디지털서비스

이슈리포트

Digitalservice Issue Report

추론 클라우드

NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2026 Vol. 06

CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

01 추론 클라우드

| 한상기 테크프론티어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

들어가며

지난 1월 아주대 윤대균 교수가 '네오 클라우드'의 개요와 주요 기업에 대해 소개했다.[1] 네오 클라우드는 IaaS의 AI 특화 버전이라고 할 수 있는 GPUaaS(Gpu as a Service)에 집중하는 차세대 클라우드 서비스를 말한다. 시너지 리서치 그룹(Synergy Research Group)의 데이터에 따르면 네오클라우드 매출이 전례 없는 속도로 성장하여 4분기에 90억 달러에 달했으며, 이는 전년 동기 대비 223% 증가한 수치이다. AI 인프라에 대한 수요 급증에 힘입어 시너지는 네오클라우드 시장이 2031년까지 4,000억 달러에 육박할 것으로 전망하며, 이는 연평균 58%의 복합 성장률을 의미한다.

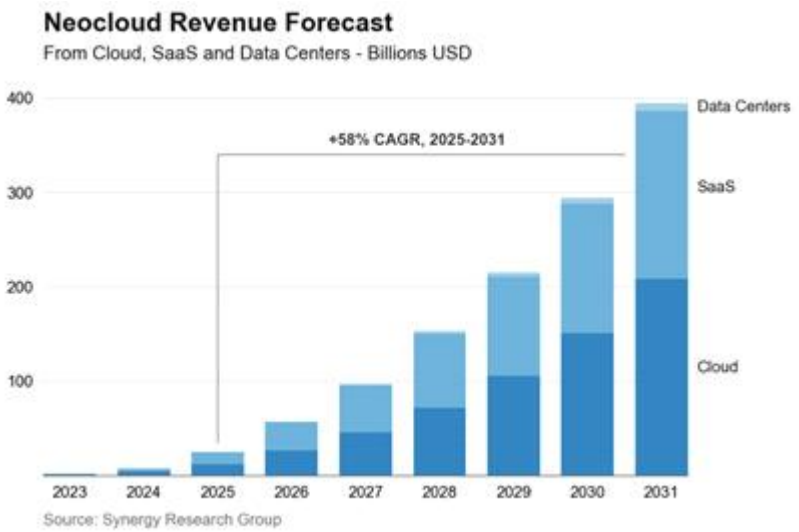


그림 1 네오클라우드 매출 예측 [출처: 시너지 리서치]

1) 윤대균, "네오클라우드 – AI 시대의 새로운 인프라 패러다임," NIA 디지털 서비스 리포트, 2026년 1월

수많은 신규 진입 업체와 전환 중인 암호화 인프라 제공업체가 계속해서 등장하는 가운데, 네오클라우드는 GPU 가속 컴퓨팅에 집중함으로써 차별화를 꾀하고 있으며, 이를 통해 AI 워크로드에 더욱 높은 성능 밀도, 빠른 배포 주기, 효율적인 확장을 제공하고 있다.

2026년 5월 네오클라우드 기업인 '제너럴 컴퓨트(General Compute)'는 '추론이 파편화되고 있다'는 백서를 발표하면서 에이전트, 특수 실리콘, 그리고 추론 시장이 어떻게 변화하고 있는지 정리했다.[2] 이번 원고는 이 백서를 중심으로 AI 추론 클라우드가 어떤 특성이 있으며 앞으로 어떻게 진화할 것인지를 정리해 보기로 한다. 제너럴 컴퓨트가 지향하는 영역이 에이전트를 위해 구축되고 있는 추론 클라우드이기 때문에 이 내용을 살펴보는 것이 의미가 있다고 생각한다.

추론 시장의 분화

모든 성숙한 컴퓨팅 시장은 분화한다. 데이터베이스가 그랬고, 네트워킹 칩 시장, 모바일 칩도 마찬가지다. 이러한 분화는 시장이 성장함에 따라 워크로드 다양성이 단일 아키텍처가 감당할 수 있는 수준을 넘어섰고, 특화된 칩은 아키텍처적 특성상 가장 적합한 분야에서 시장을 장악하게 된 것이다.

추론은 이러한 전환의 시작점에 있다. 학습은 당분간 범용 GPU에서 계속될 것이지만 추론은 요구 사항이 매우 다른 여러 작업 부하 유형으로 나뉘고 있다. 음성 추론은 배치 처리와 요구 사항이 다르고, 온디바이스 추론은 하이퍼스케일 서버 처리와 다른 문제이며, 장문 문맥 검색은 단문 생성과 다른 문제이다. 그리고 향후 몇 년 동안 가장 큰 추론 작업 부하 중 하나가 될 것으로 예상되는 에이전트 처리는 이 모든 것과 완전히 다를 것이다.

에이전트는 특정 유형의 하드웨어 특화에 보상을 제공한다. 긴 프롬프트, 짧고 구조화된 출력, 각 단계가 다음 단계를 차단하는 심층적인 순차적 경로, 그리고 채워지지 않는 배치 크기가 그 예이다. 에이전트 실행의 경제성은 경로의 각 단계에서 발생하는 지연 시간에 의해 제한되는데, 밀리초 단위의 디코딩 지연이 전체 작업에 누적되기 때문이다. 지연 시간을 잘못 설정하면 에이전트 제품은 경제성이 없거나 출시할 수 없게 되며, 반대로 제대로 설정하면 새로운 유형의 제품이 실현 가능해진다.

현재 수력 에너지로 구동하는 제너럴 컴퓨트는 미국 내 기존 코로케이션 시설에 구축된 삼바노바 SN40L에서 에이전트 워크로드를 실행하고 있다. GPT-OSS-120B에서 이 스택은 시장에서 가장

2) General Compute, "Inference is fragmenting," May 2026

강력한 GPU 기반 추론 제공업체 중 하나인 투게더 AI보다 첫 토큰 생성 시간(Time-to-First Token)이 2.6배 빠르고, 엔드투엔드 지연 시간은 4.6배 빠르다고 한다. 이는 잘 선택된 단일 실리콘 칩만으로도 잘 설계된 범용 스택 대비 이 정도의 성능 향상을 얻을 수 있음을 보여준다.

2026년 4분기에는 두 가지 핵심 요소를 추가할 예정인데 SN50을 통해 6,000억 개 이상의 파라미터 모델에서 랙 전체 처리량을 SN40L 대비 거의 20배 가까이 향상시킬 것이고, 둘째, 칩 종류를 분산시켜 AMD MI300X가 프리필을 담당하고 SN50은 디코딩에 특화하도록 한다. 이런 변화는 동일한 전제를 갖고 하는 것인데, 에이전트 워크로드는 해당 워크로드에 최적화된 하드웨어에서 최고의 성능을 발휘하며, 이러한 최적화 작업이 이루어질 만큼 시장 규모가 크기 때문이다.

에이전트가 챗봇과 다른 점

2022년부터 2025년까지 업계에서 구축한 추론 인프라는 챗봇을 위해 만들었다. 사용자가 질문을 입력하면 모델이 응답을 생성하고 사용자는 이를 읽는다. 처리량은 많은 사용자를 동시에 응대해야 하므로 중요한 반면, 사람이 실시간으로 출력 결과를 읽기 때문에 지연 시간은 상대적으로 덜 중요했다.

에이전트는 이런 가정을 모두 뒤집는다. 일단 에이전트는 입력값이 많다. 에이전트 프롬프트는 사용자 질문이 아니라 시스템 프롬프트, 도구 카탈로그, 메모리 버퍼, 검색된 컨텍스트, 그리고 이전 단계의 누적 기록이 모두 포함된 것이다. 수만 개의 토큰이 사용되는 것은 일반적이며, 수십만 개에 달하는 경우도 드물지 않다. 모든 단계에서 이러한 컨텍스트의 대부분이 다시 전송되므로, 사전 입력은 일회성 설정이 아니라 지속적인 비용이 발생한다.

두 번째 출력은 간결하고 구조화되어 있다. 에이전트는 장문의 글을 작성하는 대신 도구 호출, JSON 객체, 추론 과정 및 중간 계획을 출력하며, 일반적으로 수백 개의 토큰을 생성한다. 모델은 지속적인 출력을 생성하기보다는 시작 및 중지하는 데 대부분의 시간을 소비한다.

셋째, 작업 부하는 순차적으로 처리된다. 챗봇 대화는 사용자가 읽고 입력하는 동안 자연스러운 멈춤이 있지만, 에이전트의 처리 과정에는 그러한 멈춤이 없다. 한 단계가 완료되는 순간 다음 단계가 시작된다. 사람이 개입하여 지연 시간을 흡수하는 과정이 없기 때문에 디코딩 지연 시간 1밀리초가 전체 작업 시간에 1밀리초씩 추가되고, 이러한 지연 시간은 모든 단계에 걸쳐 누적된다.

넷째, 배치 크기가 작다. 챗봇 서비스 운영 방식은 메모리에서 모델 가중치를 읽어오는 비용을

분산시키기 위해 수십 또는 수백 개의 동시 사용자 요청을 일괄 처리하는 데 의존한다. 에이전트 워크로드는 대개 배치 크기가 1이다. 즉, 가능한 한 빠르게 실행되는 단일의 긴 궤적이며, 함께 처리할 다른 요청이 없다. 따라서 최신 GPU 서비스 스택을 정의하는 처리량 최적화는 적용되지 않는다.

종합하면 에이전트를 처리하는 작업은 긴 입력, 짧은 출력, 단계별 지연 시간에 대한 허용 오차 제로, 수십 단계에 걸쳐 누적되는 작업, 그리고 처리되지 않는 배치 처리 속도라는 특징을 갖는다. 이는 챗봇 서비스 제공과는 다른 문제이며, 단순히 난이도가 높아진 버전이 아니다.

프리필(Prefill)과 디코딩

LLM 추론에 대한 지배적인 접근 방식은 이를 단일 문제로 취급하는데, 토큰이 입력되고 토큰이 출력되며, 빠를수록 좋다는 식이다. 이런 접근 방식은 근본적으로 다른 두 가지 계산의 가중 평균에 최적화된 인프라를 만들어냈고, 이제 그 한계에 부딪히기 시작했다. 즉, 프리필과 디코딩은 하드웨어 수준에서 공통점이 거의 없기 때문이다.

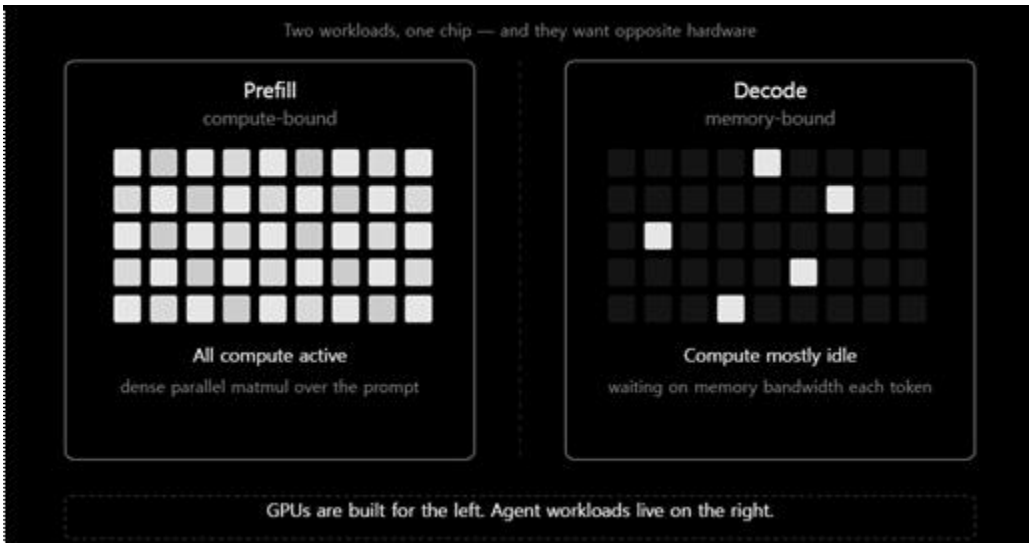


그림 2 프리필과 디코드 처리의 차이 [출처: 제너럴 컴퓨트]

프리필(Prefill)은 연산 집약적인 작업이다. 모델이 긴 프롬프트를 처리할 때, 전체 입력에 걸쳐 병렬로 밀집 행렬 곱셈을 수행하는데, 병목 현상은 FLOPs에서 발생한다. 칩은 넓은 SIMD 대역폭, 높은 연산 강도, 그리고 가중치 스트리밍을 위한 충분한 HBM 대역폭을 필요로 한다. H100이나 MI300X 같은 칩이 프리필 작업을 수행하는 것은 설계된 목적대로 작동하는 것이다.

