



CONTENTS

Digitalservice Issue Report

디지털서비스 이슈리포트

- | | | |
|----|--|----|
| 01 | AI 에이전트 시대, 서비스의 접점이 바뀐다
: SaaS 역할과 디지털서비스의 구조 전환 | 3 |
| | 이재연 선임연구원·김현우 책임연구원 한국지능정보사회진흥원 | |
| 02 | EU의 AI 분야 유럽공동이익중요프로젝트(IPCEI AI) 이니셔티브 발표 | 11 |
| | 한상기 테크프론티어 | |
| 03 | AI 위협이 바꾸는 보안 시장의 구조 | 17 |
| | 김영욱 SAP Product Engineering Product Expert | |
| 04 | AI 시대의 소프트웨어 엔지니어: 검증과 현장 중심의 역할 재편 | 26 |
| | 윤대균 아주대학교 소프트웨어학과 | |
| 05 | 2026년 클라우드 솔루션 리포트: 추론 인프라(1) 토큰 경제 | 35 |
| | 정채상 이음테크 | |

본 저작물은 한국지능정보사회진흥원이 저작권을 보유하고 있습니다.

한국지능정보사회진흥원의 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

01 AI 에이전트 시대, 서비스의 접점이 바뀐다 : SaaS 역할과 디지털서비스의 구조 전환

| 이재연 선임연구원·김현우 책임연구원 한국지능정보사회진흥원

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

1. AI 에이전트 확산과 디지털서비스 환경 변화

1-1. 생성형 AI 발전과 AI 에이전트 등장

최근 생성형 AI의 급속한 발전은 단순히 ‘정보생성’ 기능을 넘어 사용자의 업무를 대신 수행하는 AI 에이전트 형태로 빠르게 진화하고 있다. 초기 생성형 AI 서비스는 질의응답, 문서 작성, 번역, 코드 작성 등 사용자의 요청에 따라 정보를 만들어내고 제공하는 기능을 중심으로 발전해 왔으나, 최근에는 사용자의 목표를 이해하고 필요한 서비스와 데이터를 스스로 탐색·연계하는 작업을 수행하는 AI 에이전트가 새로운 서비스의 형태로 등장하고 있다.

생성형 AI가 사용자의 질의에 적절한 응답(Response)을 생성하는데 초점을 두었다면, AI 에이전트는 사용자의 목표를 이해하고 이를 달성하기 위한 행동(Action)을 수행하는 구조를 가진다. AI 에이전트는 필요한 과업을 구분하고 실행 순서와 방법을 판단한 후, 외부 서비스와 도구를 활용해 실제 업무를 처리한다. 즉, 단순히 답변을 제공하는 것을 넘어 필요한 작업을 스스로 계획하고 실행하여 결과를 만들어낸다는 점에서 기존 생성형 AI와 차이가 있다.

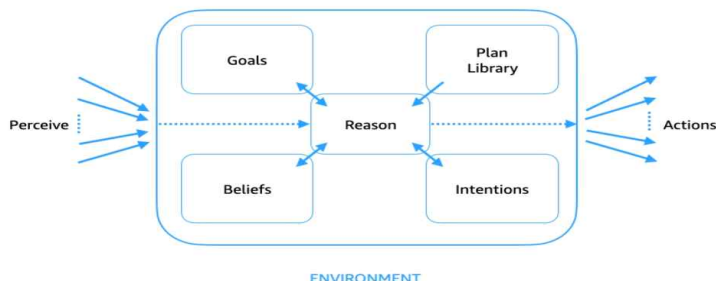


그림 1 AI 에이전트의 인지·추론·실행 구조 (AWS, Foundations of Agentic AI on AWS)

디지털서비스 이슈리포트

업무 수행을 위해 AI Agent는 사용자의 목표를 여러 세부 과업으로 나누고 실행 순서와 방법을 정하는 추론·계획(Reasoning & Planning), 외부 API와 서비스의 기능을 호출하는 도구 활용(Tool Use / Function Calling), 수행 결과를 점검하고 필요한 부분을 다시 실행하거나 보완하는 검토·반복의 과정을 거친다. 이를 통해 정해진 답변을 한 번 제공하는 데 그치지 않고, 목표가 달성될 때까지 작업을 단계적으로 수행하고 있다.

- **추론 및 계획(Reasoning & Planning)** : 사용자가 제시한 목표를 분석하여 필요한 세부 과업으로 분해하고, 이를 수행하기 위한 실행 순서와 방법을 결정한다.¹⁾
- **도구 활용(Tool Use / Function Calling)** : 외부 API를 통해 필요한 도구와 서비스의 기능을 호출하여 정보를 활용하거나 작업을 수행한다.²⁾
- **피드백 및 반복(Reflection & Iterative Refinement)** : 수행 결과를 검토하거나 외부 도구를 통해 결과를 검증하고, 발견된 오류나 개선사항을 반영하여 결과를 반복적으로 개선한다.³⁾

최근에는 일상에서 일정 관리, 문서 작성, 데이터 분석, 예약·구매 등 다양한 활동에 AI 에이전트가 여러 서비스의 기능을 연계·조합하여 업무를 수행하는 사례가 확대되고 있다. 이러한 AI 에이전트의 등장은 단순히 새로운 AI 기능이 추가되는 것을 넘어, 사용자가 개별 디지털서비스를 직접 선택하고 조작해 온 기존의 서비스 이용 방식에도 변화를 가져오고 있다.

1-2. 디지털서비스 이용 방식 변화

이러한 변화에서 주목할 점은 디지털서비스 자체가 즉시 대체되는 것이 아니라 사용자와 디지털서비스 간의 상호작용 방식이 변화한다는 점이다. 기존에는 사람이 개별 서비스의 사용 방법을 이해하고 필요한 기능을 직접 찾아 실행했다면, AI 에이전트 기반 환경에서는 에이전트가 사용자의 의도를 해석하여 적절한 서비스와 기능을 선택·활용하는 역할을 일부 담당하게 된다. 이에 따라 사람의 역할 역시 세부적인 기능 조작에서 목표 설정, 조건 및 권한 부여, 수행결과 확인과 최종 판단 중심으로 변화할 가능성이 있다.

결과적으로 디지털서비스 이용 구조는 기존의 ‘사람 → 개별 서비스’ 중심에서 ‘사람 → AI 에이전트 → 다수의 서비스’가 함께 활용되는 구조로 확대되고 있다. 이는 향후 디지털서비스가 사람뿐만 아니라 AI 에이전트도 필요한 기능과 데이터를 쉽게 탐색하고 활용할 수 있는 형태로 제공될 필요성을 높이고 있으며, 사용자의 직접적인 이용을 중심으로 발전해 온 기존 SaaS(Software as a Service)의 역할과 서비스 제공 방식에도 변화를 요구하고 있다.

1) Andrew Ng, “Agentic Design Patterns Part 4: Planning”, 2024

2) Shishir G. Patil 외, “Gorilla: Large Language Model Connected with Massive APIs”, 2023

3) Andrew Ng, “Agentic Design Patterns Part 2: Reflection”, 2024

디지털서비스 이슈리포트



그림 2 AI 에이전트 등장에 따른 디지털서비스 이용 구조 변화 (출처 : GPT-5.6 활용 제작)

2. SaaS 중심 디지털서비스 구조와 변화

2-1. SaaS 기반 디지털서비스 제공 구조

클라우드 확산과 함께 소프트웨어 이용 방식은 직접 시스템을 구축·설치하는 방식에서 인터넷을 통해 필요한 기능을 서비스 형태로 이용하는 SaaS(Software as a Service) 중심으로 변화해 왔다. SaaS는 이용자가 별도의 시스템을 직접 구축하거나 소프트웨어를 설치하지 않고도 필요한 기능을 이용할 수 있어, 시스템 구축·운영에 따른 부담을 줄이고 필요한 서비스를 신속하게 활용할 수 있다는 장점을 가진다.

현재 이메일, 문서 작성, 화상회의, 회계·인사 등 다양한 분야에서 SaaS가 활용되고 있다. 기존 SaaS는 서비스 제공자가 기능과 데이터를 하나의 애플리케이션을 통해 제공하고, 사용자가 해당 서비스에 직접 접속하여 UI와 업무 흐름(Workflow)에 따라 필요한 기능을 선택·조작하는 '사용자-애플리케이션 중심'의 구조로 발전해 왔다.

이러한 이용 구조는 SaaS의 비즈니스 모델에도 반영되어, 일정 기간 서비스를 이용하는 구독형 모델과 사용자 수를 기준으로 비용을 부과하는 방식이 대표적으로 활용됐다. SaaS는 이를 기반으로 클라우드 기반 디지털서비스의 주요 제공 방식으로 자리 잡았으며, 공공부문에서도 민간 SaaS를 필요한 시점에 선택·이용할 수 있도록 관련 제도와 이용 체계가 마련되어 왔다.

2-2. Agent가 혼드는 기존 SaaS 모델

AI 에이전트가 사용자의 업무 목표를 바탕으로 여러 소프트웨어의 기능을 대신 탐색·활용하게 되면서,

디지털서비스 이슈리포트

사용자가 개별 SaaS에 직접 접속하여 기능을 이용하는 기존 서비스 구조에도 변화가 나타나고 있다. 사용자가 각각의 서비스를 직접 이용하지 않고도 에이전트를 통해 필요한 기능을 활용할 수 있게 되면서 UI·UX 중심의 사용자 접점, 사용자 수 기반 과금 체계, 개별 애플리케이션 중심의 서비스 구조 등 기존 SaaS 모델을 구성해 온 주요 요소가 영향을 받을 가능성이 있다.

Gartner는 이러한 변화를 ‘에이전틱 아비트리리지(Agentic Arbitrage)’, 즉 AI 에이전트가 사용자를 대신해 여러 소프트웨어의 기능을 선택·활용하면서 개별 애플리케이션에 대한 사용자의 직접 이용이 감소하는 현상으로 설명한다. 이에 따라 사용자 증가와 SaaS 사업자의 수익 증가 간 기존의 연결 관계가 약화될 수 있다는 것이다. Gartner는 2030년까지 최대 2,340억 달러의 기업용 애플리케이션 소프트웨어 지출이 이러한 변화의 영향을 받을 수 있으며, 이는 2030년 기업용 애플리케이션 SaaS 지출의 약 20%에 해당할 것으로 전망했다.

특히 에이전트가 한 명의 사용자를 대신해 여러 업무를 수행할 경우 사용자 수만으로 서비스의 실제 이용 규모를 측정하기 어려워질 수 있다. 이에 따라 기존 사용자 수 기반 구독 방식과 함께 API 호출량이나 에이전트의 Action, 업무처리량 등 실제 이용을 반영한 과금방식이 확대될 가능성이 있다. 또한 에이전트가 업무 목적에 따라 여러 서비스의 기능을 선택·조합하면서, SaaS의 경쟁 요소도 기능의 다양성과 UI·UX뿐만 아니라 다른 서비스와의 연결성, 핵심 기능과 데이터의 활용성 등으로 확대될 수 있다.

최근에는 이러한 AI 에이전트 확산에 따른 기존 SaaS 시장의 변화를 ‘사스포칼립스(SaaS Spocalypse)’라고 표현하기도 한다. 그러나 이는 SaaS 자체가 사라진다는 의미라기보다, 사용자의 직접적인 이용을 전제로 형성되어 온 기존 SaaS의 가치 제공 방식과 수익·경쟁구조가 에이전트 환경에 맞게 재편될 필요성이 커지고 있음을 의미한다. 즉, AI 에이전트의 확산은 SaaS의 소멸보다는 SaaS가 이용되고 가치를 제공하는 방식의 변화를 촉진하고 있다.

3. AI 에이전트 확산에 따른 SaaS 역할 변화

3-1. AI 에이전트 기반 서비스 이용 구조

AI 에이전트 기반 환경에서는 하나의 애플리케이션이 업무 전 과정을 제공하기보다, 에이전트가 사용자의 목표에 따라 필요한 서비스와 기능을 선택하고 조합하여 업무를 수행하는 방식이 확대될 수 있다. 이 과정에서 에이전트는 업무를 세부 작업으로 나누고, 각 작업에 필요한 외부 서비스의 기능과 데이터를 호출하여 하나의 업무 흐름으로 연결한다.

디지털서비스 이슈리포트

예를 들어 출장 준비 업무의 경우 일정 확인, 교통편·숙박 검색, 예약, 일정 등록 등에 필요한 기능이 각각 다른 서비스에 존재하더라도 에이전트가 이를 순차적으로 호출·연계하여 업무를 수행할 수 있다. 이때 사용자는 개별 서비스의 기능이나 이용 방법을 일일이 파악하기보다 업무 목표와 조건을 제시하고, 각 SaaS는 에이전트가 업무를 수행하는 과정에서 필요한 기능과 데이터를 제공하는 역할을 하게 된다.

이러한 구조에서는 서비스 간 연결과 기능 호출을 위한 API를 비롯하여 Agent와 외부 서비스 간의 연계를 지원하는 기술의 중요성도 높아지고 있다. 최근에는 MCP(Model Context Protocol) 등 에이전트가 다양한 외부 도구와 데이터에 접근할 수 있도록 지원하는 연계 방식도 등장하면서, AI 에이전트가 서로 다른 서비스의 기능을 활용할 수 있는 기반이 확대되고 있다.

이에 따라 SaaS의 역할도 사용자가 직접 접속하여 이용하는 하나의 완결된 애플리케이션(Application)에 머무르지 않고, AI 에이전트가 업무 수행에 필요한 기능과 데이터를 활용할 수 있도록 제공하는 형태로 확대될 가능성이 있다.

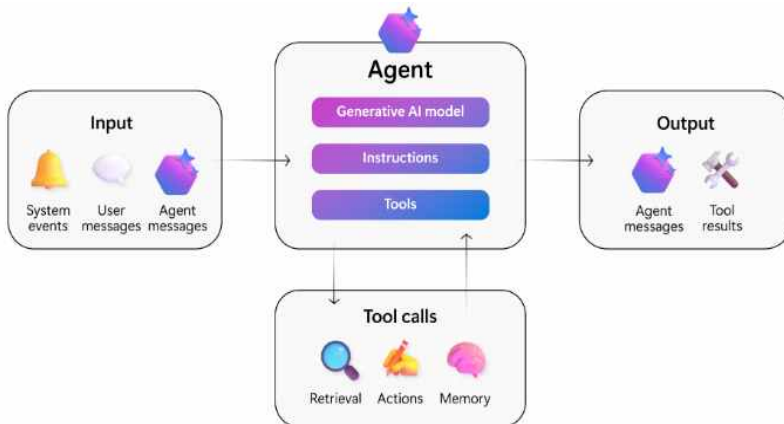


그림 3 AI 에이전트 기반 워크플로우 (출처 : Microsoft Ignite, AI 에이전트)

3-2. SaaS의 기능화·백엔드화 및 서비스 구조 재편

AI 에이전트 기반 환경에서 SaaS는 사용자가 직접 이용하는 완결된 애플리케이션인 동시에, 에이전트가 업무 수행에 필요한 기능과 데이터를 호출·활용하는 서비스로 역할이 확대될 수 있다. 예를 들어 일정 등록, 문서 생성, 고객정보 조회 등 기존에는 사용자가 SaaS 화면에서 직접 실행하던 기능을 에이전트가 외부 연계를 통해 호출하여 업무 과정의 일부로 활용하는 방식이다.

