 초당 호출 제한을 의미하는 Rate Limit, 계정과 모델, 기간별 사용 한도인 Quota가 주요 관리 지표가 된다.

지식 검색 영역 RAG

검색 정밀도(Precision), 재현율(Recall), 관련성(Relevance), Index 최신성, Vector DB 성능, 문서 권한 적용 여부를 통해 운영 품질을 판단한다.

Guardrail도 AI 서비스의 주요 구성요소다. 다만 운영 관점에서는 서비스 실행이나 지식 검색과 구분되는 정책 통제 영역으로 다루는 편이 적절하다. 민감정보 차단율, 정책 위반 탐지율, 오타율과 미탐율, 통제 우회 발생 여부가 실효성을 판단하는 기준이 된다. 비용은 독립된 운영 대상이 아니며, 모델 호출, RAG 검색, 데이터 처리 등 AI 서비스 전반을 가로지르는 관리 항목이다. 사용자와 조직, 업무별 사용량과 비용 추이, 이상 사용 여부, 예산 임계 값을 함께 추적해야 비용 변동의 원인을 파악하고 구체적인 조치 기준을 마련할 수 있다.

AI 운영의 난점은 하나의 이상 징후가 여러 운영 영역에 걸친 원인과 영향으로 연결된다는 데 있다. Latency는 플랫폼 지표처럼 보이지만 사용자 경험과 직접 연결된다. AI 응답 품질 저하도 모델만의 문제로 설명하기 어렵다. 데이터, 프롬프트, 검색, 권한이 결과에 복합적으로 영향을 주며, 평가 기준에 따라 이상 여부의 판단도 달라진다. Microsoft의 AI Observability 자료도 AI 시스템을 수명주기 전반에서 모니터링하고, 원인을 이해하며, 문제 해결과 품질 개선으로 연결할 수 있어야 한다고 설명한다.^[6]

따라서 기업은 AI 확산에 앞서 운영 대상과 책임 주체, 핵심 지표의 관계부터 정의해야 한다. 정상 범위와 경고 임계 값, 변경 승인, 보고 주기, 장애 대응 기준이 결합해야 실제 운영 체계로 기능한다. 이 기반이 마련되면 RAG Operations, LLM Ops, Agentic AI Ops도 분리된 기술 활동이 아니라 하나의 AI 운영체계를 구성하는 실행 영역으로 정렬된다. 그중 데이터 최신성과 검색 정확도, 문서 권한이 직접적인 운영 품질로 이어지는 대표 영역이 RAG다.

03 RAG Operations: 지식 기반 AI 서비스 운영

RAG(검색 증강 생성)는 생성형 AI가 외부 지식 기반을 검색해 답변 근거로 활용하도록 만드는 방식이다. Google Cloud는 RAG를 LLM과 외부 지식 기반을 결합해 출력 품질을 개선하는 접근으로 설명하며, AWS도 데이터 소스를 활용해 생성 응답의 관련성과 정확도를 높이는 방식으로 제시한다.^{[7][8]} 다만 기업 환경에서 RAG의 의미는 단순한 검색 기술에 그치지 않는다. 내부 문서, 업무 매뉴얼, 정책 자료, 장애 이력, 고객 대응 기록을 AI 서비스의 판단 근거로 연결하는 운영 구조에 가깝다.

● RAG 운영의 핵심

“검색이 되는가”가 아니라 “올바른 지식,
올바른 권한으로, 최신 상태에서 검색되는가”에 있다.



데이터 소스 관리

RAG 운영은 데이터 소스 관리에서 시작된다. 어떤 문서를 수집할 것인지, 수집된 문서를 어떻게 정제할 것인지, 승인되지 않은 문서가 유입되지 않도록 어떤 절차를 둘 것인지가 먼저 정해져야 한다. 이 기준이 없으면 검색 품질은 빠르게 흔들린다. 중복 문서, 오래된 문서, 잘못된 메타데이터, 승인되지 않은 초안이 함께 쌓이면 AI는 그중 어떤 문서가 유효한 근거인지 판단하기 어렵다. RAG 운영에서 잘못된 응답을 유발하는 주요 원인은 오염된 데이터, 갱신되지 않은 Index, 잘못된 메타데이터나 문서 분류로 부적절한 근거가 검색되는 데 있다. 따라서 모델 조정 전에 데이터 소스의 소유자, 최신성, 승인 상태와 Index 반영 이력을 추적해야 한다.

더 많은 내용을 보시려면

파일 다운받기

버튼을 눌러주세요