디코딩은 메모리 제약이 심하다. 생성 과정이 시작되면 모델은 자기회귀 방식으로 토큰을 하나씩 생성하는데, 각 토큰을 생성하려면 모델의 전체 가중치와 전체 KV 캐시를 메모리에서 읽어온 다음, 간단한 연산을 수행해야 한다. 병목 현상은 대역폭, 특히 가중치와 캐시 데이터를 연산 장치로 가져오는 데 걸리는 지연 시간이다. 연산 강도가 급격히 떨어지면서 칩의 FLOPs가 메모리를 기다리며 유휴 상태가 된다. 배치 크기가 1인 디코딩 작업을 수행하는 GPU는 마치 주차장에 서 있는 페라리처럼, 값비싼 실리콘 칩이 거의 아무것도 하지 않는 것과 같다.

에이전트 작업 부하가 증가함에 따라 이러한 불균형은 완화되는 것이 아니라 오히려 심화된다. 긴 프롬프트는 사전 입력 비용을 증가시키고, 작고 지연 시간에 민감한 생성 작업은 디코딩의 상대적 비중을 높인다. 코딩 에이전트가 도구를 20번 호출하면 디코딩 비용을 20배로 지불하게 된다. 사전 입력 속도는 첫 번째 토큰을 얻는 데 걸리는 시간의 하한선을 결정하고, 디코딩 속도는 사용자가 포기하기 전에 에이전트가 수행할 수 있는 단계 수의 상한선을 결정한다.

이에 대한 대응책은 두 가지이다. 첫 번째는 전체 워크로드에 더 적합한 실리콘, 즉 GPU보다 두 단계 모두에 더 적합한 아키텍처를 가진 칩을 선택하는 것이다. 두 번째이자 더 강력한 대응책은 두 단계에 하나의 칩을 사용하는 것을 중단하는 것이다. 프리필은 프리필 전용으로 설계된 칩에서, 디코딩은 디코딩 전용으로 설계된 칩에서 실행하여 각 실리콘 칩이 본래의 용도에 맞게 작동하도록 하는 것이다. 진정한 해결책은 하드웨어를 분산시키는 것이다.

데이터 플로우 아키텍처

LLM의 토큰 생성(디코딩)은 연산 속도보다 메모리에서 데이터를 가져오는 속도(대역폭)가 발목을 잡는 '메모리 병목 문제'를 해결해야 한다. GPU는 고대역폭 메모리(HBM)로 이를 해결하려 하지만, 하드웨어가 발전할수록 연산 능력(FLOPs) 증가 속도가 HBM 대역폭 증가 속도보다 빨라 병목 현상이 오히려 악화되고 있다. 그 결과, 배치 크기가 1일 때 GPU는 성능 저하를 겪게 된다.

외부 메모리에서 데이터를 계속 매번 가져오는 GPU와 달리, 데이터 플로우 유닛을 사용해 컴퓨팅 유닛 바로 옆에 메모리를 분산 배치하여 데이터가 파이프라인을 통해 흐르도록 만들 수 있다. 가중치와 KV 캐시를 최적의 위치에 상주시켜 HBM 대기 시간을 없앴을 수 있는데, 이는 특히 배치 크기 1 디코딩 및 에이전트 워크로드(프롬프트 캐싱 등)에서 압도적인 속도적 우위를 가진다. 실제 환경에서 동일한 120B 모델을 구동했을 때, 삼바노바의 SN40L 스택이 기존 GPU 기반 추론 제공업체보다 약 4.6배 빠른 지연 시간 성능을 보였다고 한다.

특수 추론 칩에 대한 많은 반대 의견은 범용 GPU가 CUDA를 중심으로 한 방대한 소프트웨어 생태계의 이점을 누리는 반면, 특수 목적용 칩은 그렇지 못하다는 주장이었다. 이는 3년 전만 해도 심각한 문제였지만, 지금은 그 정도가 줄어들었다. 예를 들어 삼바노바 스택은 라마, 큐웬, 딥시크, 믹스트랄 및 기타 MoE 변형을 포함하여 에이전트 워크로드에서 실제로 사용하는 모델 제품군을 지원하며, 프로덕션 에이전트 시스템에 필요한 기본 기능(KV 캐시 관리, 프롬프트 캐싱, 구조화된 출력)도 제공한다. 생태계 격차는 하드웨어 격차보다 빠르게 좁혀지고 있으며, HBM 확장 속도가 둔화됨에 따라 하드웨어 격차는 오히려 벌어지고 있다.

대량의 연산이 필요한 사전 데이터 입력(프리필) 단계는 여전히 GPU가 유리하다. 따라서 향후에는 프리필은 GPU가, 빠른 토큰 생성이 필요한 디코딩은 RDU가 담당하는 분산형 하이브리드 스택이 최적의 솔루션이 될 것이다.

마무리하며

제너럴 컴퓨트 같은 추론 클라우드를 제공하고 하는 기업은 범용 클라우드를 지향하는 것이 아니다. 에이전트라는 특정 대규모 워크로드 유형에 최적화된 추론 기반을 구축하는 데 집중하고 있는 것이다.

이를 통해 코딩 에이전트는 도구 호출을 다섯 번에서 서른 번으로 늘릴 수 있으며, 품질 차이는 선형적이지 않게 된다. 이는 코드를 작성하는 시스템과 코드를 구축하고 테스트하는 시스템의 차이와 같다. 연구 에이전트는 문서를 다섯 개에서 쉰 개까지 읽고 요약이 아닌 종합적인 결과를 도출할 수 있게 되며, 음성 에이전트는 처음 떠오르는 그럴듯한 응답을 생성하는 대신, 말하기 전에 무엇을 말할지 추론할 수 있다.

이를 위해서는 에이전트 추론을 챗봇 서비스의 부수적인 작업이 아닌, 핵심적인 작업으로 처리하는 인프라가 필요하다. 그러한 인프라는 GPU 전용 클라우드에는 존재하지 않으며, 기존 하드웨어가 문제의 핵심이 아닌 다른 부분에 최적화되어 있기 때문에 나중에 추가 구축하는 것도 불가능하다.

또한, 유용한 에이전트 인텔리전스의 양을 늘리는 것이 목표라면, 토큰당 지연 시간은 방정식의 절반에 불과하다. 나머지 절반은 새로운 용량이 얼마나 빨리 가동되는지에 대한 일정이다. 내년에 가능한 것과 3년 후에나 가능한 것을 살펴봐야 하는 이유가 여기에 있다.

02 AI 시대, 듀오폴리(Duopoly) 시장과 인재의 양극화

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

89%와 29%: 같은 코인의 양면

AI 산업은 지금 두 개의 숫자 앞에 서 있다. 하나는 89%, 다른 하나는 29%다. 전자는 5월 기준 34개 주요 AI 스타트업이 창출하는 연간 매출 800억 달러 중 앤쓰로픽과 오픈AI 두 회사가 차지하는 비중이다. 6개월 전 대비 4.5%포인트 상승했으며, 격차는 좁혀지지 않고 벌어지고 있다.[3] 후자는 WRITER가 2026년 4월 7일 발표한 2,400명 대상 글로벌 조사 결과다. 직원의 29%가, Z세대에서는 44%가 자신이 속한 회사의 AI 전략을 적극적으로 '도입 거부(sabotage)'하고 있다고 응답했다.[4]

표면적으로 이 두 숫자는 별개의 현상을 가리킨다. 89%는 AI 모델 공급 시장의 듀오폴리(Duopoly) 구조다. 29%는 기업 조직 내부의 저항과 분열이다. 그러나 이번 글의 핵심 논지는, 이 두 **숫자가 같은 코인의 양면이라는 것이다.** AI 시장에서 진행되는 공급 측 양극화와 기업·조직 안에서 진행되는 수요 측 양극화는 분리된 현상이 아니다. **시장 듀오폴리가 가격 결정권과 기능 통합 정책을 통제하는 순간, 그 압력은 기업 도입 격차로 전이되고, 다시 조직 내 인재 양극화로 굴절된다.** 두 회사가 매출의 절반 가까이 차지하면서 연간 300억 달러 이상의 현금을 태우는 모순, PwC의 'AI 경제 가치 74%를 상위 20% 기업이 독점'한다는 분석[5], 임원 75%가 자사 AI 전략을 '보여주기'로 자인하는 자기 진단, IMF가 경고하는 중간 숙련 일자리의 구조적 소멸[6]과 같은 데이터를 따로 읽으면 각각의 산업 보고서이지만, 함께 읽으면 하나의 거대한 구조 재편의 윤곽을 그린다.

3) The Information, “Anthropic and OpenAI’s Share of AI Startup Revenues Rises to 89%”, May 17, 2026
4) Fortune, “Gen Z workers are so fearful AI will take their job they’re intentionally sabotaging their company’s AI rollout”, Apr 8, 2026
5) PwC, “Three-quarters of AI’s economic gains are being captured by just 20% of companies”, Apr 13, 2026
6) IMF, “Bridging Skill Gaps for the Future: New Jobs Creation in the AI Age”, Jan 14, 2026

이 인과 사슬은 기업 의사결정자에게 두 가지 결정의 본질을 재정의한다. 첫째, AI 벤더 의존 구조는 단순한 조달 결정이 아니라 향후 5년의 비용 구조와 가격 결정권을 좌우하는 전략적 선택이다. 둘째, 조직 내 AI 도입은 단순한 도구 배포가 아니라 '슈퍼유저'와 '도입 거부'를 동시에 만들어내는 양극화 메커니즘이다. 듀오폴리의 가격 인상이 기업 마진을 압박하고, 압박받는 마진이 인력 구조조정으로 전가되며, 구조조정의 공포가 다시 '도입 거부'를 낳는 순환 구조는, 추상적 우려가 아니라 이미 작동하고 있는 현실이다.

1. 듀오폴리의 도래

34개 주요 AI 스타트업이 창출하는 연간 매출은 약 800억 달러, 월 환산 66억 달러에 달하며, 6개월 전 대비 112% 증가했다. 이 성장률도 놀랍지만 더 주목할 것은 분배 구조다. 800억 달러 중 약 712억 달러, 즉 89%가 앤쓰로픽과 오픈AI 두 회사에 집중되어 있다. 6개월 전 동일 그룹에서 이 두 회사의 점유율은 84.5%였다. 반년 만에 4.5%포인트가 더 증가한 것이다.