IDC는 이러한 변화를 SaaS가 에이전트 계층 뒤에서 보이지 않는 ‘피쳐웨어(Featureware)’로 변화할

디지털서비스 이슈리포트

가능성으로 설명한다. Agent가 사용자와의 상호작용과 업무 흐름(Workflow)을 담당하고 기존 SaaS에는 API를 통해 접근하여 필요한 기능을 실행하는 구조가 확대될 수 있다는 것이다.⁴⁾

이에 따라 SaaS의 경쟁력 역시 UI와 기능의 다양성뿐만 아니라, 신뢰할 수 있는 데이터와 핵심 업무 기능을 갖추고 있는지, 그리고 에이전트가 이를 API와 표준화된 인터페이스를 통해 안전하게 활용할 수 있도록 제공하는지가 중요해질 수 있다. MCP와 같은 표준 역시 AI 애플리케이션이 외부 데이터와 도구를 활용할 수 있도록 표준화된 연결 방식을 제공하고 있다.⁵⁾

다만 이러한 변화가 기존 SaaS의 UI나 애플리케이션 자체가 사라지는 것을 의미하지는 않는다. 복잡한 정보의 조회·분석이나 사용자의 직접적인 판단과 조작이 필요한 업무에서는 기존 SaaS의 UI와 업무 흐름(Workflow)이 계속 활용될 수 있으며, 반복적이거나 다른 서비스와의 연계가 필요한 업무에서는 에이전트가 SaaS의 기능과 데이터를 직접 호출하는 방식이 함께 활용될 가능성이 있다. 즉, SaaS는 ‘사람이 직접 이용하는 서비스’와 ‘에이전트가 활용하는 기능·데이터 제공자’라는 두 가지 접점을 함께 갖는 방향으로 변화할 수 있다.

이에 따라 SaaS의 경쟁력 역시 개별 애플리케이션의 기능과 사용 편의성뿐만 아니라, 에이전트가 필요한 기능과 데이터를 얼마나 쉽게 탐색하고 안전하게 활용할 수 있는지로 확대될 수 있다. API 등 외부 연계 수단의 제공, 데이터 접근성, 상호운용성, 권한관리 등은 에이전트 기반 환경에서 SaaS의 활용 가능성을 결정하는 주요 요소로 부각될 것으로 보인다.

결국 AI 에이전트의 확산은 SaaS를 대체하기보다 SaaS가 가치를 제공하는 방식과 서비스 간 관계를 재편하는 방향으로 작용할 가능성이 크다. 개별 애플리케이션 중심으로 제공되던 디지털서비스들이 에이전트를 매개로 다양한 기능과 데이터가 연결·조합되는 구조로 확대되면서, SaaS 역시 하나의 완결된 서비스뿐만 아니라 에이전트 기반 업무를 구성하는 기능과 데이터의 제공자로 역할이 확장될 수 있다.

4. AI Agent 시대의 디지털서비스 방향

4-1. Agent 기반 디지털서비스 구조 변화 전망

AI 에이전트의 확산은 개별 애플리케이션을 중심으로 제공되어 온 디지털서비스의 구조에도 변화를 가져올 것으로 예상된다. 기존에는 사용자가 필요한 서비스를 선택하고 각각의 애플리케이션을 직접

4) IDC, “Agent Supplier or Featureware: The Choice Every SaaS Vendor Faces Now”, 2026.6.18.

5) Model Context Protocol, “Specification”, 2024

디지털서비스 이슈리포트

이용하는 방식이 일반적이었다면, 에이전트 기반 환경에서는 사용자의 목표에 따라 여러 SaaS와 데이터, 외부 시스템의 기능이 연결·조합되어 하나의 업무를 수행하는 형태가 확대될 수 있다.

이에 따라 디지털서비스의 제공 단위도 하나의 완결된 애플리케이션에서 다양한 기능과 데이터가 서로 연결되어 활용되는 형태로 확대될 가능성이 있다. 기존 SaaS가 독립적인 서비스로 제공되는 구조와 함께, SaaS의 일부 기능이나 데이터가 API 등을 통해 에이전트에 제공되거나 여러 서비스가 에이전트를 중심으로 하나의 업무 흐름을 구성하는 방식이 공존하는 것이다. 이러한 변화가 확대될 경우 기존 SaaS에 에이전트 기능이 결합된 서비스뿐만 아니라, 에이전트가 여러 SaaS와 데이터·시스템을 연계하여 업무를 수행하는 ‘에이전트 중심의 디지털서비스’도 새로운 서비스 유형으로 등장할 수 있다.

이러한 변화는 디지털서비스의 가치와 이용방식에도 영향을 미칠 수 있다. 개별 서비스가 보유한 기능뿐만 아니라 다른 서비스 및 에이전트와의 연계와 상호운용성, 기능과 데이터에 대한 접근성 등이 중요해지고, 서비스의 이용·과금 방식 역시 사용자 수 중심에서 실제 기능 호출이나 업무 수행량 등을 반영하는 방식으로 다양화될 가능성이 있다.

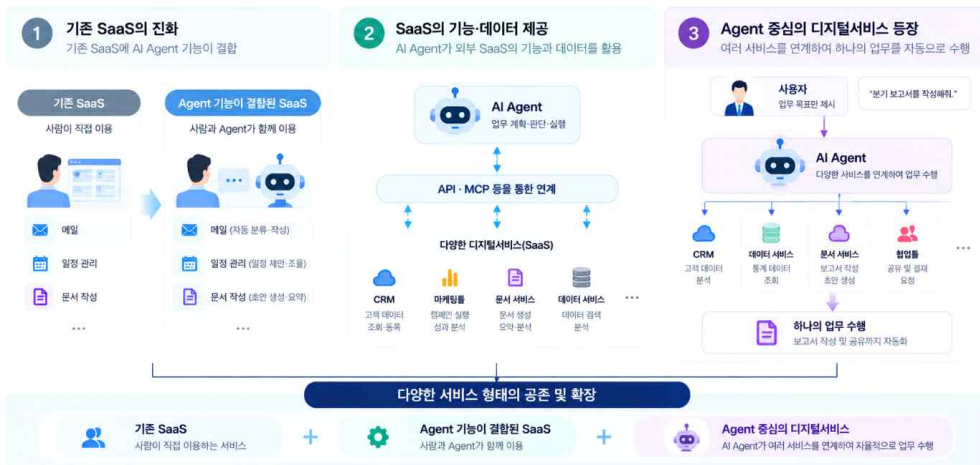


그림 4 AI 에이전트 시대의 디지털서비스 발전 방향 (출처 : GPT-5.6 활용 제작)

4-2. AI 에이전트 시대의 공공 디지털서비스의 대응 방향

향후 기존 SaaS에 AI Agent 기능이 결합되거나 Agent가 여러 SaaS와 데이터·시스템을 연계하여 업무를 수행하는 서비스가 확대될 경우, 공공부문에서도 이러한 서비스를 기존 디지털서비스 이용 체계 안에서 어떻게 발굴하고 도입·이용할 것인지에 대한 검토가 필요하다.

디지털서비스 이슈리포트

다만 AI 에이전트의 등장만을 이유로 새로운 제도나 별도의 서비스 유형을 선제적으로 마련하기보다는, 현행 디지털서비스 제도가 에이전트 기반 서비스를 어느 범위까지 수용할 수 있는지를 우선 살펴볼 필요가 있다. 현재 디지털서비스 제도는 클라우드서비스를 기반으로 하면서 클라우드와 다른 기술·서비스가 융합된 디지털서비스를 포함하고 있는 만큼, Agent의 제공 형태에 따라 현행 제도와의 관계를 구분하여 검토할 필요가 있다.

기존 SaaS 등 클라우드서비스에 에이전트 기능이 결합되어 제공되는 경우에는 기존 클라우드서비스의 기능 확장 또는 융합된 디지털서비스의 형태로 현행 제도와 연계될 가능성이 있다. 반면, 에이전트 자체가 독립적인 서비스로 제공되면서 여러 SaaS와 데이터·시스템을 연계하여 업무를 수행하는 경우에는 해당 서비스가 현행 디지털서비스의 범위에 포함될 수 있는지, 포함된다면 어떠한 기준과 단위로 등록·이용할 것인지 등에 대한 추가적인 검토가 필요하다.

특히 에이전트 자체가 하나의 서비스로 제공되거나 여러 외부 서비스를 연계하는 형태가 확대될 경우, 기존 디지털서비스와는 다른 특성이 나타날 수 있다. 하나의 에이전트가 여러 서비스를 연계하는 경우 어디까지를 하나의 디지털서비스로 볼 것인지, 외부 서비스와의 연계 범위를 어떻게 확인할 것인지, 호출량이나 업무 수행량 등 새로운 과금 방식을 기존 등록·계약 체계에서 어떻게 수용할 것인지 등에 대한 검토가 필요할 수 있다.

또한 에이전트가 실제 업무를 수행하는 과정에서 데이터 접근과 기능 실행이 이루어지는 만큼, 접근권한, 사용자 승인 및 실행기록 등 보안과 신뢰성을 확보하기 위한 요소도 함께 고려할 필요가 있다. 이러한 사항은 디지털서비스 제도뿐만 아니라 AI·보안 및 기관 내부의 업무관리 체계와도 관련되는 만큼, 향후 공공부문의 적용 가능성이 높은 분야를 중심으로 에이전트 기반 디지털서비스의 활용 사례를 발굴하고, 실제 이용하는 과정에서 나타나는 특성을 바탕으로 기존 디지털서비스의 발굴·등록·계약·이용 체계와의 연계 방안을 단계적으로 검토해 나갈 필요가 있다.

02 EU의 AI 분야 유럽공동이익중요프로젝트(IPCEI AI) 이니셔티브 발표

| 한상기 테크프론티어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

IPCEI AI 이니셔티브 발표

IPCEI는 여러 EU 회원국이 국가보조금과 민간투자를 결합해, 유럽의 AI 기술·서비스 생태계를 공동으로 육성하려는 대규모 산업정책 사업이다. IPCEI는 Important Projects of Common European Interest, 즉 ‘유럽공동이익중요프로젝트’를 뜻한다. EU의 전략적 목표에 기여하는 고위험·혁신 사업에 회원국이 공적 지원을 제공하고, 집행위원회가 이를 국가보조금 규정에 따라 심사하는 제도를 말한다.

AI는 이미 2024년 11월 IPCEI 후보로 지지받았다. 집행위원회의 9월 16일 19개 회원국이 AI 분야에서 유럽 공동 이익을 위한 최초의 중요 프로젝트(IPCEI)를 수립하고 국가보조금 규정에 따라 사전 통보하기로 한 이니셔티브를 발표했다.⁶⁾ 프로젝트는 분산형이며 누구나 접근 가능한 다양한 AI 서비스를 개발하는 것을 목표로 한다.

IPCEI AI는 AI 스택 전반을 포괄하며, 관련 분야의 최첨단 기술을 뛰어넘는 혁신적이고 고도로 발전된 AI 및 컴퓨팅 관리 기술과 응용 프로그램, 산업 제품 및 서비스를 개발하는 것을 목표로 한다. 또한 유럽의 디지털 주권에 기여하면서 기술 발전을 주도하고, 이를 통해 유럽을 AI 혁신의 선두 주자로 자리매김하며, 기술 개발부터 도입에 이르는 유럽의 AI 가치 사슬을 강화하고자 한다. 이는 유럽연합 집행위원회의 AI 대륙 행동 계획, AI 전략, 그리고 클라우드 및 AI 개발법 안에서 제시된 바와 같이, EU가 AI 분야의 글로벌 리더가 되는 목표 달성에 기여하고자 하는 것이다.

IPCEI AI는 19개 회원국(오스트리아, 벨기에, 크로아티아, 에스토니아, 핀란드, 프랑스, 독일, 헝가리, 아일랜드, 이탈리아, 라트비아, 룩셈부르크, 네덜란드, 폴란드, 루마니아, 슬로바키아, 슬로베니아,

6) EC, “Commission welcomes the design of first Important Project of Common European Interest in AI,” Sep 16, 2026

디지털서비스 이슈리포트

스페인, 스웨덴)이 공동으로 설계했으며 독일이 조정 역할을 맡고 있다. 유럽연합 집행위원회는 2025년 4월에 설립된 'IPCEI 설계 지원 허브'를 통해 이 사업을 지원해 왔다. 이 허브에서 집행위원회는 회원국들이 프로젝트에 대한 국가보조금 신고를 준비하여 처음부터 EU 규정을 준수할 수 있도록 지원한다. 이를 통해 국가보조금 신고 처리가 더욱 신속하게 진행될 수 있을 것이다.

IPCEI AI의 일환으로, 직접 참여자에게 국가 지원을 제공할 11개 회원국은 2026년 9월부터 해당 프로젝트를 유럽 위원회에 사전 통보할 것이라고 발표했다. 위원회는 이후 해당 프로젝트가 IPCEI 공고의 적격성 및 적합성 기준을 충족하는지 평가할 것이다.

회원국의 공적 지원과 참여기업의 공동투자를 바탕으로 하며, 지원의 정당성은 다음 기준으로 심사한다. 아래는 IPCEI AI에 적용되는 일반 제도 기준이다.

- 지원 필요성: 보조금이 없으면 사업 추진이 어렵거나, 규모·속도·범위가 크게 줄어드는지 확인한다.
- 혁신성: 연구개발과 최초 산업적 적용은 기존 기술 수준을 넘어서는 혁신을 요구한다.
- 유럽 전체의 편익: 성과가 지원받는 기업에 머물지 않고 다른 기업·산업·회원국으로 확산돼야 한다.
- 지원액의 적정성: 사업의 자금 부족분을 분석해 필요한 수준으로 제한한다. 정당화되는 경우 적격 비용 전액까지 가능하지만, 자동으로 전액을 지원한다는 뜻은 아니다.

이 기준들은 회원국 간 보조금 경쟁과 시장 왜곡을 억제하면서 공동투자를 가능하게 하는 장치이다.

기존 클라우드 IPCEI와 IPCEI AI와의 차이점은 다음과 같다.

구분	IPCEI CIS	IPCEI AI
사업명	Next Generation Cloud Infrastructure and Services	Artificial Intelligence
중심 영역	상호운용 가능한 클라우드·엣지 데이터 처리 생태계	AI 기술, 컴퓨팅 관리, AI 서비스와 산업적 활용
확인된 단계	2023년 12월 승인	2026년 9월 공동 설계·사전신고 추진 발표
참여 규모	7개국, 19개 기업	공동 설계 19개국
공개 재원	공적 지원 최대 12억 유로 , 추가 민간투자 예상 14억 유로	9월 발표문에 총액 미제시

그림 5 클라우드 IPCEI와 IPCEI AI의 차이 비교

디지털서비스 이슈리포트

집행위원회의 9월 16일 발표는 사업 범위를 다음과 같이 설명하고 있다. 오른쪽 열은 공식 발표의 범위를 클라우드 관점에서 해석한 것이다. 발표문은 유럽의 디지털 주권과 AI 가치 사슬 강화를 명시하지만, 개별 모델·GPU 규모·세부 기술 과제까지 공개하지는 않았다.

범위	공식 발표의 내용	클라우드와의 연결
AI 기술 전반	AI 기술 체계 전체를 대상으로 혁신 기술 개발	AI를 개발·운영하는 기반 기술과 연결
컴퓨팅 관리	첨단 컴퓨팅 자원 관리 기술 개발	분산된 연산자원을 효율적으로 활용하는 기술과 관련
AI 서비스	분산형이면서 공동으로 접근할 수 있는 AI 서비스 개발	여러 사업자·기관이 활용하는 서비스 기반과 관련
산업적 활용	AI 응용, 산업 제품·서비스 개발	기술 개발에서 기업의 실제 도입까지 연결

그림 6 ICPEI AI의 사업 범위

IPCEI 지원 체계의 발전과 설계 지원 허브

유럽 집행위(EC)는 국가보조금 규정에 따라 역내 시장과 양립 가능한 유럽 공동 이익의 중요 프로젝트(IPCEI)를 지원하기 위한 국가 보조금을 승인할 수 있다. IPCEI는 최소 4개 회원국이 참여하는 통합 대규모 국경 간 프로젝트로서, 높은 수준의 기술적 또는 재정적 위험을 수반하며, 예를 들어 유럽 생태계 조성 등을 통해 EU에 이익을 가져다주는 프로젝트가 대상이다.⁷⁾

IPCEI는 1957년 조약에 뿌리를 두고, 2014년 지원체계가 정비됐으며, 2018년부터 대규모 공동 산업 프로젝트로 본격 활용됐다. 1957년 로마조약의 유럽경제공동체 설립조약 제92조 제3항 (b)에 ‘유럽 공동 이익의 중요 프로젝트’를 위한 국가보조금을 허용할 수 있다는 근거가 들어갔으며, 2014년 집행위원회가 IPCEI 전용 지침을 마련해, 기존 분야별 기준을 통합하고 지원 대상·혁신성·공동 이익·보조금 심사 기준을 체계화했다. 2018년 12월 18일에는 여러 국가·기업의 연구개발을 연결하는 첫 통합 IPCEI로 마이크로일렉트로닉스 사업이 승인됐다. 초기 참여국은 프랑스·독일·이탈리아·영국이다.

2021년 11월 25일에는 IPCEI 실행을 촉진하기 위한 국가 지원의 역내 시장 적합성 분석 기준에 관한 개정된 공고를 발표했다.⁸⁾ 이 공고는 시장 실패를 극복하고 핵심 부문 및 기술, 인프라 투자 분야에서

7) EC, “Important Projects of Common European Interest (IPCEI),” https://competition-policy.ec.europa.eu/state-aid/ipcei_en 참고

디지털서비스 이슈리포트

획기적인 혁신을 가능하게 하며 EU 경제 전반에 긍정적인 파급 효과를 가져오는 국경을 넘는 IPCEI에 대한 회원국의 지원을 평가하는 기준을 제시했다.

개정된 IPCEI 공지에 따라 지원을 받으려면 프로젝트는 다음 조건을 충족해야 한다.

1. EU 목표 달성에 중요한 기여를 한다.
2. 중요한 시장 실패를 확실히 극복한다.
3. 프로젝트의 성격상 예외적으로 더 적은 수의 회원국이 참여해도 되는 경우가 아니라면, 최소 4개 회원국이 참여해야 한다.
4. 모든 회원국이 새롭게 떠오르는 프로젝트에 참여할 수 있는 진정한 기회를 제공하는 투명하고 포용적인 방식으로 설계되어야 한다.
5. 참여 회원국과 기업을 넘어 EU 경제와 사회 전반에 긍정적 파급 효과를 가져오는 구체적인 성과를 창출한다.
6. 이는 국가 지원을 받을 기업들의 중요한 공동 자금 조달을 수반한다.
7. '중대한 피해를 주지 않는다'는 원칙을 준수하지 않아 발생하는 부정적인 환경적 영향은, 그로 인한 긍정적인 효과가 상쇄될 가능성이 낮으므로 피해야 한다.

IPCEI는 유럽연합 산업 및 경제의 경제 성장, 일자리 창출, 청정 디지털 전환, 경쟁력 강화에 상당한 기여를 할 수 있다고 본다. IPCEI를 통해 유럽연합 전역의 지식, 전문성, 재정 자원 및 경제 주체들이 한데 모여 연합 전체에 긍정적인 파급 효과를 창출할 수 있다는 생각으로 시작한 것이다.

IPCEI 설계 지원 허브

IPCEI 설계 지원 허브는 IPCEI 후보국의 설계 단계에 참여하는 회원국에 전문적인 기술 및 전문가 지원을 제공하여 IPCEI 프로세스를 간소화하고 효율성을 높이며 후속 평가 단계를 더욱 잘 준비할 수 있도록 지원하는 것을 목표로 한다.⁹⁾

각 신규 IPCEI 후보는 자체 설계 지원 허브로부터 지원받는데, 이 허브는 경쟁총국(DG COMP)이 주도하는 팀으로 구성되며, 유럽연합 집행위원회의 여러 총국 소속 관련 전문가들과 긴밀히 협력하고, 필요한 경우 유럽공동연구센터(JRC)도 참여한다.

8) EC, "State aid: Commission adopts revised State aid rules on Important Projects of Common European Interest," Nov 25, 2021

9) 디자인 지원 허브의 주요 내용은

https://competition-policy.ec.europa.eu/state-aid/ipcei/design-support-hub_en 에서 찾아 볼 수 있다.

디지털서비스 이슈리포트

IPCEI는 국가 지원 도구이며 회원국이 주도적인 역할을 한다. 설계 지원 허브는 설계 단계에서 회원국을 지원하고 제안을 제공하지만, 국가 차원의 절차나 개별 프로젝트 선정 등 회원국의 주도권을 대체하지는 않는다.

IPCEI 후보자	후보자를 지지한 회원국 수	승인 날짜
첨단 반도체 기술	15	2024년 11월 27일
인공지능	10	2024년 11월 27일
바이오 기반 화학물질	10	2025년 11월 24일
바이오 기반 소재	11	2025년 11월 24일
바이오 기반 식품 및 사료 원료	11	2025년 11월 24일
청정 기술을 위한 순환형 첨단 소재	15	2024년 11월 27일
컴퓨팅 인프라 연속체	10	2024년 11월 27일
핵심 원자재	13	2026년 5월 20일
혁신적인 핵 기술	13	2025년 4월 9일

그림 7 디자인 지원 허브의 IPCEI 지원 후보들

디자인 지원 허브는 IPCEI 커뮤니케이션 및 사례 연구에 따라 IPCEI 기준에 대한 지속적인 피드백과 설명을 제공하며, 여기에는 다음 사항이 포함된다.

- 잠재적인 IPCEI의 범위와 구조를 파악하여 통합성을 입증할 수 있도록 한다.
- 잠재적 IPCEI가 해결해야 할 중요한 시장 실패를 식별하고 입증하는 데 필요한 지침을 제공한다.
- IPCEI를 중요하고 통합적인 프로젝트로 전반적으로 설명하는 '샤포(Chapeau)' 문서의 공동 제작
- 회원국들이 JEF-IPCEI의 국가별 모범 사례 모음집에 포함된 권고사항을 준수하면서 국가 프로세스를 적절하게 설계할 수 있도록 적극적으로 지원한다.
- JEF-IPCEI의 국가별 모범 사례 모음에 따라 시기를 포함한 조율된 방식으로 회원국들이 국가 차원의 관심 표명 요청 절차를 수립할 수 있도록 적극적으로 지원한다.

디지털서비스 이슈리포트

- 매칭 이벤트의 조직, 목적 및 내용과 관련된 전문적인 물류 및 기술 지원
- 국가별 사업 선정 후, 선정된 사업들에 대해 조기 사전 심사를 실시하여 후속 평가 단계에 필요한 충분한 품질의 문서와 증거를 확보하도록 한다.
- 회원국들이 통합을 입증하는 프로젝트의 참여를 어떻게 마련해야 하는지에 대한 분야별 지침을 제공한다.
- 회원국들이 기술, 부문 및 IPCEI별로 나타나는 긍정적 파급 효과를 파악할 수 있도록 지침을 제공한다.

IPCEI의 수명주기는 (i) 식별, (ii) 설계, (iii) 평가 및 (iv) 구현의 네 단계로 구성된다. 식별 단계를 거쳐 충분한 수의 회원국이 IPCEI 후보를 승인하면 설계 단계가 시작되며, 설계 단계에서는 조정 회원국을 중심으로 참여 회원국들이 IPCEI의 형태를 구상하고, 범위와 목적을 결정하며, 국가 차원에서 프로젝트를 선정 및 심사한다. 이 단계는 IPCEI를 유럽 위원회에 사전 통보하는 것으로 마무리되며, 이로써 평가 단계가 시작된다.

정리하며

이번 새로운 프로그램은 제조업 등 실제 산업 응용 기업의 기회에 주목해야 한다. 다만 범용 모델 개발뿐 아니라 기업의 AI 도입까지 연결하는 구조를 어떻게 가속화할 것인가를 살펴봐야 한다. 지원하는 조직은 사업의 혁신성, 보조금 필요성, 경쟁에 미치는 영향을 입증해야 하며, 사전 신고 단계에서 준비가 부족한 과제는 지원에서 제외될 수 있다.

정책적으로 주목해야 하는 부분은 여러 나라의 개별 AI 사업을 공동 기술개발과 성과 확산 의무로 연결한다는 점이다. 향후 실질적인 규모와 경쟁력을 평가하려면 총보조금보다도, 최종 승인되는 기업·과제의 구성, 국가 간 기술적 연계, 외부 기업이 결과물을 활용할 수 있는 조건을 함께 확인해야 한다. 참여국 수보다 기술 로드맵의 조정, 상호보완적인 파트너 구성, 수혜기업 밖으로의 성과 확산이 더 중요하다는 판단이다.

03 AI 위협이 바꾸는 보안 시장의 구조

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

2026년 4월, 앤트로픽이 최신 모델 미토스(Mythos)를 발표하며 이 모델이 정교하게 방어된 기업 네트워크마저 자율적으로 뚫을 수 있다고 경고했을 때, 기업 보안 책임자들의 반응은 예산을 늘리는 데 그치지 않았다. 그들은 돈을 '어디에' 쓸지를 다시 계산하기 시작했다. 이스라엘 신생 보안 기업 자프란(Zafran)은 통상 수개월이 걸리던 대형 계약을, 미토스 발표 이후 5주 만에 대형 은행 세 곳과 연달아 체결했다.¹⁰⁾ 무언가가 근본적으로 바뀌었다.

일반적인 기업의 보안 예산 총량은 완만하게 늘어난다. 가트너는 2026년 전 세계 사이버보안 지출이 2,400억 달러로, 전년의 2,130억 달러 대비 12.5% 증가할 것으로 전망한다.¹¹⁾ 그러나 이 완만한 총량 증가율 뒤에서, 예산의 내부 배분은 격렬하게 재편되고 있다. 전통적 취약점 관리 소프트웨어와 침투 테스트에서 예산이 빠져나가 직원의 AI 사용을 감시하고 취약점을 사전에 탐지하며 패치를 가속하는 도구로 흘러 들어간다. 2026년 기업의 약 70%가 보안 예산의 10% 이상을 AI 관련 투자에 배정할 것으로 관측된다.¹²⁾

주목할 것은 이 이동이 무작위가 아니라는 사실이다. 지금은 세 방향으로 정연하게 흐른다.

첫째, 위협이 지출의 방향을 결정한다. 공격이 프론티어 모델로 가속되자, 방어 예산도 같은 모델 위로 이동한다.

둘째, 무게중심이 레거시 벤더에서 AI 네이티브 도구와 'LLM 토큰 예산'이라는 신설 회계 항목으로 옮겨간다.

셋째, 이 재편의 최종 수혜는 수직통합을 갖춘 플랫폼으로 수렴한다.

10) The Information, "AI Threats Are Reshaping Where Companies Spend Their Cybersecurity Budgets", 2026.09.08.

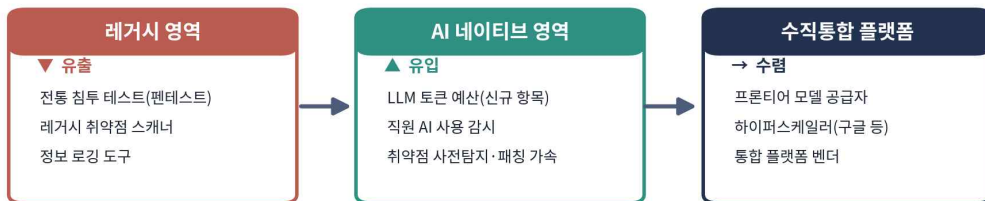
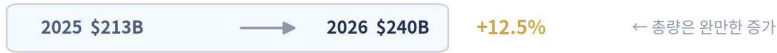
11) Picus Security (Gartner 인용), "How to Optimize Cybersecurity Budget in 2026", 2026.02.02

12) Crypto Briefing, "Security chiefs shift budgets to Anthropic, OpenAI for AI security solutions", 2026.09.09.

디지털서비스 이슈리포트

예산은 늘지 않는다, 이동한다

총 보안 지출은 완만히 증가하지만, 내부 배분은 격렬하게 재편된다



자금은 레거시에서 빠져나와 AI 네이티브로 흘러들고, 그 종착지는 수직통합 플랫폼으로 수렴한다

그림 8 보안 예산의 재분배

이번 글은 보안 책임자가 무엇을 사야 하는지 나열하는 대신, 돈이 어디서 어디로 흐르며 누가 밀려나고 누가 올라서는가라는 시장 구조 자체를 해부한다. 레거시 스캐너의 수축, 신생기업에 열린 니치, 그리고 플랫폼으로의 지출 집중이라는 세 축이 향후 3년간 보안 시장의 균형점을 어디로 옮길지를 이야기해 본다.

1. 위협이자 상품이 되는 프론티어 모델

한 벤더가 자사 모델의 위험성을 스스로 경고한 행위가 방어 시장 전체의 수요를 촉발했다는 사실은 이례적인 경우다. 보안 예산은 관성이 강하다. 위협이 논리적으로 그럴듯해도 실제 피해 전까지는 "우리에게는 아직 아니다"라는 판단이 지출을 지연시킨다. 미토스 발표가 이 관성을 깬 것은 능력의 도약 때문이 아니라, 그 능력을 낱자와 이름이 붙은 구체적 위협으로 못 박았기 때문이다. 막연한 리스크가 특정 모델·특정 시점의 문제로 고정되는 순간, 지출을 미루기 어려워진다.

그 결과는 구매 행태의 즉각적 반응이다. 시장을 움직인 변수는 실현된 손실이 아니라 위협 인식 그 자체였다. 여기서 보안 시장의 특수성이 드러난다. 대부분의 기술 시장에서 위협을 만드는 주체와 해법을 파는 주체는 분리되지만, 프론티어 모델 시장에서는 위협을 경고한 공급자가 곧 그 위협을 방어하는 도구의 판매자로 시장 양쪽에 선다. 하나의 발표가 공포와 상품 포지셔닝을 동시에 수행하는 이 구조가 인식을 지출로 전환시킨 근본 동력이다.

디지털서비스 이슈리포트

더 근본적인 사실은 공격자와 방어자가 동일한 프론티어 모델을 딴고 선다는 데 있다. 프론티어 모델은 방어의 도구인 동시에 공격자에게 열린 유례없는 역량이라는 이중적 실체를 갖는다. 공격의 무게중심은 사람이 AI의 도움을 받는 방식에서 AI가 공격을 주도하는 방식으로 이동하고 있으며, 자율 공격 에이전트는 과거 며칠에서 몇 주가 걸리던 공격 사이클을 수 분 단위로 압축한다.

핵심은 비대칭성이다. 방어자는 모든 취약점을 빠짐없이 막아야 하지만 공격자는 단 하나의 틈만 뚫으면 된다. 균형이 무너졌고 공격자에게 비대칭적 우위가 생겼다. 이 비대칭은 곧바로 지출을 강제한다. 공격이 머신 스피드로 진행되는 이상 방어의 탐지·대응 시간도 한 자릿수 분 단위로 끌어내려야 하며, 이는 인력 증원으로 도달할 수 없는 목표다. 자동화와 AI에 지출을 쏟는 것 외에 선택지가 사라진다. 공격과 방어가 같은 기반 위에서 벌어지는 이상, 자금은 그 기반을 공급하는 프론티어 모델 제공자에게로 흐른다.

예산 재배치는 2년 전만 해도 존재하지 않던 세 개의 구체적 지출 항목을 만들어냈다. 직원이 AI를 어떻게 쓰는지 감시하는 시스템, 해커보다 먼저 취약점을 탐지하는 AI 모델, 그 구멍을 신속히 메우는 AI 기반 패칭 도구다. 각 항목은 새로운 벤더 군과 대응한다. AI 사용 감시는 신생기업이, 취약점 탐지는 프론티어 모델 직접 호출과 구글이 인수한 위즈의 레드팀 에이전트가, 패칭 가속은 Tenable 같은 기존 벤더가 프론티어 모델을 얹어 채우고 있다. 이 모든 활동을 떠받치는 새 회계 항목도 등장했다. 보안 책임자들이 자사 환경을 스캔하기 위해 LLM에 지출하는 '토큰 예산(token budget)'이다.