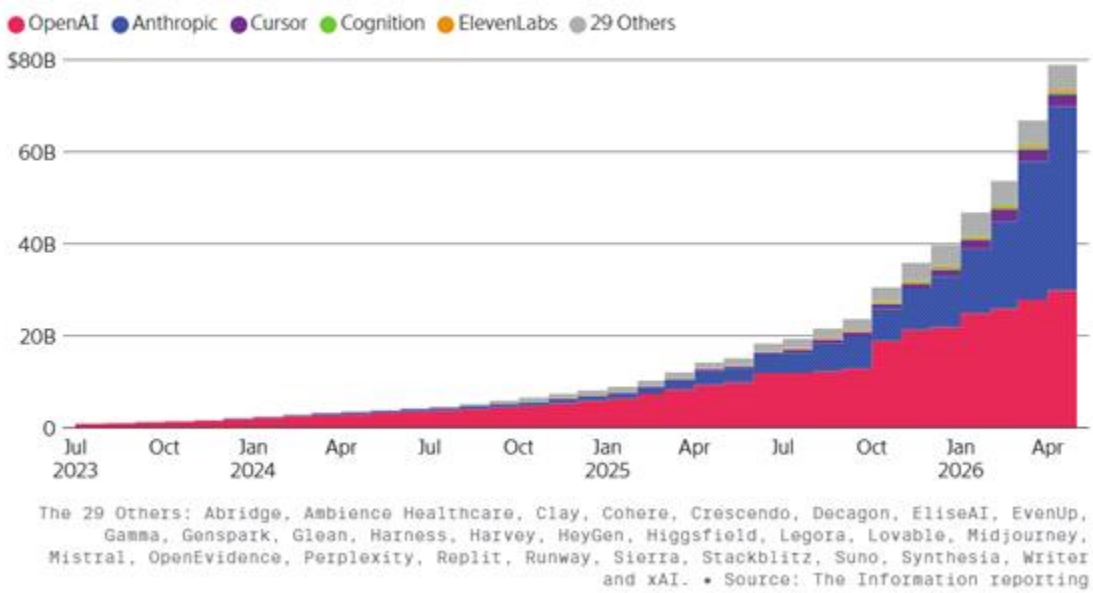


그림 3 34개 주요 AI 스타트업의 연간 매출 현황 (출처: The Information)

이 집중도는 통상적인 시장 지배의 범주를 벗어난다. 일반적으로 과점 시장은 상위 4개 기업의 점유율(CR4)이 60~80% 수준에서 형성되며, 듀오폴리로 분류하기 위해서는 상위 2개 기업의 점유율이

70%를 초과해야 한다는 것이 산업조직론의 일반적 기준이다. AI 모델 시장의 89%는 이 기준을 압도적으로 상회한다. 더 결정적인 것은 추세다. 시장이 성숙하면서 집중도가 완화되는 일반적 패턴과 반대로, AI 시장은 시장이 커질수록 집중도가 심화되는 역행적 동학(Counter-cyclical Dynamic)을 보이고 있다. 듀오폴리 내부에서는 앤쓰로픽의 매출이 최근 오픈AI를 추월한 것으로 나타나며, 그 동력은 코딩 작업과 화이트칼라 업무 영역에서의 강세에서 비롯된다. 두 회사 사이의 경쟁이 치열할수록 나머지 32개 회사에 대한 지배력은 오히려 강화되는 역설적 구조다.

동시에 무너지는 두 가지 통념: "AI 앱 가치" vs. "스타트업의 다양성"

이 매출 분포는 두 가지 통념을 동시에 무너뜨린다. 첫 번째는 'AI 시대에는 모델보다 애플리케이션 레이어에서 가치가 발생한다'였다. 인터넷 시대의 학습에 기반한 이 SaaS 업계의 주류 의견은 매출 데이터로 반증이 되고 있다. 일부 투자자들이 주장해온 명제, '현 AI 시대 소프트웨어 가치의 대부분은 첨단 AI 모델 개발자에게 귀속되며, 순수 AI 앱 개발자에게 가는 가치는 제한적이다'가 입증되고 있다.

두 번째는 'AI 스타트업 생태계가 시장의 다양성을 만든다'이다. 800억 달러 매출 그룹의 32개 회사가 거의 전적으로 두 회사의 모델에 의존하고 있다는 사실은, 생태계의 다양성이 곧 공급의 다양성을 의미하지는 않는다는 점을 드러낸다. 시장에 보이는 다양성은 인터페이스와 사용 사례의 다양성일 뿐, 그 밑에 깔린 모델 공급은 듀오폴리 구조 위에 서 있다.

이 의존 구조의 대가는 마진으로 나타난다. 커서의 사례는 상징적이다. 2026년 1월 마감 분기 기준 커서의 매출총이익률(Gross Margin)은 마이너스 23%를 기록했다. 연 매출 10억 달러를 돌파한 스타트업으로서는 이례적인 수치다. 이후 이익률이 개선되었으나, 이 데이터가 보여주는 것은 명확하다. 듀오폴리의 모델을 빌려 사업을 운영하는 구조는 매출이 아무리 빠르게 성장해도 수익성으로 전환되기 어려운 구조적 한계를 갖는다. 32개 비듀오폴리 회사들은 집합적으로 매년 수십억 달러를 앤쓰로픽과 오픈AI에 모델 사용료로 지불하고 있으며, 커서·퍼플렉시티·일레븐랩스·코그니션 등의 매출 중 상당 부분이 결국 듀오폴리의 매출로 다시 흘러 들어간다. 시장이 커질수록 듀오폴리의 매출은 직접 매출과 간접 매출(스타트업을 통한 흐름) 두 방향에서 동시에 증가하는 이중 계산 구조다.

앤쓰로픽 가격 인상이 만든 연쇄 효과, 가격 결정권

매출 점유율 89%의 진정한 무게는 현재의 매출이 아니라 미래의 가격 결정권에 있다. 앤쓰로픽의 최근 가격 인상으로 세일즈포스, 서비스나우를 포함한 주요 엔터프라이즈 소프트웨어 기업과 32개 AI

스타트업이 동시에 비용 압박을 받고 있다. 앤쓰로픽과 오픈AI는 연간 300억 달러 이상의 현금을 소진하고 있으며, 두 회사의 가격 정책은 수익성 확보가 아니라 손실 폭 축소를 위한 압박 수단으로 작동하고 있다. 그 압박은 시장 전체로 전가된다. 앤쓰로픽의 매출은 AWS·구글과 분점되고 오픈AI는 2030년까지 마이크로소프트와 매출의 20%(2026년에만 약 60억 달러)를 분점하는 구조, 그리고 엔비디아가 듀오폴리 두 회사에 합산 약 400억 달러를 투자한 자본 순환 구조가 결합하면, AI 시장은 엔비디아 → 클라우드 빅3 → 듀오폴리 → 32개 스타트업 → 엔터프라이즈 고객이라는 수직적으로 통합된 가격 전가 체인을 형성한다. 체인의 어느 한 단계에서 가격이 인상되면 압력은 하방으로 전가되며, 종착점은 항상 최종 고객이다. 이 가격 결정권의 이동은 본질적 질문을 제기한다. 우리가 오늘 산정한 AI 도입 예산은 2년 후에도 유효한가? 앤쓰로픽의 가격 인상이 단발성 사건이 아니라 듀오폴리 구조에서 반복될 패턴이라면, 기업의 AI TCO는 현재 추정치보다 상당히 높은 상방 변동성을 갖는다고 평가해야 한다.

2. 네오랩 현상: 듀오폴리를 우회하려는 자본의 시도

89%의 매출 집중도는 시장의 끝이 아니라 다음 국면을 예고하는 신호다. 듀오폴리가 굳어질수록 그 구조를 우회하려는 자본의 시도도 가속된다. 시장에서 관심있는 바라보는 네오랩(Neolab)의 정의는 단순하다. 앤쓰로픽과 오픈AI를 정면 추격하기보다 **새로운 종류의 AI 모델을 개발해 우회하려는 시도다.** 이 정의 자체가 듀오폴리의 깊이를 역설적으로 보여준다. 기존 LLM 아키텍처 위에서는 두 회사를 따라잡기 어렵다는 시장의 판단이, 다른 종류의 모델을 개발하려는 자본 배치로 이어진 것이다.

미스트랄 AI의 궤적은 네오랩 현상의 가장 가시적인 사례다. 2023년 파리에서 창업한 이 회사는 2025년 9월 ASML이 주도한 17억 유로 시리즈 C 라운드를 유치하며 110억 7,000만 유로 기업가치로 평가받았다. ASML이 단독으로 13억 유로를 투자해 최대 주주가 된 이 거래는 단순한 재무 투자가 아니라, AI 모델·반도체 장비·유럽 산업 정책이 하나의 전략적 축으로 결합되었다는 의미다. 미스트랄은 2026년 3월 추가로 8억 3,000만 달러의 부채 금융을 확보해 파리 인근에 1만 3,800개의 엔비디아 칩을 활용한 자체 데이터센터를 구축하기 시작했다.[7] 미스트랄의 연간 반복 매출(ARR)은 2024년 말 약 1,600만 달러에서 2026년 1월 4억 달러로 1년 만에 약 20배 확장되었으며, 매출의 60%가 유럽에서 발생한다는 점은 그 성장이 지정학적 자율성과 데이터 주권에 대한 유럽 기업·정부의 수요에 기반하고 있음을 보여준다.

7) CNBC, “Mistral secures $830 million in debt financing to fund AI data center”, Mar 30, 2026

미스트랄 외에도 자본 흐름은 광범위하다. 영국의 자율주행 스타트업 웨이브(Wayve)는 12억 달러, AI 데이터센터를 제공하는 엔스케일(Nscale)은 20억 달러, 프랑스의 AMI 랩스는 10억 달러를 2026년에 조달했다. 이 규모들은 듀오폴리의 누적 투자액(오픈AI 1,800억 달러, 앤쓰로픽 590억 달러)과 비교하면 여전히 작지만, 방향성은 명확하다. 자**본은 듀오폴리에 더 많이 들어가는 동시에, 그 우회 경로에도 분산되고 있다.** 네오랩들은 듀오폴리가 수평적으로 확장할 때 침투하기 어려운 데이터 주권을 요구하는 유럽 정부와 기업, 자율주행, 도메인 특화 인프라와 같은 영역에서 모델 자체를 차별화하는 전략으로 성장하고 있다.

3. 시장의 양극화가 기업의 양극화로

89%의 매출 집중과 네오랩의 우회 시도가 공급 측 양극화의 두 얼굴이라면, 그 압력을 받는 수요 측에서도 평행한 양극화가 진행되고 있다. AI 도구는 시장에 광범위하게 공급되고 있으나, 그 도구로부터 실질적 가치를 추출하는 능력은 극히 일부 기업에 집중되어 있다. 이 격차는 단순한 도입 시기의 차이가 아니라, 도구를 가치로 전환하는 통합 역량의 구조적 차이에서 비롯된다.

상위 20% 기업이 AI 경제 가치 74%를 독점

PwC 보고서는 AI가 만들어내는 경제 가치의 74%를 상위 20% 기업이 가져간다는 사실을 정량화했다. AI 리더 기업은 다른 기업보다 책임 있는 AI 프레임워크를 보유할 확률이 1.7배 높았고, 결정적으로 이 리더 기업의 직원들은 AI 출력물을 신뢰할 확률이 2배 높았다. **AI 가치 격차의 원인이 모델 접근성이나 기술 인프라가 아니라, 조직적 통합 역량과 거버넌스 성숙도라는 사실이 입증된 것이다.** 모든 기업은 동일한 앤쓰로픽·오픈AI API에 접근할 수 있고, 미스트랄의 오픈-웨이트 모델은 누구나 다운로드할 수 있다. 그러나 동일한 모델을 도입한 두 기업이 만들어내는 가치는 4배 가까이 차이 난다. AI는 기술이 아니라 변경 관리의 문제라는 명제가 매출 데이터로 입증된 것에 가깝다.

리더 기업들은 AI를 생산성 향상 도구가 아니라 성장과 사업 재창조의 촉매로 활용한다. 동일한 AI 도구가 한 기업에게는 백오피스 자동화 도구이고, 다른 기업에게는 시장 재정의의 무기다. PwC는 이 격차가 학습 속도의 차이에 기인한다고 지적한다. 리더 기업이 더 빠르게 학습하고, 검증된 사용 사례를 더 빠르게 확장하며, 의사결정을 더 안전하게 자동화하기 때문에, 격차는 자기 강화적으로 확대된다.

도입과 전환의 결정적 차이

센드투팀(SendToTeam)의 리포트에 따르면, 78%의 기업이 AI를 도입했으나 실제 전환을 달성한

기업은 34%에 불과하다.[8] 44%포인트의 격차는, 도구의 보유가 곧 가치 창출을 의미하지 않는다는 사실을 정량화한다.