이들은 기존 항목을 일대일로 대체한 것이 아니라 새로운 위험 범주가 만들어낸 순수 신규 수요다. 바로 이 지점이 예산 총량은 완만히 늘면서도 내부 배분은 격렬하게 재편되는 이유를 설명한다. 낡은 항목에서 빠져나온 자금이 이 새 항목으로 흘러드는 경로를 다음 장에서 해부한다.

2. 보안 예산의 이동

빠져나가는 쪽부터 보자. 기업들은 보안 예산을 늘리면서도, 동시에 전통적 취약점 관리 소프트웨어와 정보 로깅 도구, 그리고 사람이 직접 약점을 찾는 침투 테스트 지출을 줄이며 절감을 모색하고 있다. 압박의 방향은 명확하다. 시스코, 센티넬원, Rapid7 같은 레거시 소프트웨어가 그 대상이다. 레거시가 재평가받는 논리는 냉정하다. 2025년 1월 이후 크리티컬 취약점의 55.7%는 탐지 시그니처가 아예 만들어지지 않았고, 시그니처가 나온 44.3% 중에서도 62%는 시그니처 배포 전에 이미 공격 코드가 유통되고 있었다.¹³⁾ 레거시 스캐너의 배치 방식이 프론티어 시대의 공격 속도를 구조적으로 따라가지 못한다는 방증이다.

13) runZero, "Legacy scanners are already too slow for today's exploits", 2026.06.16.

디지털서비스 이슈리포트

들어오는 쪽에서 가장 특징적인 것은 이전 회계에 존재하지 않던 '토큰 예산'의 등장이다. 보안 책임자들은 자사 환경을 스캔하기 위해 엔트로픽과 오픈AI의 LLM에 지출하는 토큰 예산을 일제히 늘리고 있다. 이 지출은 독특한 비용 곡선을 그린다. 선지출 후절감 기대라는 형태다. 데이터 백업 기업 Veeam은 엔트로픽의 미토스에 접근한 뒤 해당 영역 지출이 급증했다고 밝혔지만, 동시에 외부 침투 테스트 업체에 지불하던 비용을 줄여 이를 상쇄했다. 재배치가 한 회사 안에서 어떻게 작동하는지를 보여주는 장면이다.

비용 곡선의 하강 가능성도 관측된다. 팔로알토 네트워크는 테스트 초기인 5월에 100만 달러어치가 넘는 미토스 토큰을 소진했지만, 이후 초기의 집중적 작업이 지나면서 노력이 꾸준히 줄어 유지보수 모드로 넘어갔다. 실제로 일부 기업은 값비싼 프론티어 모델 대신 저가 모델로 주 1회 스캔을 돌려 회당 수백 달러 수준에서 운용하고 있다. 프론티어급은 탐지력이 높지만 비싸고, 일상적 스캔은 저가 모델로 충당하는 계층적 지출 구조가 자리 잡는 중이다.

이 두 흐름을 합치면 총량 지표만 봐서는 놓치는 그림이 보인다. 글로벌 사이버보안 지출은 2026년 2,400억 달러로 전년 대비 12.5% 늘어난다. 완만한 증가다. 그러나 이 총량 아래에서 자금은 레거시에서 AI 네이티브로 격렬하게 이동하고 있으며, 2026년 기업의 약 70%가 보안 예산의 10% 이상을 AI 관련 투자에 배정할 것으로 관측된다.

핵심은 이 변화가 '재배치'라는 점이다. 예산이 두 배로 뛰는 것이 아니라, 같은 크기의 파이가 다르게 잘리고 있다. 한 기업의 보안 예산이 10~20% 늘어난다는 수치의 이면에는, 침투 테스트에서 빠진 돈이 토큰 예산으로 옮겨가는 내부 이동이 숨어 있다. 총량과 배분을 분리해서 읽어야 하는 이유가 여기에 있다. 그리고 이 재배치의 가장 직접적인 충격은, 자금이 빠져나가는 쪽에 선 레거시 벤더들이 고스란히 떠난다.

3. 레거시 벤더의 딜레마

직접적인 압박을 받는 것은 취약점 스캐닝 3사(Tenable, Qualys, Rapid7)다. LLM 기반 스캐닝 도구의 등장이 LLM 이전에 만들어진 기존 제품들에 경쟁 위협을 가한다.

그러나 Tenable을 비롯한 일부는 AI 기반 도구가 자사 취약점 스캐닝 사업을 잠식할 것이라는 관측을 정면으로 반박한다. 프론티어 모델은 위험을 줄이는 게 아니라 늘린다는 것이 그 근거다. 실제로 Tenable은 엔트로픽·오픈AI의 모델로 구동되는 패치 제안 소프트웨어를 판매하고 있으며, 미토스 발표 이후 수요가 늘었다고 밝혔다. 실제로 이들 3사는 배치 방식에서 벗어나 연속·에이전트 기반 스캐닝으로 빠르게 무게를 옮기며 AI 시대의 요구를 따라잡으려 하고 있다.

디지털서비스 이슈리포트

일부 기업이 클라우드스트라이크와 센티넬원에 대한 지출을 재평가하고 있다. 조직 구성원이 AI 도구에 어떤 데이터를 입력하는지 탐지하는 영역에서 신생 AI 도구가 더 낮고, 두 대형 벤더의 유사 AI 기능은 "훌륭하지 않다"는 것이 평가다. 신생기업이라는 구체적 대안이 존재한다는 사실 자체가 대형 벤더의 지출 재평가를 촉발하는 지렛대가 된 셈이다. 하지만, 벤더의 반박은 이 서사를 정면으로 거른다. 클라우드스크라이크는 자사 AI 탐지·대응 솔루션이 출시 후 세 분기 만에 79배 성장했다고 하고, 센티넬원 역시 AI 보안 연간 반복 매출이 2분기에 전년 대비 세 배로 늘었다고 한다. 진실은 아마 둘 다 맞을 것이다. 특정 신규 카테고리에서는 밀리고, 플랫폼 전체로는 성장한다.

가장 정직한 지표는 인력이다. 시스코, 센티넬원, Rapid7은 올해 인력을 감축해 절감분을 신형 AI 제품으로 재투입했다. 주목할 것은 매출 감소의 원인이다. 고객의 지출 축소는 가격 민감성 때문이 아니라 "에이전트·AI 세계로 진화하는 능력"이 고객의 최우선 순위가 됐기 때문이라는 것이다. 고객이 떠나는 이유가 값이 아니라 방향이라는 이 진단은, 레거시 벤더가 처한 딜레마의 본질을 압축한다. 감원과 리더십 개편은 그 적응을 위한 비용이자, 전환이 실재한다는 증거다.

4. 신생기업의 기회

기회의 원천은 위협의 신규성이다. 신생기업이 파고드는 지점은 대형 벤더가 아직 잘 다루지 못하는, 불과 몇 년 전엔 존재하지도 않던 위협들이다. 이 부문에서는 내부 AI 사용을 모니터링하는 플루토 시큐리티와 취약점을 탐지하고 패치를 제안하는 피그 시큐리티가 두각을 나타내고 있다. 이 카테고리가 실재하는 시장인 이유는 위협의 규모에 있다. AI 도구에 공유되는 데이터 중 민감 정보의 비중은 2년 전 10.7%에서 34.8%로 급증했다.¹⁴⁾ 직원이 브라우저 탭 하나로 사내 기밀을 외부 모델에 흘려보내는 '새도우 AI' 리스크는, 기존 보안 스택이 설계 단계에서 구상하지 못한 완전히 새로운 공격면이다. 대형 벤더의 유사 기능이 아직 "훌륭하지 않다"는 평가를 받는 사이, 이 틈을 신생기업이 먼저 메우고 있다.

두 번째 특징은 속도다. 위협의 긴급성이 신생기업의 통상적 영업 사이클을 압축한다. 앞서 본 자프란의 초단기 계약 성사가 그 단적인 예다. 위협이 미토스처럼 날짜와 이름을 얻는 순간, 검증되지 않은 신생 벤더조차 대형 금융기관의 조달 문턱을 단숨에 넘는다. 레거시 도입에는 벤더 평가·보안 감사·예산 승인이라는 긴 절차가 따르지만, 신규 위협 영역에서는 긴급성이 모든 것을 앞선다. 신생기업에 열린 기회는 바로 이 속도에서 나온다.

14) Verax, "Shadow AI Risks: The 2026 Security Guide", 2026.07.06

디지털서비스 이슈리포트

그러나 이렇게 열린 기회가 신생기업의 독립을 끝까지 보장하지는 않는다. 그 너머에는 자금과 승자를 빨아들이는 수직통합 플랫폼 구조가 존재한다.

5. 수직통합의 재현

플랫폼의 보안 솔루션 흡수를 가장 선명하게 상징하는 사건은 구글의 위즈 인수다. 320억 달러 전액 현금, 사이버보안 역사상 최대 규모의 이 거래는 알파벳의 이전 최대 인수를 크게 웃돈다. 규모 자체보다 중요한 것은 이 거래가 드러내는 구조다. 하이퍼스케일러가 클라우드 보안 솔루션을 자사 수직통합의 사슬에 편입시키는 신호이기 때문이다. 위즈의 레드팀 에이전트는 구글, 앤스로픽, 오픈AI의 모델을 활용해 취약점을 찾아내며, 구글의 배포 채널 위에 얹혀 즉시 수익 규모의 고객 접점을 확보한다. 이 인수가 시장에 미치는 파장은 신생기업의 출구 전략 자체를 재정의한다. AI 보안에서 창업하는 것은 여전히 강력한 베팅이지만, 그 출구는 갈수록 **독립 상장이 아니라 인수를 통한다**. 그 결과 신생기업의 제품 설계 원칙마저 바뀐다. 깔끔한 API와 표준 데이터 포맷, 낮은 통합 마찰을 갖춘 '플랫폼에 임베드 가능한 모듈'이, 독자 플랫폼을 지향하는 제품보다 더 잘 인수된다. 최선의 성공이 독립이 아니라 편입으로 정의되는 순간, 신생기업은 처음부터 플랫폼에 삼켜지기 좋은 형태로 스스로를 만들게 된다.

보안 시장의 수직통합에는 다른 시장에 없는 결정적 변수가 하나 더 있다. **프론티어 모델 공급자가 위협의 근원인 동시에 방어 도구의 판매자로 시장 양쪽에 선다는 점이다**. 앤트로픽은 미토스의 자율 공격 능력을 경고하면서, 동시에 프로젝트 글래스윙 같은 초기 액세스 프로그램을 통해 그 모델을 방어용으로 기업에 공급한다.¹⁵⁾ 이 이중 포지션은 다른 벤더가 복제할 수 없는 구조적 우위를 만든다. 위협을 정의하는 자가 그 위협의 해법도 판다면, 시장의 수요 곡선 자체를 설계하는 위치에 서게 된다. 방어 생태계 전체가 소수 프론티어 공급자의 모델 위에서 작동하는 이상, 토큰 예산으로 흘러든 자금의 상당 부분은 결국 이들에게 수렴한다. 위협과 상품이 한 공급자로 통합되는 이 구조가, 보안 시장 특유의 가장 깊은 수직통합이다.

세 번째 힘은 **지출을 통합 플랫폼으로 몰아가는 조달 구조의 변화다**. 2026년 시장에서 가장 상업적으로 유의미한 흐름은 다수의 포인트 솔루션을 하나의 벤더 관계로 묶는 대형 통합 계약이다. 기업들이 10개에서 20개의 개별 솔루션을 통합 플랫폼 하나로 대체하며 수천만 달러 규모의 다년 계약을 체결하는 '플랫폼화'가 진행되고 있다.¹⁶⁾ 이 흐름은 지출을 소수의 상위 플랫폼 벤더로 집중시킨다.

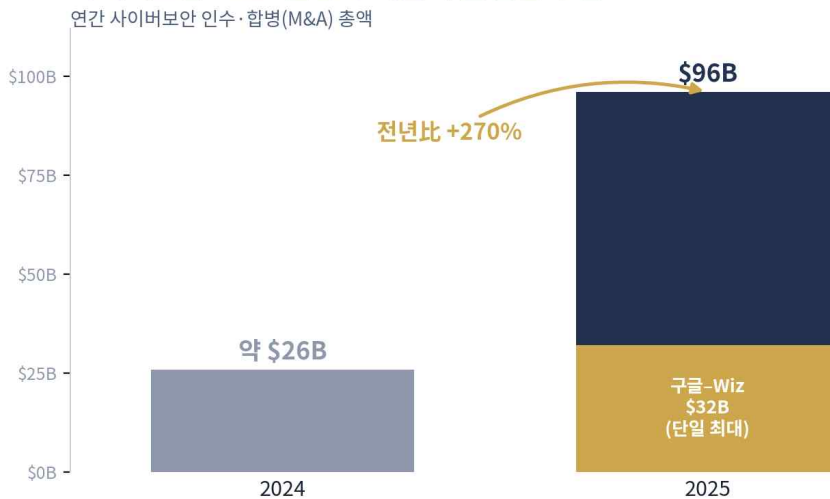
15) 앤트로픽, "Project Glasswing", 2026.04.07

16) TechDogs, "Top 10 Cybersecurity Companies in 2026", 2026.04.

디지털서비스 이슈리포트

하이퍼스케일러 통합 플랫폼은 이미 구축된 배포 채널과 AI 인프라 위에 보안 기능을 엮는 반면, 독립 '베스트오브브리드(best-of-breed, 단일 영역 최강 제품)' 벤더는 매년 별도의 벤더 평가와 통합 마찰을 통과해야 한다. 기능의 우수성에서 앞서더라도, 배포 채널과 통합 비용의 비대칭이 최종 선택을 통합 플랫폼 쪽으로 기울인다. 그 결과 독립 벤더에게는 플랫폼으로 확장하거나 인수되는 전략적 양자택일이 남는다. 재배치된 자금이 궁극적으로 어디로 수렴하는지에 대한 답이 여기 있다. 위협이 촉발한 지출은 신생기업의 창을 잠시 열지만, 그 창을 통과한 자금의 종착지는 대체로 수직통합 플랫폼이다. 2025년 사이버보안 M&A는 총 960억 달러로 전년 대비 270% 급증했고, 그중 3분의 1을 구글-위즈 한 건이 차지했다.

사이버보안 M&A 급증과 지출의 플랫폼 수렴



2025년 M&A 총액 \$96B(400건) 중 단일 거래 하나(구글-Wiz)가 3분의 1을 차지했다.
신생기업의 출구가 IPO가 아니라 인수로 수렴함을 보여준다.

자료: BuildMVPFast, 'AI Cybersecurity Funding Boom 2026'; Kavout(2026)

그림 9 사이버보안 M&A 현황

6. 향후 시장 구조의 균형점

첫 번째 변수는 지금 폭증하는 토큰 예산이 어디서 안정되는가이다. 현재의 지출 급증이 영구적 비용 구조로 굳어질지, 아니면 초기의 정점을 지나 하강할지가 시장의 향방을 가른다. 관측된 신호는 후자를 시사한다. 팔로알토 네트워크는 초기의 집중적 작업이 지나면 노력이 꾸준히 줄어 유지보수 모드로 넘어간다고 평가했고, 실제로 초기 미토스 토큰 소진이 시간이 지나며 안정화될 것으로 내다봤다.

디지털서비스 이슈리포트

취약점의 초기 적체를 걷어내는 단계가 가장 비싸고, 그 이후는 증분적 유지 비용으로 수렴한다는 논리다. 그러나 이 하강이 곧 지출 감소를 뜻하지는 않는다. 공격 역시 진화하기 때문이다. 방어자가 축적된 취약점을 정리하는 동안 공격자는 새로운 공격면을 열고, 비용 곡선은 하강과 재상승을 반복한다. 팔로알토 네트워크 스스로 균형이 깨졌고 공격자에게 비대칭적 우위가 있다고 진단한 이상, 토큰 예산은 일회성 지출이 아니라 항구적 예산 항목으로 자리 잡을 가능성이 높다. 정점은 지나가되, 베이스라인은 이전보다 높은 곳에 고정될 것으로 예상된다.