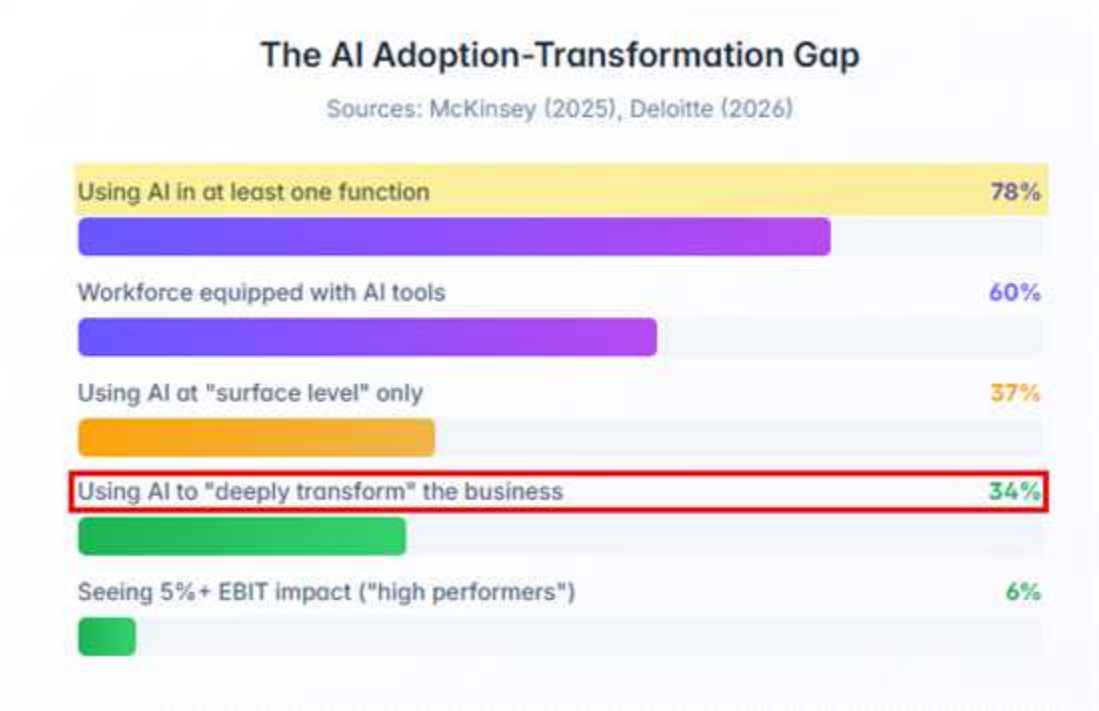


그림 4 AI를 도입과 실제 전환을 달성한 기업의 차이 (출처: SendtoTeam)

딜로이트의 리포트도 동일한 패턴을 확인한다. 2026년 기업의 직원 AI 접근권은 2025년 대비 50% 증가했고, 외형적으로 AI 도입은 빠르게 확산되고 있다. 그러나 진정으로 사업을 AI로 구조화하는 기업은 34%에 불과하다.[9] 도구를 더하는 것과 프로세스를 재설계하는 것은 다른 차원의 작업이다. 전자는 기존 조직 구조 위에 AI 레이어를 얹는 일이고, 후자는 조직의 의사결정 흐름·역할 정의·성과 평가 기준 전체를 재구성하는 일이다. 인간-AI 협업 팀이 인간만으로 구성된 팀 대비 60% 더 높은 생산성을 보이고, 에이전트 AI 도입 기업이 매출은 6~10% 증가하고 70% 비용 절감을 보고하는 결과는 워크플로우 재설계가 동반될 때에만 실현된다. 도구를 더하는 기업은 1년 후 정체에 빠지지만, 프로세스를 재설계하는 기업은 3년 차에 AI 네이티브 워크플로우에 도달하며 이 시점에서 두 그룹의 격차는 봉합이 불가능해진다.

8) SendToTeam, “AI Workforce Trends 2026: 7 Data-Backed Shifts”, Mar 23, 2026
9) Deloitte, “The State of AI in the Enterprise”, Jan, 2026

4. 인재의 양극화: 슈퍼유저와 '도입 거부'의 공존

기업 간 양극화가 도구의 사용 방식에서 비롯된다면, 그 도구를 직접 사용하는 인간 차원에서도 평행한 양극화가 진행된다. 같은 조직 안에서 한 직원은 AI 슈퍼유저로 거듭나며 5배의 생산성을 기록하고, 다른 직원은 회사의 AI 전략을 적극적으로 '도입 거부'한다. **AI 시대의 인재 양극화는 기술의 문제가 아니라 신뢰의 문제다.**

슈퍼유저: 5배의 생산성, 3배의 승진

WRITER 조사에서 AI 슈퍼유저는 일반 직원 대비 5배의 생산성을 보였고, 지난 1년간 승진과 임금 인상을 동시에 받을 확률이 약 3배 높았다.[10] 이 격차는 임금 시장에서 가격화된다. 동일 직무에서 고급 AI 스킬을 보유한 노동자는 그렇지 않은 동료보다 56% 높은 임금을 받는다.[11] 가트너의 2026년 1분기 조사는 슈퍼유저 우위의 원천을 분해한다.[12] 다수의 사용 사례에서 AI를 활용하는 직원은 단일 사용 사례에 머무는 직원 대비 2~3배 더 높은 생산성·품질·프로세스 개선 성과를 내며, AI에 긍정적 관점을 가진 직원은 부정적 관점을 가진 직원 대비 3.4배 더 생산적이다. **슈퍼유저의 우위는 단지 더 많은 시간을 AI에 투입한 결과가 아니라, 다양한 맥락에서 AI를 사용하는 폭과 AI에 대한 태도라는 두 변수의 상호작용에서 나온다.**

그러나 개인의 우위가 조직의 성과로 자동 전환되지는 않는다. 조직 차원에서 생성 AI로부터 유의미한 ROI를 실현한 비율은 29%, AI 에이전트의 경우 23%에 그친다. 슈퍼유저는 5배의 가치를 만들지만, 그 가치가 조직 전체로 확산되지 못하는 구조 안에서, 나머지 직원들의 저항이 슈퍼유저의 우위를 조직의 성과가 아닌 개인의 격차로 고정시킨다.

29%의 '도입 거부', 그 중 44%의 Z세대

WRITER 조사의 가장 충격적인 발견은 슈퍼유저의 반대편에 있다. 직원의 29%가, Z세대에서는 44%가 적극적으로 'AI 도입 거부'하고 있다고 응답했다. 구체적 양상은 공식 AI 도구 사용 거부, 승인되지 않은 외부 AI 도구 사용, 회사 기밀 정보를 공개 AI 도구에 입력, 의도적인 저품질 결과물 생성 등으로 분류된다.

---

10) Writer, “The invisible crisis in enterprise AI”, Apr 2026
11) Gloat, “10 Key AI Workforce Trends In 2026”, Mar 4, 2026
12) Gartner, “50% of Enterprises Without a People-Centric AI Strategy Will Lose Their Top AI Talent”, May 13, 2026

이 저항을 단순한 세대 갈등으로 해석하는 것은 표면적 분석이다. '도입 거부'를 인정한 직원 중 30%가 AI 자동화로 인한 일자리 우려를 사유로 들었다. 무용지물이 될 것에 대한 공포는 추상적 불안이 아니라 실제 데이터에 근거한다. 2025~2026년 미국에서만 16만 5,000명 이상의 기술직 정리해고가 있었고, 2026년 3월에는 추적 시작 이래 처음으로 AI가 정리해고 1위 사유가 되었다. WRITER 조사에 응한 임원의 69%가 AI 관련 인력 감축을 계획하고 있다.

그러나 **'도입 거부'의 더 결정적인 원인은 따로 있다.** '도입 거부'를 인정한 직원 중 26%가 회사의 빈약한 AI 전략을 사유로 지목했다. 그리고 임원의 75%가 자사의 AI 전략이 실질적 가이드가 아닌 "보여주기"라고 인정했다. 직원의 '도입 거부'가 비합리적 저항이 아니라, 전략의 부재에 대한 합리적 반응일 수 있다는 가능성을 임원 자신이 인정한 셈이다. 전략의 부재가 강제 도입을 낳고, 강제 도입이 '도입 거부'를 낳고, '도입 거부'가 다시 ROI 실패를 낳고, ROI 실패가 다시 전략이 보여주기에 불과하다는 인식을 강화하는 순환 구조 안에서, 슈퍼유저의 5배 생산성은 조직 성과로 전환되지 못한다. 부수적 위험도 정량화되어 있다. 임원의 67%가 자사가 승인되지 않은 AI 도구로 인해 데이터 유출 또는 침해를 이미 경험했다고 믿으며, 36%의 조직은 AI 에이전트 감독을 위한 공식 계획이 없다. '도입 거부'는 단순한 생산성 손실이 아니라 보안 사고와 거버넌스 실패로 직접 연결되는 위험이다.

5. 듀오폴리는 어떻게 인재 양극화로 나타나는가?

지금까지 이야기한 89%의 매출 집중, 74%의 가치 독점, 5배의 슈퍼유저 생산성, 29%의 '도입 거부'는 각각 다른 산업 보고서에서 다른 시점에 발표된 데이터들이다. 그러나 이 데이터들 안에서는 인과 사슬이 보인다. **시장의 듀오폴리 구조가 가격 결정권을 통해 기업 마진을 압박하고, 압박받는 마진이 인력 결정을 강요하며, 강요된 결정이 다시 조직 내 신뢰 구조를 재편한다.**

가격 인상 → 마진 압박 → 인력 구조조정

앤쓰로픽의 가격 인상은 매출의 89%를 점유한 듀오폴리의 일상적 경영 의사결정이지만, 그 효과는 32개 AI 스타트업과 주요 SaaS 기업들에게 동시에 전파된다. 모델 사용료 인상으로 매출 원가가 증가하면, 같은 매출 성장률을 유지하기 위해 가장 빠르고 가시적인 절감이 가능한 인건비에서 절감을 찾게 된다. 특히 AI 도입으로 자동화가 가능해진 직무들은 이중적 압력에 노출된다. 그 직무 자체를 자동화함으로써 비용을 줄이는 동시에, 자동화의 정당성을 AI 도입의 가시적 성과로 제시할 수 있기 때문이다.

AI 시대의 정리해고는 경기 침체에 따른 일시적 조정이 아니라, **AI 도입 자체가 인력 감축의 근거가 되는 새로운 종류의 구조조정이다**. 도미노의 종착점은 노동자 개인의 심리다. 임원의 69%가 AI 관련 인력 감축을 계획하는 조직 안에서, 직원의 64%가 AI 전환 실패로 인한 일자리 상실을 두려워한다고 답한 것은 같은 현실의 두 얼굴이다. 그리고 이 공포가 다시 '도입 거부'의 동력이 된다. 도미노는 듀오폴리에서 시작해 개별 직원의 책상에서 끝나지 않고, 그 책상에서 다시 조직으로 역방향의 압력을 만들어낸다.

한국 기업의 위치: 통합 역량 격차의 의미

이번 글에서 분석한 인과 사슬은 글로벌 데이터에 기반하지만, 한국 기업에게도 직접적이고 구체적인 함의를 갖는다. 국내 대기업의 AI 도입률은 글로벌 평균을 빠르게 따라가고 있다. 그러나 **도입률은 양극화의 결정 변수가 아니다. 결정 변수는 도입한 AI가 조직 안에서 작동하는 방식이며, 그 방식은 신뢰 환경과 거버넌스 성숙도에 의해 좌우된다.**

한국 기업이 직면한 위험은 두 갈래로 나뉜다. **첫 번째는 도입 속도와 통합 역량의 격차, 그리고 가격 결정권 부재의 결합이다.** 글로벌 리더 기업들이 책임 있는 AI 프레임워크, 교차기능 거버넌스 위원회, AI를 사업 재창조의 촉매로 활용하는 전략을 동시에 구축하는 동안, 도입 자체에 집중하는 한국 기업은 리더의 조건에서 격차가 생길 수 있다. 동시에 국내 대기업의 AI 도입은 상당수가 SI 기업을 통해 진행되며, SI 기업은 다시 마이크로소프트·구글·SAP 등 글로벌 플랫폼 위에서 솔루션을 구축하고, 그 위에서 작동하는 모델은 다시 앤쓰로픽과 오픈AI의 API다. 결과적으로 한국 기업이 AI 도입에 지출하는 비용은 국내 SI 마진 + 글로벌 통합 SaaS 마진 + 듀오폴리 모델 사용료의 삼중 구조로 분배되며, 한국 기업이 가격 결정권을 갖는 부분은 국내 SI 마진 한 층에 불과하다.