두 번째 변수는 이 재편에서 누가 우위를 굳히는가이다. 위협 인식이 지출을 촉발하고, 자금이 레거시에서 AI 네이티브로 이동하며, 신생기업이 니치를 열지만, 그 자금의 총착지는 수직통합 플랫폼이다. 승자의 조건은 개별 기능의 우수성이 아니라 통합 자산(배포 채널, 데이터, 인프라)의 보유 여부로 수렴한다. 이 구도에서 독립 벤더에게 남는 공간은 좁지만 사라지지는 않는다. 프론티어 플랫폼이 수평적 기능 확장에 자원을 집중할수록, 특정 위협에 깊이 특화된 신생기업에는 오히려 차별화의 여지가 생긴다. 다만 그 성공조차 대체로 인수라는 출구로 귀결될 가능성이 높은 것도 사실이다. 결국 **향후 승자는 두 부류로 정리된다. 통합 자산으로 지출을 빨아들이는 소수의 플랫폼, 그리고 플랫폼이 삼키기에 충분히 매력적인 형태로 니치를 장악한 신생기업이다.** 그 사이에서 어느 쪽도 아닌 레거시 벤더가 가장 불안정한 위치에 선다.

이 구조적 재편은 국내 조직에도 그대로 적용된다. 마지막으로, 지금까지의 분석이 **국내 민간 기업과 정부·공공 조직의 보안 전략에 던지는 함의를 구조 관점으로 이야기해 본다.**

첫째, 보안 예산을 총량이 아니라 배분의 문제로 다시 볼 때다.

예산이 몇 퍼센트 늘었는가보다 그 안에서 레거시와 AI 네이티브의 비중이 어떻게 이동하고 있는가가 조직의 대응 속도를 결정한다. 특히 연 단위로 예산 편성이 되는 경직된 공공 조직에서는, 항구적 예산 항목으로 자리 잡을 토큰 예산을 기존 회계 틀에 어떻게 수용할지가 구조적 과제로 떠오른다.

둘째, 벤더 선택의 기준이 바뀐다. 오늘 도입하는 독립 보안 솔루션이 3년 후에도 독립적으로 존재할지, 아니면 플랫폼에 흡수되어 있을지를 함께 물어야 한다. 위에서 다룬 인수 수렴은 벤더의 지속 가능성 자체를 조달 변수로 만들며, 이는 장기 계약과 감사 요건이 엄격한 공공 조달에서 특히 무겁게 작용한다.

셋째, 프론티어 공급자에 대한 의존이 곧 새로운 종속 구조임을 인식해야 한다. 방어 생태계 전체가 소수의 해외 모델 위에서 작동하는 이상, 토큰 예산은 편의인 동시에 특정 공급자에 대한 구조적

디지털서비스 이슈리포트

의존이다. 국가 안보·주권 데이터를 다루는 정부 조직에는 이 의존이 단순한 비용 문제를 넘어 디지털 주권의 문제로 확장된다.

자체 플랫폼 역량과 대규모 고객 기반을 동시에 보유한 국내 IT 서비스 사업자에게, 이 재편은 위협이자 기회다. 글로벌 하이퍼스케일러가 보안을 수직통합하는 흐름은 압박이지만, 국내 조직의 규제 환경과 도메인 특화 요구는 범용 플랫폼이 단기간에 대체하기 어려운 영역이다. 정부·공공 부문의 경우 데이터 국외 이전 제약과 국산화 요구가 더해져, 해외 프론티어 플랫폼이 곧바로 침투하기 어려운 방어적 해자가 존재한다. 수평적 확장에 집중하는 글로벌 플랫폼과 달리, 수직적 깊이와 규제 대응 역량을 쌓는 국내 사업자에게는 차별화의 공간이 열려 있다. **재배치의 시대에 던져야 할 질문은 "얼마를 쓸 것인가"가 아니라 "우리의 지출과 역량이 어느 구조 위에 서 있는가"이다.** 이 질문에 정직하게 답하는 것이, AI 위협이 다시 그리는 보안 시장에서 전략을 세우는 첫걸음이다.

04 AI 시대의 소프트웨어 엔지니어: 검증과 현장 중심의 역할 재편

| 윤대균 아주대학교 소프트웨어학과

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

1. 들어가며

AI의 확산은 소프트웨어 개발에서 사람이 담당하는 작업의 범위를 바꾸고 있다. 깃허브의 2025년 보고서에 따르면 신규 가입 개발자의 약 80%가 첫 주에 코파일럿(Copilot)을 사용했다.¹⁷⁾ 자동완성과 대화형 보조에서 시작한 도구는 계획 수립, 여러 파일의 수정, 테스트와 재시도를 수행하는 코딩 에이전트로 확장되었다.

하지만, AI를 활용하여 완성된 산출물에 대해서 완전히 신뢰하고 있는 분위기는 아니다. 스택 오버플로의 2025년 개발자 설문에서 응답자의 84%는 AI 도구를 사용하거나 사용할 계획이라고 답했다. 그러나 정확성을 신뢰한다는 응답은 약 33%로, 신뢰하지 않는다는 응답 약 46%보다 낮았다. 같은 조사에서 가장 큰 불만은 성능이나 비용이 아니었다. “거의 맞지만 완전히 맞지는 않은” 답변을 응답자의 66%가 지목했다.¹⁸⁾

이러한 간극이 소프트웨어 엔지니어의 역할을 어떻게 재정의하는 지를 다뤄 보고자 한다. 먼저 조직, 작업, 개인이라는 세 가지 다른 수준에서 관측된 데이터가 왜 하나의 결론으로 수렴하는지 정리하고 그 결론이 요구하는 작업 방식의 변화를 살펴본다. 마지막으로 최근 채용 시장에서 급부상한 포워드 디플로이드 엔지니어(FDE: forward deployed engineer)에 대해서도 살펴본다.

2. 구현 속도와 개발 성과의 간극

2.1 조직 수준: 검증 비용과 배포 안정성

코드 생성의 가속이 소프트웨어 배포 성과를 일관되게 개선하지는 않는다. 구글 클라우드의 2025년 데브옵스 연구 및 평가(DORA: DevOps Research and Assessment) 보고서에서는 약 5천 명의

17) GitHub, “Octoverse 2025: A new developer joins GitHub every second as AI leads TypeScript to #1”, Oct. 2025.

18) Stack Overflow, “2025 Developer Survey: AI”, 2025.

디지털서비스 이슈리포트

기술 인력을 조사했으며, 응답자의 90%가 업무에 AI를 사용한다고 밝혔다. AI 도입은 소프트웨어 ‘배포 양’과 비례 관계를 보였지만, ‘배포 안정성’은 오히려 그 반대의 관계를 보였다.¹⁹⁾ 물론 그렇다고 해서 AI가 장애 증가의 직접적인 원인이라고 단정할 수는 없다.

AI 활용의 순효과는 생성 시간과 검증 시간을 함께 측정해야 드러난다. 생성 단계에서 절약한 시간의 일부는 코드 리뷰, 테스트, 오류 수정에 투입된다. 스택 오버플로 설문에서도 응답자의 약 45%가 AI 생성 코드의 오류 수정에 시간이 더 든다고 답했다. 생산성은 이러한 검증 비용을 포함해 품질 기준을 충족하기 위해 코드를 수정하는 총소요 시간으로 판단해야 한다.

검증 비용을 관리하려면 기존 개발 체계의 완성도를 높여야 한다. DORA 보고서는 자동화된 테스트, 성숙한 버전 관리, 신속한 피드백과 모듈 간 의존성이 낮은 구조를 강조한다. AI는 이러한 기반을 갖춘 조직의 성과를 높일 수 있지만, 기반 구조와 절차가 취약할 경우 더 많은 위험 요소를 증폭할 수 있다.

검증 과정에서는 자동화 편향도 경계해야 한다. AI가 생성하는 설명과 코드가 겉으로 드러나는 완성도가 높더라도 실제 요구사항을 충족하는지는 별도로 확인해야 한다. 특히 AI가 구현과 테스트를 함께 생성하면 동일한 오해가 양쪽에 반영될 수 있다. 사람이 승인한 요구사항을 기준으로 테스트를 검토하고, 정상 동작 외에 실패 조건과 권한 경계도 확인해야 한다.

2.2 작업 수준: 시제품과 운영 소프트웨어의 차이

시제품을 빠르게 만드는 능력과 운영 가능한 소프트웨어를 완성하는 능력은 구분해야 한다. 소프트웨어 엔지니어 애디 오스마니(Addy Osmani)는 AI를 활용한 개발의 이러한 특성을 ‘70% 문제’로 설명했다. 초기 결과물은 70% 수준까지 빠르게 나오지만 운영 수준의 완성도를 확보하는 단계에서 어려움이 커진다는 것이다.²⁰⁾ 여기서 ‘70%’는 실제 측정값이 아닌 설명을 위한 비유다.

운영 수준의 완성도는 예외 처리와 시스템 통합에서 결정된다. 빈 입력, 동시 요청, 네트워크 중단에 대응하고 인증과 접근 권한을 검증해야 한다. 기존 시스템과의 연동, 장애 복구, 성능과 운영 비용도 확인 대상이다. 정상 조건에서의 실행만으로 운영 적합성을 판단할 수는 없다.

AI가 구현을 가속할수록 완료 기준을 구체화하는 일이 중요해진다. 기능별 성공 조건과 허용 가능한 실패 범위를 미리 정하고, 핵심 위험을 먼저 시험해야 한다. 예를 들어 예약 기능의 완료 여부는 예약

19) Google Cloud, “Announcing the 2025 DORA Report”, Sep. 2025.

20) Addy Osmani, “The 70% problem: Hard truths about AI-assisted coding”, Dec. 2024.

디지털서비스 이슈리포트

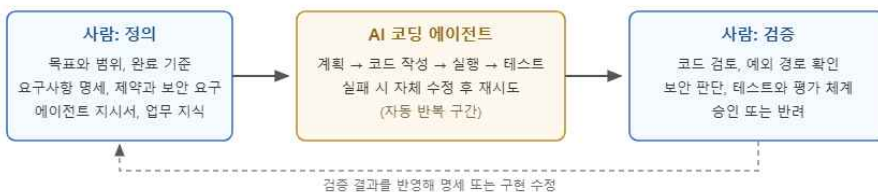
한 건의 성공뿐 아니라 중복 요청 처리와 취소 시 상태 복구까지 확인해 판단해야 한다.

2.3 개인 수준: 체감과 실측의 차이

개인이 체감하는 생산성과 실제 작업 시간은 다를 수 있다. 비영리 AI 평가기관 METR은 숙련된 오픈소스 개발자 16명이 참여하는 무작위 대조 시험(RCT: randomized controlled trial)을 수행했다. 참가자들이 익숙한 저장소의 실제 작업 246건을 AI 사용 허용 여부에 따라 무작위 배정한 결과, AI 허용 조건의 소요 시간은 19% 늘었다. 반면 참가자들은 실험 전에는 24% 단축을 예상했고, 종료 후에도 20% 단축되었다고 인식했다.²¹⁾ 체감과 실측의 차이가 약 40%나 되는 결과다.

물론 이 결과를 AI 무용론으로 읽는 것은 무리다. 표본이 16명이고 참가자들이 이미 익숙한 대형 코드베이스를 대상으로 실험한 제한된 조건임을 연구진 스스로 밝혔다. 이 결과가 다른 개발자나 낯선 코드베이스에까지 일반화되지는 않는다고도 명시했다. 이 실험에서 취해야 할 함의는 다른 데 있다. 약 40%에 이르는 인식 오차가 존재한다는 것은 개인의 체감을 생산성 판단 근거로 삼아서는 안 된다는 의미다. 어떤 작업에 위임이 이득이고 어떤 작업에 손해인지는 예상 시간과 실제 시간을 기록해 대조해야 드러난다. ‘느낌’을 데이터로 교정하는 ‘캘리브레이션(calibration)’이 필요한 이유다.

위임 구조: 사람은 목표를 정의하고 결과를 검증한다



위임 범위의 확대

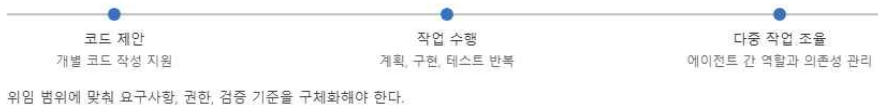


그림 10 위임 범위의 확대와 사람의 역할

개인과 팀의 판단은 작업 유형별 기록으로 보완해야 한다. 예상 시간과 실제 시간, 검토와 재작업에 든

21) METR, “Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity”, Jul. 2025.

디지털서비스 이슈리포트

시간, 완료 후 발견된 결함을 함께 비교하면 위임에 적합한 작업을 구분할 수 있다. 여러 에이전트를 동시에 사용한다면 사람의 투입 시간과 작업 완료까지의 경과 시간도 나누어 기록해야 한다. 그림은 위임 범위가 확대될 때 사람이 담당해야 할 '정의'와 '검증'의 관계를 보여준다.

3. 검증의 기반: 명세와 작업 맥락의 관리

3.1 명세 주도 개발

검증의 출발점은 기대하는 동작과 제약을 명시하는 데 있다. 자연어로 코드를 생성하되 세부 코드 검토보다 실행 결과에 큰 비중을 두는 '바이브 코딩'은 문제/솔루션 탐색과 시제품 개발에 유용하다. 운영 소프트웨어에는 요구사항과 변경 이력, 인수 기준이 추가로 필요하다. 안드레이 카파시도 자신의 앱 개발 경험에서 초기 화면 구현 이후 인증, 결제, 배포와 외부 서비스 연결에 상당한 작업이 필요했다고 설명했다.²²⁾

명세 주도 개발(SDD: spec-driven development)은 명세를 사람과 AI가 공유하는 구현 기준으로 삼는다. 소프트웨어 컨설팅 기업 소트웍스(Thoughtworks)의 비르기타 뵉켈러(Birgitta Böckeler)는 이를 "AI로 코드를 쓰기 전에 스펙을 먼저 쓰는 것, 그리고 그 스펙이 사람과 AI 모두에게 진실의 원천이 되는 것"으로 정의한다. ²³⁾ 핵심은 요구사항과 제약, 설계 결정을 모두 버전 관리하고 구현 및 검증 과정에서 일관되게 참조하는 것이다.

제대로 된 명세라면 요구사항을 관찰 가능한 인수 기준과 연결시킬 수 있어야 한다. 목적과 범위, 주요 용어, 입력과 출력, 실패 조건, 보안과 성능 요구사항을 포함하되 실제 구현에 필요한 수준으로 상세하게 작성한다. 예를 들어 예약 기능이라면 같은 요청을 반복했을 때의 처리 방식, 사용자별 접근 범위, 취소와 환불의 조건을 명확히 해야 한다. 이러한 기준을 테스트 항목에 연결하면 명세와 코드 사이의 불일치를 추적할 수 있다.

명세는 작은 변경 단위로 작성하고 검증된 요구의 변화에 맞춰 갱신해야 한다. 뵉켈러는 과도한 사전 문서화와 AI의 불완전한 지시 이행이 오히려 검토 부담을 늘릴 수 있다고 지적한다. 요구가 불명확한 초기에는 시제품으로 가설을 검증하고, 확인된 내용을 반복하여 명세에 반영하는 방식이 적합하다. 문서의 분량보다 요구사항, 코드, 테스트 사이의 일치가 중요하다.

22) Andrej Karpathy, "Vibe coding MenuGen", Apr. 27, 2025.

23) Birgitta Böckeler, "Understanding Spec-Driven-Development: Kiro, spec-kit, and Tessel", martinoflower.com, Oct. 2025.