두 번째 위험은 세대 간 신뢰 격차의 한국적 양상이다. WRITER 조사의 29%·44% 비율을 한국 시장에 단순 적용할 수는 없지만, **한국 조직 문화에서 '도입 거부'는 직접적 거부보다는 명시적 순응과 암묵적 비협조의 형태로 나타날 가능성이 높다.** 한국 기업의 AI 거버넌스가 이러한 양상을 감지하고 대응할 수 있는 수준에 도달했는지는 별도의 검토가 필요한 주제다.

이 두 가지 위험은 절망적 진단이 아니라, 한국 기업이 향후 12개월 안에 통합 역량과 거버넌스 성숙도에 집중적으로 투자할 합리적 근거다. 격차를 좁히는 것은 더 많은 도구를 사는 일이 아니라, 조직의 신뢰 환경을 재설계하는 일이다.

6. 리더십을 위한 함의: 양극화를 어떻게 대응할 것인가?

두 양극화는 분리된 현상이 아니라 같은 구조의 두 표현이다. C-레벨 임원과 전략기획 담당자는 향후 12개월 동안 두 가지 핵심 의사결정 — 벤더 전략과 조직 전략 — 의 본질을 재정의해야 한다.

벤더 전략: 듀오폴리에 의존하되, 의존성을 관리하라

듀오폴리를 회피해야 하는가, 받아들여야 하는가의 이분법은 잘못 설정된 질문이다. 89%의 시장 점유율과 모델 품질을 고려하면 회피는 현실적 선택이 아니지만, 가격 체인을 그대로 받아들이면 AI TCO는 통제 불가능한 변동성에 노출된다. 올바른 질문은 **의존하되 의존성을 어떻게 관리할 것인가이다.**

세 가지 원칙이 작동한다. **첫째, 모델 추상화 레이어의 구축이다.** 애플리케이션과 모델 공급자 사이에 교체 가능한 인터페이스를 두어, 앤쓰로픽의 클로드, 오픈AI의 GPT, 미스트랄, 구글의 제미나이 등 여러 모델을 동일한 인터페이스로 호출할 수 있도록 설계하면, 가격·품질·정책 변동에 대응할 유연성이 확보된다. 4월 말 마이크로소프트와 오픈AI의 독점 종료는 듀오폴리 내부에서조차 다변화가 진행되고 있음을 시사한다. **둘째, 한국산 모델을 포함한 네오랩 모니터링이다.** 주요 네오랩의 ARR 성장, 자본 조달, 기업 고객 확보 현황을 분기 단위로 모니터링해야 한다. 듀오폴리에 대한 협상력은 대안의 존재를 인지하고 있는가에 의해 결정된다. **셋째, TCO의 상방 변동성 반영이다.** 듀오폴리가 연간 300억 달러 이상의 현금을 소진하는 구조 안에서 가격 인상은 반복될 수 있으므로, 보수적 시나리오로 연 15~30%의 모델 사용료 인상을 가정한 TCO 모델을 병행 운영하고 이 시나리오에서도 ROI가 유지되는 영역에 우선 투자해야 한다.

슈퍼유저 양성과 '도입 거부' 관리는 같은 문제다

슈퍼유저-'도입 거부' 양극화는 통상 별개의 HR 과제로 다뤄진다. 슈퍼유저 양성은 교육 프로그램의 문제로, '도입 거부' 관리는 보안과 컴플라이언스의 문제로 분리된다. 이 분리가 잘못되었다. **두 현상은 같은 신뢰 환경의 두 결과이며, 따라서 두 가지를 분리해서 다루는 한 두 가지 모두 해결되지 않는다.**

네 가지 함의가 도출된다. **첫째, AI 도입의 첫 단계는 도구 배포가 아니라 신뢰 환경 진단이어야 한다.** 손상된 신뢰 환경 위에 AI 도구를 배포하면, 슈퍼유저는 등장하지 않고 '도입 거부'만 증가한다. **둘째, AI 전략의 투명성이다.** 효과적인 AI 전략은 어떤 직무가 어떻게 변할 것인가, 어떤 새로운 역할이 만들어질 것인가, 그 전환에서 회사가 어떤 약속을 하는가에 대한 구체적이고 검증 가능한 내용을

포함해야 한다. 추상적 안심 메시지는 효과가 없다. **셋째, 슈퍼유저의 가치를 조직 자산으로 전환하는 메커니즘이다.** 슈퍼유저가 발견한 효과적 사용 사례를 조직 차원에서 수집·검증·확산하는 시스템과, 시간 절약이 아닌 사용 사례의 깊이와 다양성을 추적하는 지표가 필요하다. **넷째, '도입 거부' 대응의 본질은 처벌이 아니라 원인 제거여야 한다.** 관리의 본질은 공식 도구의 효율성 향상과 직원의 신뢰 회복이지, 처벌과 통제 강화가 아니다.

향후 12개월의 관찰 지점

두 양극화의 미래는 결정되어 있지 않다. 의사결정자들이 추적해야 할 네 가지 지표를 제안한다. **첫째, 앤쓰로픽·오픈AI의 매출 점유율 변화다.** 89%가 더 상승한다면 듀오폴리 구조는 굳어지는 방향이고, 정체하거나 하락한다면 시장 재편의 신호다. **둘째, 네오랩의 자본 흐름과 ARR 성장률이다.** 미스트랄의 ARR이 12개월 안에 10억 달러를 돌파한다면, 듀오폴리 외 대안이 의미 있는 시장 점유율을 확보하기 시작한 신호다. **셋째, AI 관련 정리해고 추이다.** 2026년 3월 AI가 정리해고 1위 사유가 된 사건이 일회성 통계인지 구조적 추세인지가 향후 노동 시장 양극화의 핵심 변수다. **넷째, 상위 20% 기업의 가치 점유율 변화다.** PwC가 측정한 74%가 더 상승한다면 신뢰 자본의 자기 강화 동학이 가속되는 신호다.

마지막 메시지는 다음과 같다. **AI 시대의 양극화는 기술이 만들어내는 것이 아니라, 기술 앞에서 우리가 내리는 결정이 만들어낸다.** 듀오폴리에 대한 의존을 어떻게 관리할 것인가, 슈퍼유저의 가치를 어떻게 조직 자산으로 전환할 것인가, '도입 거부'의 원인을 처벌이 아니라 신뢰 회복으로 해결할 것인가? 이 세 가지 결정의 누적이 향후 한국 기업의 위치를 결정한다. **도구는 누구나 살 수 있다. 그러나 그 도구로부터 가치를 추출하는 능력은 조직이 매일 내리는 작은 결정들의 누적에서 만들어진다.**

03 2026년 클라우드 솔루션 리포트: API 관리

| 정채상 KAIST

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

들어가며: MCP는 표준이 되기 전에 섀도우 IT가 됐다

20여 년 전 회사 PC를 운영하던 IT 부서의 풍경은 어수선했다. 직원들은 검색 한 번으로 찾은 사이트에서 화면보호기, 무료 코덱, 출처 모를 유틸리티의 실행 파일(.exe)을 받아 설치했고, 그 안에 무엇이 들어 있는지를 회사가 추적할 방법은 거의 없었다. 사태가 정리된 것은 백신이 좋아져서가 아니라, "어떤 실행 파일이 어떤 경로로, 어떤 권한을 가지고 회사 안에 들어오는지"를 한 점에서 통제하는 엔드포인트 관리·그룹 정책·소프트웨어 카탈로그라는 통제점(Control Plane)이 자리 잡았기 때문이었다.

2026년 5월의 풍경은 그 시절과 닮았다. 개발자는 깃허브와 npm 레지스트리에서 별점이 높은 순으로 모델 컨텍스트 프로토콜(MCP, Model Context Protocol) 서버를 골라 회사 데이터에 연결한다. 그러나 별점은 인기의 신호일 뿐 안전의 보증이 아니고, 부풀리기도 어렵지 않다. 정작 AI 에이전트는 자기가 호출한 도구가 어디서 왔는지조차 알지 못한다. 퀄리스(Qualys)는 2026년 3월 보고서에서 MCP 서버를 "AI 시대의 새로운 섀도우 IT(Shadow IT)"라고 표현했다. 비유처럼 들리지만 실은 정확한 진단이었다.[13] MCP는 2024년 말 앤트로픽이 공개한 이후 1년 반 만에 활성 공개 서버가 1만 개를 넘겼고, 오픈AI, 구글 딥마인드, 마이크로소프트가 이 표준을 채택하며 사실상의 산업 표준으로 자리 잡았다. 동시에 받아쓰기가 너무 쉬운 탓에 IT 부서는 누가 무엇에 어떻게 연결하는지 더는 추적하지 못한다.

추적하지 못하는 이 사각지대가 얼마나 넓은지는 수치가 말해 준다. 노스틱(Knostic)의 공개 인터넷 스캔에 따르면 1,800개 이상의 MCP 서버가 인증 없이 동작했고[14], 애스트릭스(Astrix) 분석에 따르면

13) Qualys. "MCP Servers: The New Shadow IT for AI in 2026." Qualys Blog, 2026.03.19. https://blog.qualys.com/product-tech/2026/03/19/mcp-servers-shadow-it-ai-qualys-totalai-2026
14) Hou, J., et al. "Securing the Model Context Protocol (MCP): Risks, Controls, and Governance." arXiv:2511.20920, 2026. https://arxiv.org/html/2511.20920v1

MCP 서버의 53%가 정적 시크릿(Static Secrets, 코드나 환경 변수에 고정된 장기 토큰)에 의존했다. 인증 없이 노출된 서버, 장기 토큰에 의존하는 연결, 정식 승인 없이 운영에 자리 잡은 의존성. 모두 '누가 무엇에 연결돼 있는지'를 통제하지 못해 생기는 문제이고, 통제점이 들어서야 비로소 드러나고 막을 수 있다.

성격이 다른 위협도 있다. 패키지 자체가 공급망 경로로 들어와 승인된 서버 안에서 작동하는 사고다. 한 번 승인한 서버라도 업데이트가 일어날 때마다 다시 점검하지는 않기 때문에, 신뢰하던 패키지의 업데이트 한 번이 그대로 위협으로 바뀐다. mcp-remote 패키지의 원격 코드 실행 취약점(CVE-2025-6514)은 43만 7,000건의 다운로드 환경에 영향을 미쳤고, 주간 다운로드 1,500건에 이르던 인기 패키지 포스트마크(Postmark) MCP 서버는 사용자가 발송하는 모든 이메일을 공격자에게 BCC로 조용히 전달했다. 마이크로소프트 365 코파일럿의 CVE-2025-32711은 숨겨진 프롬프트로 데이터 유출을 가능하게 했고, 아사나(Asana) MCP의 버그는 서로 다른 고객 인스턴스 사이에 데이터를 노출시켰다. 이 부류는 승인된 서버 내부에서 일어나기 때문에, 뒤에서 보듯 게이트웨이를 통과시키는 것만으로는 막히지 않는다.

MCP를 정의한 앤트로픽 2026년 4월 로드맵도 같은 공백을 자기 고백처럼 적었다. 엔터프라이즈 준비 상태를 4가지 우선순위 중 하나로 꼽으면서도, 그것이 '아직 정의가 가장 덜 된 영역'이라고 인정했다. 무엇이 부족한지는 감사 추적, SSO(Single Sign-On) 통합, 게이트웨이 동작, 설정 휴대성이라는 항목으로 정리했다.[15] 표준의 세부가 정해지기도 전에 광범위한 채택이 먼저 일어난 것이다.

20여 년 전과 다른 점은 통제점이 어디에 들어설지가 비교적 분명하다는 점이다. 그 자리에 들어서고 있는 것이 API 게이트웨이다. 4월호에서 다룬 코드형 인프라(IaC)가 인프라를 코드로 정의하는 계층이었고 5월호의 쿠버네티스가 워크로드의 운영체제였다면, API 게이트웨이는 그 위에서 움직이는 AI 트래픽이 반드시 통과하는 관문이다. 이 관문을 두고 6개 벤더가 경쟁한다. 이번 글에서 다룰 벤더는 Kong, 아피지(Apigee, 구글 클라우드), 아마존 베드락 에이전트코어(AgentCore), 뮬소프트(MuleSoft, 세일즈포스), 타이크(Tyk), 포스트맨(Postman)이다. 모두가 "MCP 지원"을 발표한 2026년에 정작 중요한 기준은 지원 여부가 아니라, 이미 퍼진 섀도우 MCP를 어떻게 게이트웨이의 통제 아래로 들이는가에 있다.

15) Model Context Protocol. "The 2026 MCP Roadmap." Model Context Protocol Blog, 2026.04. https://blog.modelcontextprotocol.io/posts/2026-mcp-roadmap/

2026년 API 관리의 세 가지 큰 흐름

섀도우 MCP의 실상

MCP 서버는 깃허브와 npm에서 받아 즉시 실행할 수 있는 가벼운 자바스크립트, 파이썬 프로세스인데, 이 가벼움이 폭발적 성장의 동력이면서, IT 부서가 연결 관계를 파악하지 못하는 원인이기도 하다. 퀄리스 보고서는 회피 패턴을 세 가지로 정리한다. 감시가 닿지 않는 높은 번호의 포트를 무작위로 잡아 로컬에서 도는 사례, IDE 플러그인과 개발자 도구에 내장돼 IT의 시야 밖에 머무는 사례, 테스트로 시작했다가 정식 승인 없이 프로덕션 의존성이 된 사례이다. 노스틱이 발견한 1,800개 이상의 무인증 서버는 공개 인터넷에 노출된 것만 집계한 숫자이며, 사내망에 머무는 무관리 서버는 여기에 포함되지 않는다. mcp-remote와 포스트마크의 사례는 패키지 공급망 자체가 위협 요소가 됐음을 보여 주는데, 인기 패키지의 업데이트 한 번이 수십만 개의 설치 환경을 동시에 위험에 빠뜨린다.

거버넌스는 본질적으로 프로토콜 계층이 아니라 조직 계층의 문제다. DX 히어로즈의 2026년 초 분석은 "프로토콜 개선이 조직의 정책을 대신할 수는 없다"고 정리하면서, 도구 단위로 통제하는 방법은 표준화되지 않은 채 프로토콜이 기본으로 제공하는 것은 서버 단위의 차단과 허용에 그친다는 점, 로컬 MCP 설정이 중앙 레지스트리의 통제를 우회한다는 점을 짚는다.[16] 정책으로서의 거버넌스가 실제로 작동하려면, 그것을 강제하는 운영 차원의 통제점이 있어야 한다. 그 통제점은 프로토콜이 아니라 운영의 단일 출입구에서 만들어진다.

통제점의 귀환 — API 게이트웨이가 AI 통제점이 되다

API 게이트웨이는 새로운 발명이 아니다. 인증, 속도 제한(Rate Limiting), 로깅, 트래픽 라우팅, 정책 적용은 지난 20년간 게이트웨이의 표준 기능이었다. 2026년 들어 달라진 것은 이 기능들이 새로운 대상에 적용된다는 데 있다. 인증의 대상은 사람과 앱에서 AI 에이전트로 옮겨 가고, 속도 제한의 단위는 초당 요청에서 분당 토큰으로 확장되며, 감사의 단위는 호출당 페이로드에서 에이전트의 의사결정 추적으로 넓어진다.

여기서 한 가지 분명히 해 두고 싶은 것은, 게이트웨이가 보안 도구를 대체하지 않는다는 점이다. 위협을 탐지하고 권한을 판별하는 일은 보안 계층의 몫으로, 3월호에서 다룬 CNAPP(Cloud-Native

16) DX Heroes. "MCP Governance in the Enterprise: What the Landscape Looks Like in Early 2026." 2026. https://dxheroes.io/insights/mcp-governance-landscape-early-2026

Application Protection Platform, 클라우드 네이티브 애플리케이션 보호 플랫폼), 제로 트러스트, 비인간 ID의 영역이다. 게이트웨이의 고유한 역할은 모든 AI 트래픽이 반드시 한 점을 통과하도록 강제하는 인라인 통제(Inline Enforcement)다. 보안 도구가 무엇이 위험한지를 판정한다면, 게이트웨이는 그 판정이 적용될 단일 경로를 만든다. 그래서 앞 챕터의 무인증이나 미승인 부류는 게이트웨이로 가시화되고 차단되지만, 포스트마크 BCC처럼 승인된 서버 내부에서 벌어지는 행위는 게이트웨이가 아니라 보안 계층이 잡아야 한다. 이후 연재 글에서 다룰 ID와 접근 관리가 '누가'를 판정한다면, 게이트웨이는 '어느 경로로'를 강제한다.

이 역할을 두고 6개 벤더가 저마다 뛰어들고 있는데, 다음 챕터의 벤더 분석에서 평가축별로 비교한다. 다만 모든 조직이 지금 이 통제점을 새로 만들어서 관리해야 하는 것은 아니다. MCP가 이미 운영에 들어와 있거나 에이전트가 사내 데이터에 닿고 있다면 그 시점이 앞당겨지고, 그렇지 않다면 인벤토리부터 시작해도 늦지 않다(도입 시점을 가르는 기준은 맺으며에서 구체화한다). 기존 API 게이트웨이가 있다면 이를 교체할 일은 아니지만, 그 게이트웨이가 MCP와 A2A(Agent-to-Agent, 에이전트 간 통신) 트래픽을 같은 평면에서 다룰 수 있는지, LLM 토큰 비용을 정책으로 제한할 수 있는지, 에이전트별 권한 분리를 지원하는지를 로드맵에서 확인해야 한다. 5월호에서 쿠버네티스가 AI 워크로드의 운영체제 자리에 올랐다고 짚었는데, 그 위에는 통제점이 따로 들어서야 한다.

"MCP-ify"의 올바른 방법

"MCP-ify"는 기존에 쓰던 API를 AI 에이전트가 호출할 수 있는 MCP 도구 형태로 바꿔 내놓는 과정이다. 단순한 변환으로 보일 수 있으나, 게이트웨이를 거치지 않은 MCP-ify는 회사 데이터로 직접 이어지는 새 출입구를 내는 일이며, 들어가며에서 본 .exe를 직접 받아 설치하던 시절과 다를 바 없다.

다음의 세 단계를 따르는 게 올바른 순서이다.

1. 발견(Discovery) 단계: 사내 API 인벤토리를 정비한다.
2. 통제 노출(Controlled Exposure) 단계: 게이트웨이를 단일 경로로 강제하되, 노출하려는 MCP 서버의 출처와 무결성을 함께 검증한다.
3. 감사 가능 운영(Auditable Operation) 단계: 에이전트 호출 추적과 토큰 비용 추적을 함께 운영한다.

이 순서를 건너뛰어 발견 단계 없이 노출부터 시작하면 섀도우 MCP가 오히려 자라난다. Kong 컨텍스트 메시, 에이전트코어 게이트웨이, 아피지 API 허브의 스펙 부스트(Spec Boost) 같은 벤더

자동화가 통제 노출 단계의 마찰을 줄여 주지만, 발견과 감사 가능 운영의 책임은 여전히 도입 조직의 몫이다. 4월호에서 "AI가 인프라 코드를 쓸 때 정책형 코드(Policy as Code)로 가드레일을 설정하라"고 제언한 것과 같은 원칙이 여기에도 그대로 적용된다. 에이전트가 API를 호출할 때에도 게이트웨이 정책이 같은 가드레일 역할을 한다.

벤더 심층 분석: 통제점 역량의 비교

벤더들은 아래 표 1과 같이 다섯 가지 축으로 비교한다. 1) 발견 자동화(섀도우 MCP를 가시화하는가), 2) 인증과 권한 분리(에이전트별 권한을 분리하는가), 3) 비용 통제(LLM 토큰 비용을 사전 정책으로 제한하는가), 4) 감사 추적(에이전트 호출을 추적 가능한가), 5) 다중 환경 적용성(특정 클라우드 서비스 제공사나 생태계에 종속되지 않는가)이다.

다만 여섯 벤더가 모두 같은 비교 선상에 서는 것은 아닌데, 포스트맨은 런타임 통제점이라기보다 발견과 카탈로그 단계를 보강하는 보완재에 가까워, 권한 분리와 비용 통제 비교 항목에서는 빠지고, 나머지 다섯이 통제점 자체를 두고 경쟁하는 구도이다.

구분	Kong	아피지	AWS	퓰소프트	타이크	포스트맨
핵심 포지셔닝	AI 연결성 플랫폼	가장 포괄적 API 관리	API/Lambda의 즉시 MCP 변환	통합 + 에이전트 오케스트레이션	오픈소스 통제점	API 카탈로그 + 에이전트 모드
발견 자동화	컨텍스트 메시	API 허브 스택 부스트	에이전트코어 게이트웨이	Agent Registry	AI 스튜디오	API 카탈로그
권한 분리	MCP Registry 정책	Apigee 정책 + IAM	에이전트코어 Policy	Flex Gateway MCP·A2A	오픈코어 정책	MCP 인증 위임
비용 통제	AI Gateway 토큰 한도	LLM 워크로드 정책	베드락 비용 정책	Anypoint 모니터링	커뮤니티 한도	미지원
감사 추적	Konnect 호출 로그	Apigee 분석	CloudWatch / 평가	Anypoint 분석	오픈소스 로그 + 상용 분석	API 카탈로그 활동
약점	신제품 다수 프리뷰	가격·GCP 종속	베드락 생태계 종속	세일즈포스 종속·가격	인지도·인프라 한정	런타임 게이트웨이 부재

그림 5 2026년 API 게이트웨이 6개 벤더의 통제점 역량 비교

Kong — API 게이트웨이에서 AI 연결성 플랫폼으로

Kong은 2026년 2월 "AI 연결성"이라는 새 카테고리를 선언했다. 그 한가운데에 컨텍스트 메시(Context Mesh)가 있다. 사내 API를 자동 발견해 MCP 도구로 변환하고 런타임 거버넌스와 함께 배포하는, Kong이 업계 최초로 내세우는 시도다[17]. MCP 레지스트리는 Kong의 통합 API 관리

17) Kong. "Kong Launches Context Mesh to Connect Enterprise Data to AI Agents." PRNewswire, 2026.02.10. https://www.prnewswire.com/news-releases/kong-launches-context-mesh-to-connect-enterprise-data-to-ai-agents-302683492.html

플랫폼인 Kong Konnect의 API 서비스 카탈로그에 통합돼 MCP 서버·도구의 등록, 발견, 거버넌스를 한 곳에 모은다. 사내에 흩어진 기존 API를 스캔하고, 인증을 포함한 MCP 도구 정의를 자동 생성한 뒤, Kong AI 게이트웨이에 배포하며 기존 접근 제어와 정책을 상속시키는 흐름은 발견 → 통제 노출 → 감사 가능 운영의 세 단계를 한 제품 안에서 처리하려는 접근이다. 솔레이스(Solace, 실시간 이벤트 스트리밍 업체)와의 파트너십으로 API·이벤트 스트리밍·AI 서비스를 단일 통제점으로 묶기로 했다.[18]

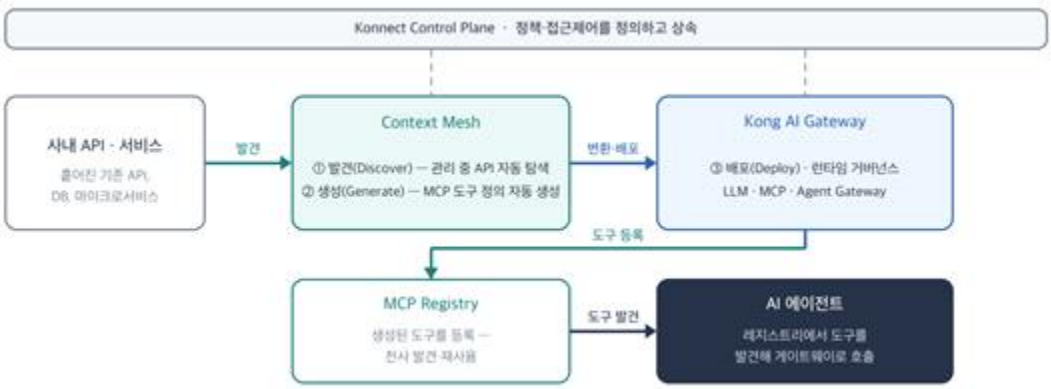


그림 6 Kong 컨텍스트 메시 — 사내 API를 자동 발견해 MCP 도구로 변환·배포하는 흐름

다만 도입 전에 살펴야 할 점들이 있는데, 먼저 컨텍스트 메시와 MCP 레지스트리가 모두 테크 프리뷰 단계여서 프로덕션 검증이 아직 부족하다. 다음으로 "AI 연결성"이라는 새 카테고리의 시장은 아직 형성 중이어서 도입하려 해도 예산 확보의 근거가 약해진다. 마지막으로 API 트래픽 기반 과금에 AI 에이전트 트래픽이 더해지면 비용 예측이 어려워지는 점이 있다.