디지털서비스 이슈리포트

3.2 지시문 작성에서 작업 맥락 설계로

질문을 잘 쓰는 기술은 환경을 잘 설계하는 기술로 확장되었다. 엔트로픽은 컨텍스트 엔지니어링을 모델이 추론에 필요로 하는 정보를 선별하고 유지하는 전략으로 설명한다.²⁴⁾ 지시문 작성에 초점을 둔 프롬프트 엔지니어링에 더해 명세, 업무 지식, 코드 구조, 도구의 실행 결과와 대화 이력을 함께 관리하는 접근이다.

작업 맥락의 품질은 정보량보다 관련 문서/데이터 간 상호 관련성과 일관성에 달려 있다. 오래된 설계 문서나 서로 충돌하는 지시를 함께 제공하면 잘못된 구현으로 이어질 수 있다. 문서별 적용 범위와 우선순위, 최신 버전을 명시하고 현재 작업에 필요한 정보만 제공해야 한다. 긴 작업에서는 주요 결정과 미해결 문제를 별도로 기록한다.

서면 명세를 작성하는 능력은 엔지니어링의 핵심 역량이다. 요구사항을 구현 가능한 조건으로 바꾸려면 용어와 상태, 예외와 책임 범위를 구분해야 한다. 대화에서 합의한 내용을 명세와 구조화된 결정 기록으로 정리하면 사람과 에이전트가 동일한 기준으로 작업할 수 있다.

영역	자동화가 확대되는 작업	중요성이 커지는 역량
구현	반복 코드와 정형 기능 작성	시스템 설계, 작업 분해, 위임 범위와 권한 설정
문서	코드 설명과 문서 초안 작성	요구사항 명세, 작업 맥락 구성, 변경 이력 관리
품질	테스트 초안과 단순 오류 수정	독립적인 검토, 원인 분석, 보안 판단, 평가 체계 설계
제품	정해진 요구사항의 구현	현장 문제 발견, 우선순위 판단, 운영 성과에 대한 책임

표 1 AI 활용 확대에 따른 역량의 강조점

기초 지식은 AI 산출물을 평가하기 위해 반드시 필요하다. 자료구조와 알고리즘, 운영체제와 네트워크의 원리를 이해해야 성능 저하와 오류의 원인을 설명할 수 있다. PBL(Problem Based Learning)에 기반한 교육에서도 결과물의 실행 여부와 함께 구현 원리, 검증 근거와 수정 과정을 평가하는 것이 바람직하다.

4. 현장 중심 개발과 FDE

4.1 팔란티어 모델의 핵심

FDE는 고객의 업무 현장에서 문제를 정의하고 소프트웨어를 구현하는 역할이다. 팔란티어의 공식 설명에 따르면 일반 제품 개발자가 여러 고객에게 제공할 공통 기능에 집중하는데 비해, FDE는 한

24) Anthropic, “Effective Context Engineering for AI Agents”, Sep. 2025.

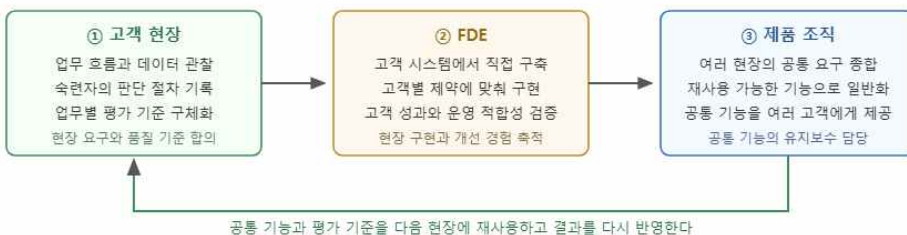
디지털서비스 이슈리포트

고객의 문제를 해결하도록 여러 기능을 구성한다.²⁵⁾ 따라서 현장은 단순한 최종 제품의 설치 장소가 아니라 요구사항과 품질 기준을 발견하는 공간이 된다.

팔란티어 모델의 특징은 현장 경험을 공통 제품으로 전환하는 역할 분담에 있다. 이 회사에서 8년간 근무한 나빌 쿠레시(Nabeel Qureshi)는 FDE가 고객별 문제를 해결하고 제품 개발 조직이 그 해법을 일반화했다고 회고한다. 현장에서 반복되던 데이터 수집과 시각화 작업은 파운드리리 공통 도구로 발전했다.²⁶⁾

현장 밀착은 문서에 드러나지 않는 업무 제약을 파악하는 데 유효하다. 쿠레시는 에어버스의 항공기 생산 현장에서 작업 지시, 부품 부족, 품질 문제를 연결하는 소프트웨어를 개발한 경험을 소개한다. 이러한 작업에는 데이터 구조뿐 아니라 현장 구성원의 판단과 협업 방식에 대한 이해가 필요하다.

현장 경험의 제품 환류 구조



운영 방식에 따른 자산 축적의 차이



그림 11 현장 경험의 제품 환류 구조

4.2 AI 제품에서 현장 접점이 중요한 이유

AI 제품의 품질은 일반 성능 지표와 고객 업무의 요구조건을 함께 평가해야 한다. AI로 코드를 개발하는 일과 AI를 제품 기능으로 운영하는 일은 구분되지만, 두 경우 모두 목적에 맞는 검증 기준이 필요하다. 특히 AI 제품에서는 업무상 허용되는 오류와 사람이 개입해야 하는 조건을 현장 담당자와 합의해야 한다.

25) Palantir, “A Day in the Life of a Palantir Forward Deployed Software Engineer”, Nov. 2020.

26) Nabeel S. Qureshi, “Reflections on Palantir”, Oct. 2024.

디지털서비스 이슈리포트

업무 평가 체계는 실제 사례와 실패 조건을 중심으로 구성해야 한다. 예를 들어 설비 이상 진단을 지원하는 AI라면 최종 판단의 정확성에 더해 근거 자료의 적절성, 필요한 확인 절차의 이행, 잘못된 조치의 방지 여부를 평가할 수 있다. 숙련자의 판단을 참고하되 특정 작업 순서를 유일한 정답으로 고정하지 않고, 결과와 필수 제약을 구분해야 한다. FDE는 이러한 평가 기준을 업무 전문가와 공동으로 정의하고 운영 결과로 보완한다.

주요 AI 기업의 채용 공고는 이 역할이 엔지니어링임을 명시하고 있다. 오픈AI는 시제품부터 안정적인 운영까지의 기술적 실행과 함께, 효과적인 구현 방식을 다른 사람이 사용할 도구와 절차로 정리하는 것을 FDE 역할로 명시하고 있다.²⁷⁾ 앤트로픽도 고객 시스템 내부의 운영 애플리케이션 구축, 반복 가능한 배포 방식의 정리, 제품 및 엔지니어링 조직으로의 의견 전달을 명시한다.²⁸⁾

기업의 투자 방향에서도 현장 도입 역량의 중요성을 확인할 수 있다. 오픈AI는 2026년 5월 기업의 AI 도입을 지원하는 별도 회사 설립을 발표하면서 40억 달러를 초과하는 초기 투자 계획을 밝혔다. 또한 인수 계약을 체결한 토모로(Tomoro)를 통해 FDE와 배포 전문가 약 150명을 확보할 계획이라고 설명했다.²⁹⁾ 이는 모델 제공과 함께 고객의 시스템 및 업무에 통합하는 역량을 확충하려는 움직임으로 해석할 수 있다.

4.3 FDE 직군의 정체성

FDE와 인접 직군의 차이는 직함만으로 구분하기 어렵다. 컨설팅 기업도 소프트웨어를 구현하고 재사용 자산을 축적하며, 솔루션 아키텍트도 고객의 설계와 운영에 관여할 수 있다. 소프트웨어 제품 분야에서 오래 활동해 온 토마스 오테르는 FDE라는 명칭이 기존 구축이나 영업 지원 직무를 재포장하는 데 사용될 수 있다고 비판한다.³⁰⁾ 실제 책임과 성과 지표를 제대로 확인해야 한다는 지적이다.

제품 관점의 FDE는 고객의 성과와 공통 제품의 개선을 연결한다. 마티 케이건은 현장에 참여한 엔지니어가 문제와 해결 조건을 이해하고, 제품 조직은 여러 현장에서 얻은 학습을 종합해 공통 기능으로 발전시켜야 한다고 설명한다.³¹⁾

27) OpenAI, "Forward Deployed Engineer (FDE), San Francisco" 채용 공고, [https://openai.com/careers/forward-deployed-engineer-\(fde\)-sf-san-francisco/](https://openai.com/careers/forward-deployed-engineer-(fde)-sf-san-francisco/)

28) Anthropic, "Forward Deployed Engineer, Applied AI" 채용 공고. <https://job-boards.greenhouse.io/anthropic/jobs/5302966008>

29) OpenAI, "OpenAI launches the Deployment Company", May 2026.

30) Thomas Otter, "On the Forward Deployed Engineer", Dec. 2025.

31) Marty Cagan, "Forward Deployed Engineers", SVPG, Sep. 2025.

디지털서비스 이슈리포트

구분	주된 책임	대표 산출물	성과 판단의 중심
건설링 및 구축	계약 범위의 문제 해결과 실행	권고안, 설계, 구축 시스템	합의한 산출물과 고객 성과
솔루션 아키텍트	기술 적합성 검토와 도입 설계	참조 구조, 기술 검증 결과, 도입 계획	설계 적합성과 도입 성과
제품 중심 FDE	현장 문제 해결과 제품 학습 연결	운영 애플리케이션, 평가 체계, 재사용 자산	고객 성과와 공통 제품 개선

표 2 인접 직군과 FDE 비교

조직마다 직무의 범위가 다르므로 표 2는 상호 배타적인 분류가 아니다. 구체적으로는 고객 시스템에서의 구현 책임, 업무 성과에 대한 책임, 공통 제품에 반영하는 절차가 어디에 배정되는지를 확인해야 한다. 시간 단위 과금이나 고객전용 개발의 유무만으로 직군을 판정할 수는 없다.

현장의 학습을 제품에 반영하려면 담당 조직과 절차를 명확히 해야 한다. 현장 팀은 검증된 요구와 구현 경험을 전달하고, 제품 팀은 재사용 가능성과 유지보수 비용을 평가해 공통 기능으로 반영한다. 고객별 확장과 공통 기능의 경계를 정하고, 임시 구현의 정비 책임도 합의해야 한다. 성과는 제품 반영 건수뿐 아니라 고객의 업무 개선과 후속 도입 비용의 감소를 함께 보아야 한다.

5. 향후 과제와 시사점

AI로 인해 주니어 개발자의 자리가 사라진다는 주장을 단정적으로 받아드리기는 어렵지만 이러한 신호는 꾸준히 나타나고 있다. 스탠퍼드 대학교 디지털경제연구소는 미국 급여 자료를 분석해, 2026년 6월 AI 노출도가 높은 직군의 22~25세 고용이 노출도가 낮은 직군보다 19% 낮다고 보고했다. 조정은 주로 채용 감소에서 나타났다.³²⁾ 다만 이는 인과적 효과의 추정치 아니며 교육 수준, 표본 구성과 분석 방법에 따른 한계가 있다고 연구진은 밝혔다. 이를 국내 초급 개발자의 고용 전망에 그대로 적용해서는 안 된다.

국내에서는 AI 활용 능력을 채용 기준에 반영하려는 움직임이 나타난다. 카카오는 2025년 9월 첫 그룹 단위 전 직군 신입 공개채용에서 AI 기술로 새로운 가치를 창출하는 인재를 선발하겠다고 밝혔다.³³⁾ 잡코리아 분석에 따르면 2026년 1분기 기준 AI 키워드를 포함한 공고는 5년 전 대비 112% 늘었고 신입 대상 AI 공고는 162% 증가했다.

32) Stanford Digital Economy Lab, "No Widespread Displacement, but the AI Employment Gap for Young Workers Has Widened to 19%", Aug. 12, 2026.

33) 카카오, "카카오, 창사 첫 그룹 통합 신입 공개채용 실시", 2025.9.3.
<https://www.kakaocorp.com/page/detail/11688>

디지털서비스 이슈리포트

조직은 AI 도입 효과를 품질과 운영 성과까지 포함해 평가해야 한다. 자동화된 테스트와 명시적인 완료 기준을 도구 도입과 함께 마련하고, 변경 완료 시간, 재작업, 운영 중 결함을 지속적으로 관찰해야 한다. 초기에는 범위가 작고 결과를 확인하기 쉬운 작업에서 시작해 측정 결과에 따라 위임 범위를 확대하는 방식이 적합하다.

개인은 구현 능력에 문제 정의와 검증 역량을 결합해야 한다. 업무 담당자와 요구사항을 구체화하고, AI에 맡길 작업과 직접 판단할 부분을 나누며, 결과의 적절성을 설명할 수 있어야 한다. 초급 개발자에게도 이러한 경험을 제공하되 숙련자의 검토와 피드백을 통해 책임 범위를 점진적으로 넓혀야 한다.

이제 교육의 과제는 기초 역량, AI 활용, 업무 이해를 함께 발전시키는 것이다. 이를 소프트웨어 교육에 적용하면 실제 사용자의 문제를 조사하고 명세, 구현, 검증을 반복하는 과제가 적합하다. 학습자는 생성한 코드를 설명하고 오류의 원인과 수정 근거를 제시할 수 있어야 한다.

앤드루 응은 소규모 AI 활용 팀에서 제너럴리스트가 유리해지는 이유를 이렇게 설명했다. “두 사람으로 구성된 팀이 다섯 가지 전문성이 필요한 일을 해내야 한다면, 그중 누군가는 자기 전문 영역 밖의 역할을 맡아야 한다.” 그는 일부 팀에서 엔지니어와 제품 관리자의 비율이 8대 1에서 1대 1까지 내려가고 있다고 관찰한다.³⁴⁾ 구현 비용이 떨어질수록 차별화는 무엇을 만들지 결정하는 능력과 만든 것이 맞는지 판정하는 능력에서 나온다. 소프트웨어 엔지니어가 가야 할 방향도 여기에 있다.

34) Andrew Ng, “AI-Native Software Development Needs Generalists”, The Batch, Apr. 2026.

05 2026년 클라우드 솔루션 리포트: 추론 인프라 (1) 토큰 경제

| 정채상 이음테크

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.

한국지능정보사회진흥원이 저작권을 보유하고 있으며 승인 없이 이슈리포트의 내용 일부 또는 전부를 다른 목적으로 이용할 수 없습니다.

들어가며: 청구서가 먼저 온다

에이전트를 프로덕션에 올린 조직이 그다음 달에 가장 먼저 받아 드는 것은 거버넌스 대시보드가 아니라 청구서다. 지난 편에서 에이전트(스스로 계획을 세워 여러 단계를 실행하는 AI 프로그램)를 만드는 일은 쉬워졌고 남은 문제는 통제라고 정리했는데, 그 통제 문제가 실무에서 가장 먼저 비용의 형태로 드러난다. 에이전트는 한 번의 결과를 내놓기 위해 계획을 세우고, 도구를 호출하고, 돌아온 결과를 다시 읽고, 필요하면 처음으로 돌아가 다시 시도하는데, 이 과정이 쌓이면서 같은 일을 단발성 질의로 처리할 때보다 10배에서 30배에 이르는 토큰을 소비한다고 업계는 본다. 파일럿 시절에는 반올림 값이나 오차에 가까웠던 이 비용이 길지 않은 시간 후에 예산에서 별도 항목으로 관리해야 하는 상시 운영비가 되었다.

이 변화를 업계의 언어로 가장 선명하게 정리한 쪽은 공교롭게도 칩을 파는 회사였다. 엔비디아의 젠슨 황은 2026년 3월 GTC 2026 기조연설에서 데이터센터를 'AI 팩토리'로 재정의했는데, 투입되는 원료는 전기와 데이터이고 공장에서 나오는 완제품은 토큰이며, 생산성은 와트당 토큰으로 잴다는 것이다³⁵⁾. 칩 회사가 스스로를 토큰 생산 인프라 사업자로 설명하기 시작했다는 것은 토큰이 규격화된 원자재로 취급되기 시작했다는 뜻이고, 구매자의 언어로 옮기면 추론은 이제 기술 선택의 문제이기 이전에 조달의 문제가 되었다는 것이다.

가트너는 2026년 8월 전망에서 AI에 최적화된 인프라 서비스(IaaS) 지출이 2026년 423억 달러로 전년 대비 96% 늘고, 그 가운데 추론이 233억 달러로 학습의 190억 달러를 처음 넘어서 전체의

35) Data Center Frontier. "Jensen Huang Maps the AI Factory Era at NVIDIA GTC 2026." 2026.03.

디지털서비스 이슈리포트

55%를 차지할 것으로 내다봤다³⁶⁾. 기업의 예산 편성에서도 같은 흐름이 확인되는데, a16z는 기업 CIO 100명을 조사해 기업 한 곳의 평균 대규모 언어 모델 예산이 2025년 700만 달러에서 2026년 말 1,160만 달러로 늘어날 것으로 전망했다³⁷⁾.

이번 리포트는 그 토큰을 '사는 쪽'의 이야기다. 추론 산업은 세 개의 층으로 쌓여 있고, 이 글은 CIO와 CTO가 실제로 지갑을 여는 맨 위층을 다루며, 공장을 '짓는 쪽'의 사정은 다음 편에서 이어진다. 이번 보고서에서의 질문은 다음의 셋이다. 1) 어디서 살 것인가, 2) 어떤 방식으로 살 것인가, 그리고 3) 얼마에 살 것인가? 그런데 2026년의 조직은 이 셋을 한 번 정해 두고 끝내지 못하는데, 애플리케이션이나 에이전트가 모델을 호출할 때마다 셋을 다시 판단하는 일을 사람 대신 소프트웨어가 맡기 시작했기 때문이다. 이 질문들에 답할 수 있는 상태, 곧 무엇을 얼마나 쓰는지 알고 필요하면 알아낼 수 있는 상태를 이 글은 구매 역량이라 부르고, 그 역량을 어떻게 갖출 것인가로 글을 정리한다.

토큰은 어디서 와서 어떻게 팔리는가?

토큰이 오는 곳, 세 개의 층



그림 12 추론 산업의 세 개 층과 그 위의 구매 데스크

36) Gartner. "Gartner Forecasts Worldwide AI-Optimized IaaS Spending to Grow 96% in 2026." 2026.08.10.

37) a16z. "How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025."

디지털서비스 이슈리포트

맨 아래층은 AI 데이터센터로, 부지와 전력을 갖춘 사업자와 그 위에 GPU를 놓고 연산 시간을 파는 사업자를 합쳐 부르는 이름이며 공장에 비유하자면 건물과 설비다. 그 위의 추론 인프라는 같은 GPU에서 더 많은 토큰을 뽑아내는 서버 엔진과 최적화의 층이고 생산 라인에 해당한다. 맨 위의 토큰 팩토리는 완제품인 토큰을 API로 판매하는 층이며, 엔비디아가 AI 팩토리라 부르는 것이 세 층을 합친 전체라면 이 글의 토큰 팩토리는 그 맨 위층만을 가리킨다. 세 층은 회사의 경계라기보다 기능의 경계여서 한 회사가 두 층을 품기도 하고, 구매자가 GPU를 직접 두고 오픈 모델을 돌리는 셀프호스트 방법으로 가운데 층을 스스로 떠안기도 한다. 그리고 이 세 층으로 쌓인 구조의 맨 위에는 모델을 호출할 때마다 어느 공급자에게서 살지를 고르는 구매 데스크가 새로 얹어진다.

이 세 층은 가동률을 기준으로 한 지점에서 만난다. 데이터센터의 자본 지출은 그 자체로는 비용일 뿐이고 추론 인프라가 GPU를 쉬지 않고 돌려 토큰으로 바꿔낼 때에만 매출이 되는데, 그래서 이 산업의 마진은 GPU 시간을 빌려주는 아래층보다 토큰을 파는 위층에서 커질 수 있다. 하지만, 그것은 가동률을 채웠을 때에만 실현되는 잠재적 마진이고 커모디티화(어디서 사든 같은 물건이 되어 가격으로만 경쟁하게 되는 것)의 압력이 그 마진을 계속 깎아내린다. 구매자에게 이는 선택지로 여겨지는데, 토큰을 사면 GPU가 쉬든 말든 신경 쓸 일이 없는 대신 남의 마진을 함께 지불하고, GPU 시간을 사면 마진을 줄이는 대신 가동률을 스스로 책임져야 한다.

가동률이 정하는 세 가지 구매 모드: 택시, 대절 버스, 자가용

실무에서 추론을 사는 방식은 크게 셋이며, 탄 만큼 내는 택시(서버리스 API)와 시간당 빌리는 대절 버스(전용 엔드포인트), 사서 직접 모는 자가용(셀프호스트)에 해당한다. 셋을 가르는 것은 무엇을 남에게 맡기고 무엇을 직접 지느냐인데, 다만 택시에도 분당 토큰 한도라는 사실상의 용량 계약이 붙는다.

구분	서버리스 API	전용 엔드포인트	셀프호스트
비유	탄 만큼 내는 택시	시간당 빌리는 대절 버스	사서 직접 모는 자가용
과금 단위	토큰당	GPU 시간당	GPU 구매·임대 + 운영 인건비
지연	가변	안정	튜닝에 따라 최적화 가능
처리량	유연	보장	자체 용량 한도 내 자유
모델 선택	제공 목록 한정	제공 목록 + 커스텀 모델 반입	오픈 모델 전면 자유
데이터 주권	낮음~중간	중간	완전
관리 부담	최소	낮음	높음

표 3 세 가지 구매 모드 비교(칸의 값은 대표적인 경우이며 사업자와 리전에 따라 다르다)

디지털서비스 이슈리포트

세 모드 사이의 경계선을 긋는 것이 가동률(확보한 처리량 대비 실제로 쓴 처리량)이다. 서버리스와 전용 엔드포인트의 손익분기점은 대체로 GPU 가동률 40~50% 부근에 놓인다는 분석이 있고³⁸⁾, 셀프호스트는 GPU 임대 원가만 놓고 보면 응답 속도 요구에 따라 22~48% 구간에서 손익이 갈리지만³⁹⁾ 운영 비용을 얹으면 그 경계는 위로 올라간다. 같은 분석은 거의 놓고 있는 셀프호스트 GPU가 서버리스보다 2~4배 비쌀 수 있다고 계산하는데, 그 차이는 인건비와 피크에 맞춰 확보해 두는 유휴 용량에서 온다. 사실상의 표준이 된 vLLM과 프리픽스 재사용 구조(RadixAttention)로 이름을 알린 SGLang 같은 오픈소스 서빙 엔진이 진입 문턱을 낮추기는 했지만, 양자화 수준을 정하고 병렬 구성과 문맥 캐시(KV 캐시) 메모리를 조정하고 장애를 복구하는 일에는 여전히 전담 인력이 필요하다.

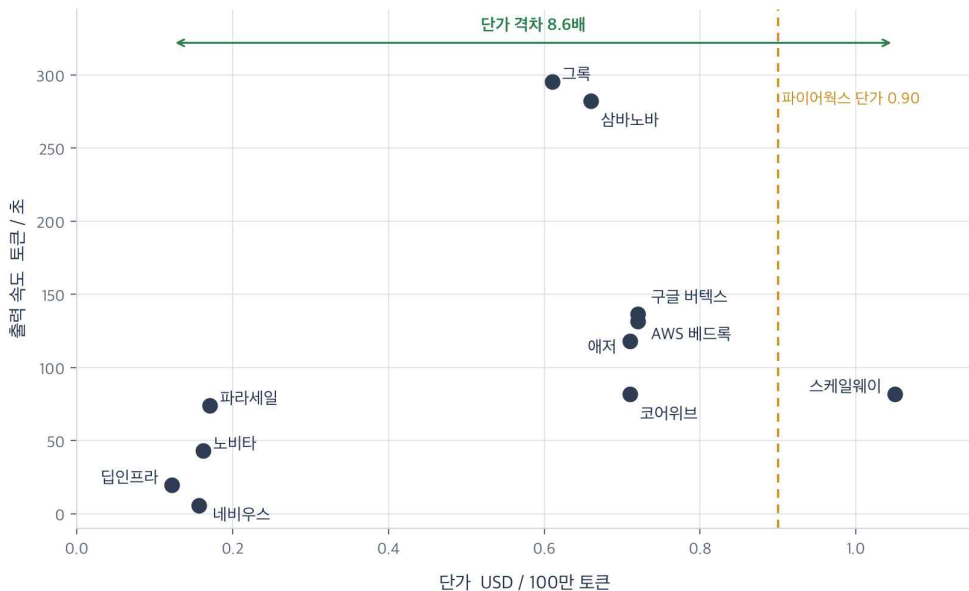


그림 13 동일 오픈 모델의 사업자별 단가와 생성 속도 분포.

Llama 3.3 70B, 단가는 캐시·입력·출력 7:2:1 혼합 기준, 파이어웍스는 속도 미측정
(출처: Artificial Analysis, 2026년 9월 2일 조회)

이 경계는 어느 서버리스 단가와 견주느냐에 따라서도 크게 달라진다. 같은 오픈 모델을 여러 사업자가 동시에 서빙하는 구조가 자리 잡으면서, 널리 쓰이는 한 오픈 모델을 기준으로 최저가 사업자와 최고가 사업자의 100만 토큰당 단가는 8배 넘게 차이가 난다. 초당 생성 속도 역시 사업자마다 크게 다른데,

38) GMI Cloud. "Serverless vs Dedicated Inference: Which Setup Is Better for Production LLM Workloads?"

39) DigitalOcean. "Serverless vs Dedicated vs Self-Hosted LLM Inference Cost."

디지털서비스 이슈리포트

단가와 속도가 비례하지 않아서 가장 빠른 사업자가 가장 비싼 사업자는 아니다⁴⁰⁾. 단가표의 한 줄도 들여다보면 입력과 출력, 캐시 적중, 긴 문맥 구간에 따라 요금이 나뉘고, 사고 단계를 거치는 추론(리즈닝) 모델은 화면에 보이지 않는 사고 토큰까지 출력 요금으로 청구한다. 구매자가 고를 여지가 그만큼 넓다는 사실이 구매 데스크가 존재하는 이유다.

구매 데스크: 호출마다 공급자를 고르는 총

같은 상품의 가격이 8배 넘게 벌어져 있다면, 모든 호출을 최상위 모델로 보내는 것은 모든 출장을 일등석으로 보내는 일과 다르지 않다. 그래서 애플리케이션과 모델 사이에 새로운 층이 끼어들었는데, 호출이 일어날 때마다 이번 토큰은 어느 공장에서 살지를 자동으로 결정하는 계층이며 앞의 층 위에 놓인 구매 데스크라 부를 만하다.

이런 자동 선택이 편의 못지않게 위험도 가져온다는 사실은 2025년 8월 오픈AI의 GPT-5 출시가 보여줬다. 오픈AI는 챗GPT에서 쉬운 질의는 소형 모델로, 어려운 질의는 대형 모델로 보내는 실시간 라우터를 기본값으로 켜는데(API에서는 사용자가 세 가지 크기의 모델을 직접 고른다⁴¹⁾), 초기 라우팅이 제대로 작동하지 않으면서 복잡한 질의까지 소형 모델로 흘러갔고 새 모델이 오히려 퇴보했다는 반발이 일자 이전 모델의 선택지를 되살리는 것으로 대응했다⁴²⁾. 소비자 제품에서 먼저 벌어진 이 일은 관리형 라우팅을 사려는 기업에서도 그대로 되풀이될 수 있는데, 내 질의를 무엇이 처리했는지 모르는 상태에서는 품질 저하를 진단할 수도 책임을 물을 수도 없다.

이 편의와 위험을 함께 다루는 상품은 세 갈래로 나뉜다. 품질과 비용 사이의 다이얼을 사용자에게 넘기는 독립 게이트웨이, 라우팅을 기본 기능으로 흡수한 하이퍼스케일러, 그리고 GPT-5처럼 라우팅을 제품 안에 넣어 소비자가 모델 대신 결과의 품질 등급을 사게 하는 모델 벤더다. 앞의 두 갈래는 다음 장의 오픈라우터 절과 AWS 베드록·마이크로소프트 파운드리 절에서 본다.

시장이 라우팅을 상품으로 팔아 온 동안, 학계는 같은 문제를 공개된 연구 과제로 다뤄 왔다. 값싼 모델부터 시도하는 캐스케이드 방식을 처음 정식화한 프루갈GPT(FrugalGPT, 2023)에서 2026년의 LLM라우터(LLMRouter)까지, 개별 기법이 표준 벤치마크를 거쳐 통합 라이브러리로 수렴했다. 학습된 라우터가 성능 지표에서 최강의 단일 고정 모델보다 상대적으로 14.6% 나은 결과를 냈다는

40) Artificial Analysis. "Llama 3.3 Instruct 70B: API Provider Benchmarking & Analysis."

41) OpenAI. "Introducing GPT-5 for developers." 2025.08.07.

42) Fortune. "OpenAI's GPT-5 Model Router Backlash." 2025.08.12.

디지털서비스 이슈리포트

LLM라우터의 보고⁴³⁾는, 비용과는 별개로 모델마다 잘하는 질의가 달라 조합이 단일 최강 모델을 이길 수 있다는 뜻이다. 라우팅은 특정 업체의 비밀로 남지 않는 표준화 가능한 학습 문제이며, 자사의 질의 분포와 그 위의 품질 판정 데이터로 직접 학습시켜 내재화할 수 있는 역량이기도 하다.

라우팅 기술이 그렇게 범용화되는 동안 구매 데스크라는 자리는 오히려 비싸졌다. 팔로알토 네트워크는 게이트웨이 스타트업 포트키(Portkey)를 인수해 자사 보안 플랫폼의 통제점으로 삼았고⁴⁴⁾, 스트라이프는 2026년 8월 오픈라우터 인수에 합의했는데 보도된 가격은 70억 달러 이상으로 석 달 전 기업가치의 5배가 넘는다⁴⁵⁾. 라우팅의 가치를 다른 곳에서 지불하고 있는 셈이다.

여기서 한 가지 질문이 남는데, 스위치 하나로 수요가 공장 사이를 즉시 옮겨 다닐 수 있고 그 스위치를 만드는 기술마저 오픈소스로 풀려 있다면, 토큰을 만들어 파는 쪽에는 무엇이 남는가? 다음 장의 여섯 벤더는 각자의 방식으로 이 질문에 답하는 중이다.

주요 벤더별 솔루션 심층 분석

이 영역의 주요 플레이어인 여섯 벤더를 가치사슬의 위치 순서로 살펴본다. 독립 추론 클라우드, 전용 실리콘 진영, 구매 데스크, 하이퍼스케일러의 순이며, 각 벤더가 커모디티화의 압력 앞에서 무엇을 해자로 내세우는지에 초점을 맞춘다.

파이어웍스 AI: 속도에 값을 매기다

독립 추론 클라우드인 파이어웍스 AI는 속도를 해자로 삼는 전략의 가장 성공한 사례다. 회사에 따르면 2026년 7월 연간 반복 매출이 10억 달러를 넘겼고 하루에 처리하는 토큰은 40조 개를 넘는다⁴⁶⁾. 커서·노션·쿼라·버셀 같은 고객 명단이 파이어웍스의 위치를 잘 보여 주는데, 코드 편집기와 문서 도구처럼 사용자가 응답을 기다리는 시간이 곧 제품 품질인 회사들이다.

기술의 중심에는 자체 개발한 어텐션 커널 파이어어텐션(FireAttention)이 있고, 그 위에 같은 GPU에서

43) Jiaxuan You et al. "LLMRouter." arXiv:2608.06867, 2026.