아피지(Apigee, 구글 클라우드) — 검증된 API 관리에 LLM 비용 통제를 더하다

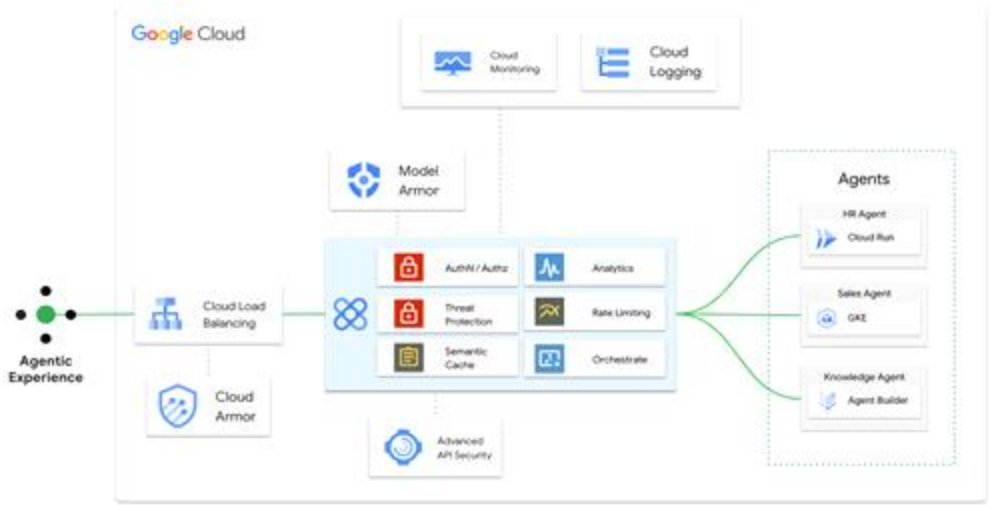
아피지는 2025년 가트너 API 관리 매직 쿼드런트 리더에 10년 연속으로 자리 잡았고 평가 부문 중 '실행 능력'에서 가장 높은 자리에 올랐다.[19] 빅쿼리, 컴퓨트 엔진을 포함한 구글 클라우드 전체

18) Kong. "Kong and Solace Announce Partnership to Unify API and Real-Time Data and Event Streaming." PRNewswire, 2026.02.18. https://www.prnewswire.com/news-releases/kong-and-solace-announce-partnership-to-unify-api-and-real-time-data-and-event-streaming-302691193.html

19) Google Cloud. "Apigee, a Leader in 2025 Gartner Magic Quadrant for API Management." Google Cloud Blog, 2025. https://cloud.google.com/blog/products/ai-machine-learning/apigee-a-leader-in-2025-gartner-api-management-magic-quadrant

서비스에 MCP 지원이 통합된 생태계의 일부로 운영되며, 관리형 API를 커스텀 MCP 서버로 변환하는 기능이 있어 기존 아피지 고객은 자사 API를 즉시 AI 에이전트에게 노출할 수 있다. 그리고 GKE용 APIM 오퍼레이터는 정식 출시되어 쿠버네티스 환경에 경량 API 관리를 제공한다.[20)]

다른 솔루션들과의 차별점은 LLM 워크로드 정책 2종이 정식 출시되어 토큰 사용량 모니터링과 속도 제한으로 AI 호출의 비용과 리소스를 통제할 수 있게 되었는데, 이는 "AI 시대의 API 핀옵스(FinOps)"에 해당하는 기능이다. API 허브의 스펙 부스트는 AI가 기존 API 스펙을 분석해 예시·설명·에러 문서를 자동 보강하며, 에이전트가 API를 더 잘 이해하도록 돕는 발견 단계의 자동화를 이루어 낸다.



21)

그림 7 아피지가 AI 에이전트 트래픽을 중개하며 인증·속도 제한·분석을 적용하는 구조

약점도 분명하다. 가격이 경쟁 대비 높아 소규모 트래픽에서 진입 비용이 부담으로 작용한다. 구글 클라우드에 종속돼 멀티 클라우드 환경에서 단독으로 운영하기 어렵다. 아피지 X(차세대)와 아피지 엣지(레거시) 사이의 마이그레이션이 복잡하다는 점이 고객 불만으로 남아 있다.

---

20) Google Developers. "Google Cloud Announces APIM Operator for Apigee General Availability." Google Developers Blog, 2025.05.12. https://developers.googleblog.com/google-cloud-announces-apim-operator-for-apigee-general-availability/

21) https://cloud.google.com/blog/products/api-management/using-apigee-api-management-for-ai

아마존 베드락 에이전트코어 — API에서 에이전트 게이트웨이로

아마존 베드락 에이전트코어 게이트웨이는 API 게이트웨이 엔드포인트, 람다 함수, 기존 서비스를 몇 줄의 코드로 MCP 호환 도구로 변환해 에이전트에게 제공한다. AWS 워크로드를 가진 조직에는 마찰이 가장 적은 통제 노출 경로인 셈이다.

통제 역량의 핵심은 2026년 3월 3일 정식 출시된 정책이다. 에이전트의 게이트웨이 도구 호출을 사전에 가로채 세밀한 권한을 적용하는 기능으로, 에이전트 거버넌스를 처음으로 프로덕션 수준에서 구현한 것으로 꼽힌다.[22] 평가는 에이전트의 실제 행동에 기반한 품질 평가를 제공하며, 에피소딕 메모리는 에이전트가 경험에서 학습해 유사 상황에 적응하는 장기 메모리이고, 양방향 스트리밍은 음성 에이전트의 자연스러운 대화를 지원한다.

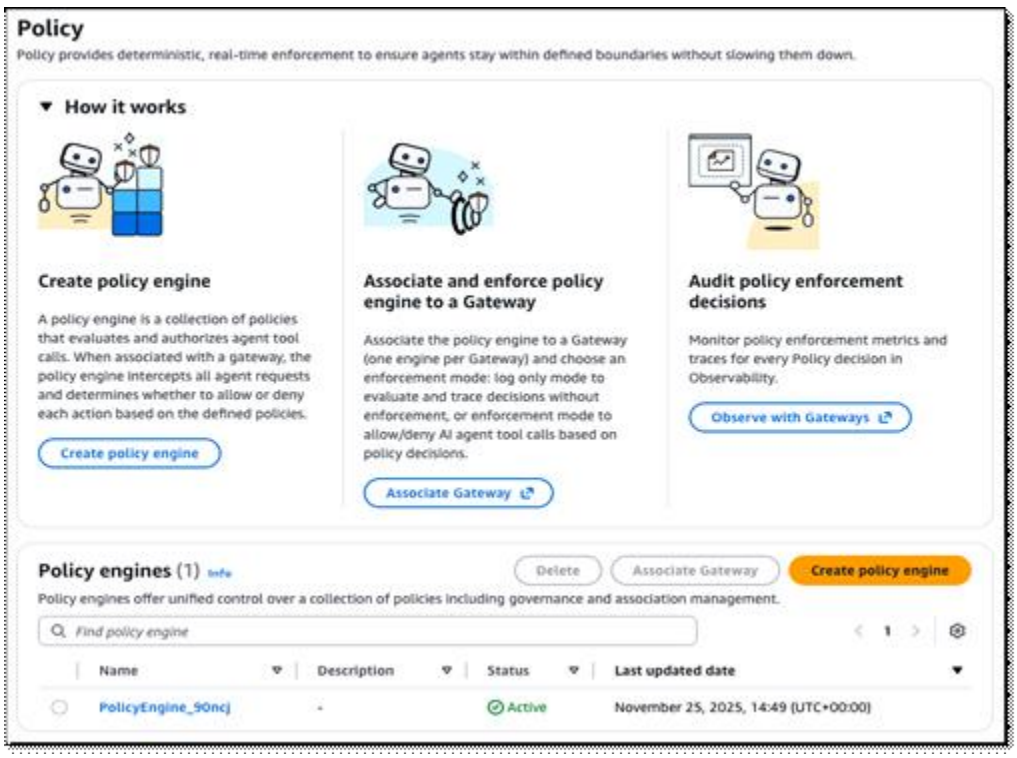


그림 8 에이전트코어 게이트웨이의 정책(Policy) 설정 — 도구 호출 전 권한 적용

22) AWS. "Amazon Bedrock AgentCore adds Quality Evaluations and Policy Controls for Deploying Trusted AI Agents." AWS Blog, 2026.03.03. https://aws.amazon.com/blogs/aws/amazon-bedrock-agentcore-adds-quality-evaluations-and-policy-controls-for-deploying-trusted-ai-agents/

다만 살펴야 할 점이 두 가지 있다. 먼저 에이전트코어 게이트웨이가 베드락 생태계 안에서만 완전한 기능을 발휘해, 앤트로픽 API 직접 호출이나 애저 오픈AI같은 다른 LLM 제공사와의 통합이 제한적이다. 다음으로 기존 API 게이트웨이와 에이전트코어 게이트웨이의 관계가 불명확해 둘을 동시에 운영해야 하는 상황이 발생할 수 있다.

뮬소프트(MuleSoft, 세일즈포스) — 통합 플랫폼에서 에이전트 오케스트레이터로

뮬소프트 에이전트 패브릭(Agent Fabric)은 에이전트 시대의 통합 허브를 지향한다. 에이전트 레지스트리(에이전트·AI 자산 카탈로그), 에이전트 브로커(에이전트 간 지능형 라우팅), 플렉스 게이트웨이(MCP·A2A 보안·관리), 그리고 에이전트 비주얼라이저(에이전트 상호작용 가시화)의 네 기둥으로 구성된다.[23] MCP와 A2A를 함께 다루는 드문 구성이며, 세일즈포스 에이전트포스 3의 네이티브 MCP 지원과 결합돼 CRM 데이터에서 AI 에이전트 동작까지 끊김 없는 흐름을 제공한다. 애니포인트 코드 빌더(Anypoint Code Builder)는 커서와 윈드서프 같은 MCP 지원 IDE에서 뮬소프트 개발을 가능하게 하며, 애니포인트용 아인슈타인이 API 스펙과 데이터위브(Dataweave) 변환을 자동 생성한다. 뮬소프트는 2026년 가트너 iPaaS(통합 플랫폼 서비스) 매직 쿼드런트에서도 줄곧 리더로 평가받아 왔다.

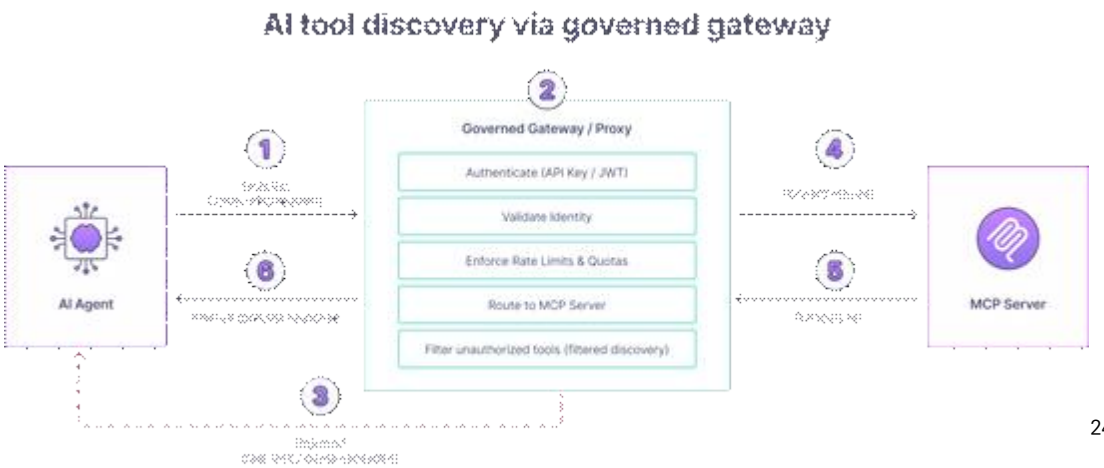
그림 9 뮬소프트 에이전트 패브릭의 네 기둥 — Discover·Orchestrate·Govern·Observ

23) Salesforce. "Salesforce Launches MuleSoft Agent Fabric to Orchestrate and Govern Any AI Agent Across the Agentic Enterprise." Salesforce News, 2026. https://www.salesforce.com/news/stories/mulesoft-agent-fabric-announcement/

약점도 짚어 보면, 세일즈포스 생태계에 종속돼 세일즈포스를 쓰지 않는 조직에는 과잉 투자로 작용한다. 이 도구들을 아우르는 핵심 제품인 애니포인트 플랫폼의 가격이 엔터프라이즈 전용 수준이어서 중소 규모 조직에 진입 장벽이 된다. 통합 플랫폼의 복잡성 때문에 API 관리만 필요한 조직에는 전체 스택이 과하다는 점을 이야기한다.

타이크(Tyk) — 오픈소스 AI 게이트웨이의 대안

타이크는 REST, 그래프QL(GraphQL), TCP, gRPC를 지원하는 오픈소스 API 게이트웨이로, 기능 잠금 없는 '배터리 포함(Batteries-included)' 철학이 차별점이다. CNCF(클라우드 네이티브 컴퓨팅 재단), 리눅스 재단, OpenAPI 스펙 멤버로 오픈소스에 진정성 있게 참여한다. 타이크 AI 스튜디오는 오픈코어 AI 게이트웨이로, LLM·에이전트·MCP 툴체인·RAG 워크로드의 라우팅, 거버넌스, 보안을 단일 통제점에서 처리한다. 4월호에서 다룬 오픈토푸(OpenTofu)가 테라폼의 오픈소스 대안이었던 것과 같은 맥락에서, 타이크는 Kong의 상용 전략에 대한 오픈소스 대안이고, 속도 제한 같은 핵심 통제는 커뮤니티 에디션에서도 문제없이 제공된다. 따라서 데이터 주권이 중요한 규제 산업 등 자체 호스팅을 선호하는 조직에게 가장 자연스러운 옵션으로 고려된다.



24)

그림 10 타이크의 거버넌스 게이트웨이를 통한 MCP 도구 발견·통제 흐름

반대로 살펴야 할 점도 몇 가지 있는데, 먼저 Kong, 아피지, AWS 대비 시장 인지도가 낮아 엔터프라이즈 도입을 설득하기 어렵다. 다음으로는 AI 스튜디오가 오픈코어(핵심 오픈소스에 상용 확장이

24) https://tyk.io/learning-center/mcp-gateway-architecture-technical-guide/

더해진 형태)여서 오픈소스 부분에서 AI 기능의 경계가 어디까지인지 모호하다. 그리고 글로벌 PoP(Point of Presence, 접속 거점) 수와 인프라가 클라우드 서비스 제공사(CSP) 게이트웨이 대비 한정적이다.

포스트맨(Postman) — API 테스팅 도구에서 AI 네이티브 API 플랫폼으로

포스트맨은 2026년 3월 "AI 네이티브 API 개발의 새 시대"를 선언하며 플랫폼을 전면 개편했다.[25] 에이전트 모드는 AI가 포스트맨과 연결된 리포지터리 전반에서 컬렉션·테스트·목(Mock)을 편집하고 생성하는 에이전틱 기능이며, MCP 프로토콜의 네이티브 지원으로 아틀라시안, 클라우드워치, 깃허브 같은 MCP 서버와 통합된다. API 카탈로그는 스펙·컬렉션·테스트 실행·CI/CD 활동·프로덕션 관측성(Observability)을 단일 뷰로 제공하는 'API의 단일 출처'를 지향한다. AI 테스트 생성 기능은 API에 대한 계약·부하·단위·통합·E2E(종단 간) 테스트를 자동으로 만들어 주며, 에이전트 모드는 MCP 서버에서 컨텍스트를 받아 다단계 변경을 자동화한다.

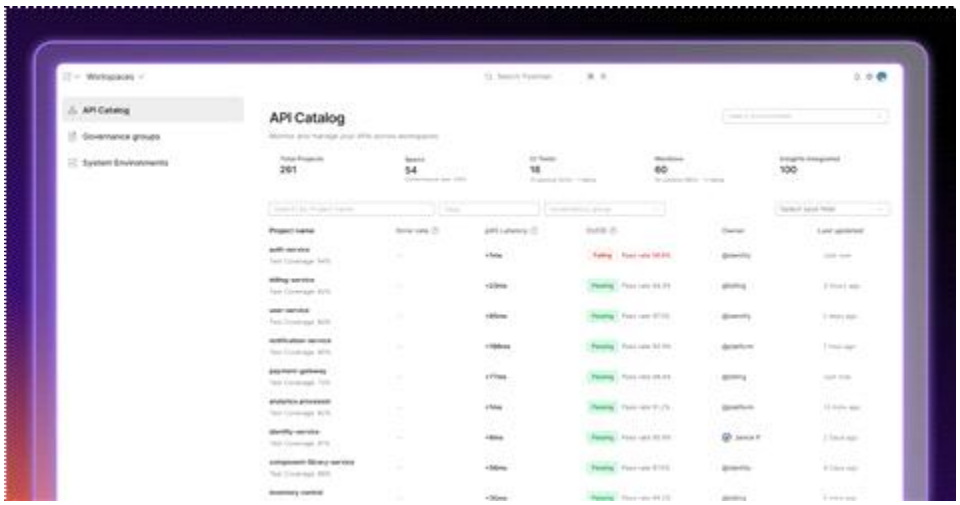


그림 11 포스트맨 API 카탈로그 — 스펙·테스트·관측성을 단일 뷰로 모은 'API 단일 출처'

반면에 약점으로는 포스트맨의 전통적 강점이 API 테스팅에 있고 API 관리에는 있지 않아 Kong, 아피지, AWS 같은 런타임 게이트웨이 기능을 제공하지 않는다는 점이다. 그리고 AI 네이티브 기능은 2026년 3월에 출시돼 대규모 엔터프라이즈 레퍼런스가 아직 쌓이지 않았다는 평이고, 유료 플랜의 가격이 팀 규모에 따라 빠르게 오르는 점도 자주 지적된다.

25) Postman. "New Postman is Here." Postman Blog, 2026.03.02. https://blog.postman.com/new-postman-is-here/

맺으며: CIO·CTO를 위한 API 전략 제언

2026년의 API 전략은 새 표준의 도입이 아니라 통제점의 복원에서 시작한다. 다만 모든 조직이 지금 당장 게이트웨이를 새로 들여야 하는 것은 아니다. 사내에서 MCP 서버가 이미 실제 운영에 깊이 자리 잡았거나, AI 에이전트가 사내 데이터에 접근하고 있거나, LLM 토큰 비용이 무시할 수 없는 수준에 이르렀다면 통제점의 복원은 미룰 수 없는 과제다. 셋 중 어디에도 해당하지 않는다면 지금은 도입을 서두르기보다 인벤토리와 모니터링으로 도입 시점을 가늠하기를 추천한다. 통제점의 복원이 필요한 조직에게 할 질문은 "이미 사내에 퍼진 섀도우 MCP를 어떻게 게이트웨이 뒤로 끌어들이고, 감사 가능한 운영으로 정착시킬 것인가"가 될 것인데, 아래의 네 가지를 권한다.

1. 섀도우 MCP의 인벤토리를 먼저 수행한다: 통제점을 세우기 전에 회사 안에 어떤 MCP 서버가 어디에 연결돼 있는지를 파악하는 것이 우선이다. 깃허브와 npm 사용 분석, 사내 네트워크 트래픽 분석, 개발자 인터뷰의 세 경로로 인벤토리를 작성할 것을 권한다. 이 작업 없이 게이트웨이를 도입하면 통제 대상이 무엇인지 모르는 상태로 시작한다.
2. 게이트웨이를 단일 출입구로 강제한다: 인벤토리가 확보되면 모든 MCP 트래픽이 게이트웨이를 거치도록 네트워크 정책을 설정한다. 이는 .exe 시대에 그룹 정책으로 미승인 실행 파일의 설치를 차단했던 것과 같은 원리다. 기존 API 게이트웨이가 MCP와 A2A를 지원하지 않으면 단기적으로는 어댑터를 두고, 중기적으로는 게이트웨이 교체 또는 추가 도입을 검토할 것을 권한다.
3. LLM 토큰 비용을 API 핀옵스의 일부로 설계한다: 에이전트가 API를 호출할 때 발생하는 LLM 토큰 비용은 기존 API 비용 구조와 다른 변동성을 갖는다. 아피지의 LLM 워크로드 정책이나 Kong AI 게이트웨이의 토큰 한도 같은 게이트웨이 레벨 통제를 핀옵스의 일부로 설계할 것을 권한다. 사후 청구서 분석이 아니라 사전 정책 통제가 필요한 영역이다.
4. API 팀과 AI 팀의 거버넌스를 통합한다: API 팀과 AI 팀이 분리된 거버넌스로 움직이면, API는 에이전트가 안전하게 쓰기 어려운 상태로 남고, 에이전트는 통제 없이 API를 호출한다. 두 전략을 하나의 거버넌스 프레임워크 안에서 운영할 것을 권한다. 앞에서 인용한 DX 히어로즈의 진단처럼 거버넌스의 실패는 프로토콜 계층이 아니라 조직 계층에서 일어나며, 통제점의 복원은 결국 조직의 일이다.

반대로 피해야 할 것도 분명하다. 인벤토리 없이 게이트웨이부터 들이는 일, 발견을 건너뛰고 노출부터 여는 일, "MCP 지원" 발표만 보고 게이트웨이가 사고까지 막아 준다고 기대하는 일이다. 셋 다 통제점을 세우는 대신 새 출입구를 늘린다.

20여 년 전 .exe 시대의 끝을 만든 것은 더 좋은 백신이 아니라 통제점의 등장이었다. 2026년 AI 인프라에서 같은 일이 다시 일어나고 있으며, API 게이트웨이가 그 통제점 자리에 들어서고 있다.

Appendix

구분	Kong	아피지	AWS	뮬소프트	타이크	포스트맨
핵심 포지셔닝	AI 연결성 플랫폼	가장 포괄적 API 관리	API/Lambda 의 즉시 MCP 변환	통합 + 에이전트 오케스트레이션	오픈소스 통제점	API 카탈로그 + 에이전트 모드
발견 자동화	컨텍스트 메시	API 허브 스펙 부스트	에이전트코어 게이트웨이	Agent Registry	AI 스튜디오	API 카탈로그
권한 분리	MCP Registry 정책	Apigee 정책 + IAM	에이전트코어 Policy (GA)	Flex Gateway MCP·A2A	오픈코어 정책	MCP 인증 위임
비용 통제	AI Gateway 토큰 한도	LLM 워크로드 정책 (GA)	베드락 비용 정책	Anypoint 모니터링	커뮤니티 한도	미지원
감사 추적	Konnect 호출 로그	Apigee 분석	CloudWatch / 평가	Anypoint 분석	오픈소스 로그 + 상용 분석	API 카탈로그 활동
약점	신제품 다수 프리뷰	가격·GCP 종속	베드락 생태계 종속	세일즈포스 종속·가격	인지도·인프라 한정	런타임 게이트웨이 부재

표 2