44) Palo Alto Networks. "Palo Alto Networks Completes Acquisition of Portkey to Secure AI Agents." 2026.05.29.

45) Stripe. "Stripe agrees to acquire OpenRouter to help businesses optimize token routing and usage." Stripe Newsroom, 2026.08.19.

46) Fireworks AI. "Announcing Our Series D." 2026.07.

디지털서비스 이슈리포트

토큰 산출을 높이는 최적화 기법들이 쌓여 있다. 회사는 이 커널의 FP8 구현이 오픈소스 엔진 vLLM보다 4배 빠르다고 밝혀 왔고⁴⁷⁾, 고객인 노션은 응답 지연을 약 2초에서 350밀리초로 줄였다고 말한다⁴⁸⁾. 서버리스 외에 전용 엔드포인트와 온프레미스 배포까지 세 가지 구매 모드를 모두 갖췄고, 2026년에는 마이크로소프트 파운드리 안에서 정식 서비스로 제공되기 시작해 독립 추론 클라우드가 하이퍼스케일러의 계약과 청구서 안으로 들어간 드문 사례가 됐다.

속도에는 값이 붙는다. 파이어웍스의 서버리스 단가는 16B 초과 모델 기준 100만 토큰당 0.90달러로, 최저가 사업자의 0.12달러와 견주면 7.5배에 이르러⁴⁹⁾, 밤새 돌리는 배치 요약처럼 속도가 필요 없는 워크로드에는 과잉이고, 양자화 수준과 캐싱 옵션이 많아 최적의 조합을 찾는 데에도 학습 비용이 든다. 빠른 토큰은 아직 원자재가 되지 않았다는 것이 파이어웍스의 주장이고, 응답 속도가 곧 제품 품질인 워크로드에서만 그 값을 치를 이유가 있다.

투게더 AI: 백화점이 공장을 갖기 시작할 때

투게더 AI(Together AI)는 오픈소스 모델 생태계의 백화점을 자처한다. 폐쇄형 API보다 크게 저렴하다는 것을 정면에 내세우고, 수백 개의 오픈 모델을 서빙하는 카탈로그 위에 파인튜닝 서비스와 전용 GPU 클러스터까지 있는 스펙트럼을 갖췄으며, 회사는 2026년 7월 기준 수주 잔고가 11억 5,000만 달러이고⁵⁰⁾, 8월에는 월간 처리 토큰이 400조 개에 이른다고 밝혔다⁵¹⁾.

2026년 7월의 투자 유치 발표에서 눈여겨볼 숫자는 따로 있는데, 투게더가 확보했다고 밝힌 500메가와트(업계는 확보한 계산 능력을 전력 용량으로 센다)의 커밋 컴퓨트다⁵²⁾. 위층의 마진은 가동률로만 지켜지므로 투게더는 가동률의 근원인 공급을 잡으려 가치사슬 아래로 내려간 것이고, 2026년 8월에는 IBM 클라우드에 추론 전용 GPU 클러스터를 두는 2억 4,000만 달러 규모의 다년 계약을 맺어 남의 공장까지 빌린 셈이다. 투게더가 내세우는 강점은 규모의 경제와 파인튜닝인데, 플레이그라운드에서 모델을 비교하다 같은 화면에서 파인튜닝 작업을 시작하는 흐름이 사용자 경험의 강점이다.

47) Fireworks AI. "FireAttention: Serving Open Source Models 4x faster than vLLM by quantizing with ~no tradeoffs." 2024.01.08.

48) Fireworks AI. 홈페이지 고객 사례.

49) Fireworks AI. "Serverless Pricing." Fireworks Docs.

50) TechCrunch. "Neocloud Together AI Raises \$800M, Leaps to \$8.3B Valuation." 2026.07.01.

51) IBM. "IBM and Together AI Sign Multi-Year Agreement to Scale Open-Source AI Inference with NVIDIA AI Infrastructure on IBM Cloud." IBM Newsroom, 2026.08.11.

52) Together AI. "Announcing our \$800M Series C to accelerate the shift to open-source AI." 2026.07.01.

디지털서비스 이슈리포트

반면 위와 아래에서 동시에 압력을 겪는데, 위로는 폐쇄형 최상위 모델이 없으므로 오픈 모델로 충분한지를 검증하는 일이 고객의 몫이고, 아래로는 자체 GPU를 직접 운영해 단가로 경쟁하는 사업자들이 있어 오픈 모델끼리 견주면 투게더는 오히려 비싼 측에 속한다. 6~20배 저렴하다는 회사의 주장은 폐쇄형 모델과 견줄 때의 이야기다.

그록과 세레브라스: GPU에서는 나오지 않는 속도

GPU 대신 전용 실리콘으로 토큰을 만드는 이 두 회사가 내세우는 것은 클러스터 전체의 처리량보다 요청 하나가 체감하는 속도, 곧 첫 토큰이 나오기까지의 지연과 그 뒤의 생성 속도다. 토큰을 하나 만들 때마다 모델 가중치 전체를 메모리에서 읽어야 하므로 생성 단계는 연산보다 메모리 대역폭에 묶이는데, 두 회사는 가중치를 GPU의 외장 메모리(HBM) 대신 칩 위의 SRAM에 두어 이 병목을 피한다. 그록(Groq)의 LPU는 실행 순서가 미리 정해져 대기 시간이 없는 구조로 소형·중형 모델에 특화되어, 독립 벤치마크 기준 널리 쓰이는 70B급 오픈 모델을 초당 300토큰 이상으로 생성해 GPU 기반 사업자의 속도를 큰 차이로 앞선다. 실시간 에이전트나 음성 대화처럼 응답이 몇백 밀리초만 늦어져도 제품으로서 실패하는 워크로드가 그록의 영역이다.

세레브라스는 같은 원리를 훨씬 큰 칩에 적용한다. 웨이퍼 하나를 통째로 칩으로 쓰는 WSE-3는 GPU의 칩 내장 메모리보다 수백 배 큰 SRAM을 갖춰 405B급 초대형 모델까지 초당 900토큰 넘게 생성한다. 회사는 2026년 5월 나스닥에 상장해 55억 달러 안팎을 조달했고⁵³⁾, 8월에 발표한 2분기 실적에서 클라우드 부문 매출이 전년 동기 대비 4배 가까이 늘었다고 밝혔으며⁵⁴⁾, 오픈AI와는 750메가와트 규모, 200억 달러 이상의 다년 공급 계약을 맺었다⁵⁵⁾.

반면 두 회사는 같은 값을 치르는데, 가중치를 SRAM에 담아야 하므로 큰 모델일수록 칩 수가 급증하고 새 모델을 올릴 때마다 포팅 비용이 들어 모델 커버리지가 GPU 사업자보다 좁으며, 공급 능력이 수요를 따라가지 못하는 시기가 반복된다. 그러나 다이얼을 아무리 돌려도 GPU에서는 나오지 않는 속도가 있고, 그 속도의 값을 가장 크게 매긴 쪽은 GPU 회사 자신이었었는데, 엔비디아는 2025년 12월 그록 추론 기술의 비독점 라이선스를 보도에 따르면 약 200억 달러에 사들이고 창업자 조너선 로스를 포함한 핵심 인력을 데려 갔으며, 그록은 독립 회사로 남아 GroqCloud를 계속 운영하고 있다⁵⁶⁾.

53) Cerebras. "Cerebras Systems Announces Pricing of Initial Public Offering." 2026.05.

54) Cerebras. "Cerebras Systems Fast Inference Cloud Business Nearly Quadruples in Second Quarter 2026." 2026.08.12..

55) Cerebras. "Cerebras Systems Announces Strong First Quarter 2026 Results." 2026.06.23.

디지털서비스 이슈리포트

오픈라우터: 결제 회사가 산 구매 데스크

오픈라우터는 여섯 벤더 가운데 유일하게 GPU가 없다. 400개가 넘는 모델과 80곳이 넘는 공급자를 오픈AI 호환 단일 API 뒤에 모아 놓고, 공급자의 토큰 단가는 그대로 넘기되 크레딧을 결제할 때 5.5%의 수수료를 얻는다⁵⁷⁾. 앞에서 언급한 공장이 아니라 대표적인 구매 데스크이며, 2026년 5월 기준 사용자는 800만 명, 처리량은 월 100조 토큰으로 6개월 만에 5배가 됐다⁵⁸⁾. 그리고 2026년 8월, 결제 회사 스트라이프가 이 구매 데스크를 사기로 합의했다.

핵심 기능은 폭넓은 모델 선택지, 한 공급자가 장애를 일으키면 자동으로 다른 공급자로 넘기는 폴백, 그리고 품질과 비용 사이의 다이얼을 사용자에게 넘기는 자동 라우터다. 구매 데스크에 서면 시장 전체가 보인다는 점도 오픈라우터만의 자산인데, 회사가 공개하는 사용 통계에 따르면 프로그래밍 워크로드가 2025년 초 전체 토큰의 11%에서 연말에는 절반 이상으로 늘었고⁵⁹⁾, 2026년 2월에는 중국 모델이 주간 토큰 처리량에서 미국 모델을 처음 앞질러 6월 기준 라우팅된 토큰의 46%를 차지했다⁶⁰⁾.

반면 구매 데스크라는 위치의 약점은 그대로다. 사용량이 커진 고객은 볼륨 할인과 약정 요금을 얻기 위해 공급자와 직접 계약하려 하고, 라우팅 기술은 오픈소스로 풀려 있어 자체 구축이 가능하며, 감사 로그와 예산 통제 같은 엔터프라이즈 거버넌스 기능은 아직 성숙 중이다. 그렇다면 스트라이프는 무엇을 산 것인가. 오픈라우터의 수수료가 결제할 때 붙는다는 사실이 답의 절반이고, 스트라이프의 패트릭 콜리슨이 인수 발표에서 "토큰은 AI로 제품을 만드는 회사들의 중심 통화"라고 말한 것이 나머지 절반이라 할 수 있다. 오픈라우터의 창업자는 자사를 'AI의 스트라이프'라 불렀는데, 스트라이프가 산 것은 라우팅 기술이라기보다 800만 사용자가 토큰을 결제하는 계산대이고, 구매 데스크의 자리가 그 자리에 서 있는 기술보다 비싸다는 것이 이 거래의 뜻이다.

AWS 베드록: 이미 맺어 둔 계약이라는 자산

이전 보고서에서 베드록 에이전트코어를 에이전트 실행 인프라로 다뤘다면, 이번에 보는 것은 같은 베드록의 아래층, 곧 토큰을 파는 층이다. AWS의 관리형 추론 서비스인 베드록은 수백 개의

56) Groq. "Groq and Nvidia Enter Non-Exclusive Inference Technology Licensing Agreement to Accelerate AI Inference at Global Scale." Groq Newsroom, 2025.12.

57) OpenRouter. "FAQ." OpenRouter Docs.

58) TechCrunch. "OpenRouter More Than Doubles Valuation to \$1.3B in a Year." 2026.05.26.

59) OpenRouter. "State of AI."

60) Yahoo Finance. "Chinese AI Models Now Capture Up to 46% of US Enterprise Token Usage." 2026.

디지털서비스 이슈리포트

파운데이션 모델을 단일 API로 제공하며, 아마존에 따르면 12만 5,000곳이 넘는 고객과 포춘 100대 기업의 약 80%가 이용한다⁶¹⁾. 이미 체결된 AWS 계약과 IAM 권한, VPC 경계, 규정 준수 인증, 곧 심사가 끝난 보안 울타리 안에서 토큰을 산다는 것은 조달 부서와 보안 부서에 새 벤더를 심사하는 절차를 통째로 생략해 준다는 뜻이기도 하다.

구매 모드의 관점에서 베드록은 세 요금제를 한 제품 안에서 다룬다. 토큰당 과금하는 온디맨드, 지연을 감수하는 대신 업계 관행대로 절반 가격인 배치 추론, 처리량을 약정하고 그 용량을 채워 쓰면 30~50%를 절감하는 프로비저닝 처리량이 그것이며, 앞의 가동률 논의를 벤더를 바꾸지 않고 요금제만 바꿔 따라갈 수 있게 한 것이다. 라우팅도 안으로 흡수했는데, 2025년 4월 정식 출시된 지능형 프롬프트 라우팅은 요청별로 응답 품질을 예측해 같은 모델 계열 안에서 크기가 다른 모델을 오가며, AWS에 따르면 내부 테스트에서 계열에 따라 16~56%의 비용 절감을 보였다⁶²⁾.

Router settings

Models for routing
 Choose two base models from the same provider for your router.

 **Nova Pro 1.0** ⓘ ✕
APAC Nova Pro

 **Nova Lite 1.0** ⓘ ✕
APAC Nova Lite

Fallback model
 Choose the default model for requests that do not meet the routing criteria. The response quality difference is measured based on the fallback model.

Fallback model
 Nova Pro ▼

Routing criteria
 Define the thresholds to use for routing requests between models.

Response quality difference (%)
The quality difference between the fallback model and other model responses for each prompt, as predicted by a model. Lower values indicate similar responses from other models to the fallback model.



그림 14 AWS 베드록 콘솔의 지능형 프롬프트 라우팅 설정 (출처: AWS)

반면 심사가 끝난 울타리 안에 머무는 대가는 선택의 제한이다. 새 모델이 시장에 나온 뒤 베드록 카탈로그에 오르기까지 시차가 있고, 라우팅은 같은 계열 안에서만 작동하므로 계열을 넘는 선택은 여전히 사람의 몫이며, 모델별 단가가 AWS 정가로 고정되어 있어 사업자 간 8배가 넘는 단가 격차는

61) Amazon. "Andy Jassy weighs in on the rapid growth of Amazon's chips business." About Amazon, 2026

62) AWS. "Amazon Bedrock Intelligent Prompt Routing Is Now Generally Available." 2025.04.22.

디지털서비스 이슈리포트

이 울타리 안으로 들어오지 않는다. 베드록을 붙드는 것은 토큰의 값보다 이미 맺어 둔 계약이며, 그것은 구매자 입장에서 가장 경계해야 할 종류의 편의이기도 하다.

마이크로소프트 파운드리: 유통망이 된 애저

마이크로소프트 파운드리는 2026년 1월 애저 AI 파운드리에서 이름을 바꾸며 애저의 한 서비스에서 마이크로소프트의 핵심 제품으로 격상됐다⁶³⁾. 이전 보고서에서 에이전트 플랫폼으로 다룬 그 파운드리의 추론 충전데, 오픈AI 모델은 물론 엔트로픽의 클로드, 딥시크, xAI, 메타, 미스트랄까지 한 카탈로그에서 고를 수 있다⁶⁴⁾.

전략적으로 눈에 띄는 것은 유통 구조의 재편이다. 파이어웍스가 파운드리에 편입된 것은 파이어웍스에게는 새 유통 경로이고, 파운드리에게는 딥시크나 큐원 같은 오픈 모델을 자체 서빙 인프라 없이 카탈로그에 올리는 방법이 되었다. 라우팅도 갖추었는데, 2026년 기준으로 모델 라우터는 하나의 배포 뒤에 오픈AI·엔트로픽·딥시크·메타·xAI의 모델을 두고 운영자가 균형·비용·품질 가운데 한 전략을 고르게 하므로, 같은 계열 안에서만 오가는 베드록보다 선택 범위가 넓다. 다만 라우터 자체에 입력 토큰 기준의 요금이 따로 붙는다⁶⁵⁾.

반면 사용자 측면에서 기존 애저 오픈AI 엔드포인트와 파운드리 엔드포인트가 병존해 어느 쪽을 써야 하는지가 여전히 혼란스럽고, 모델별·배포 방식별로 갈리는 과금 체계도 복잡한데, 2026년 8월에는 일부 오픈 모델의 토큰당 과금을 중단하고 프로비저닝 처리량으로만 남기기도 했다⁶⁶⁾. 베드록이 AWS에 정착한 조직에게 주는 편의를 파운드리는 애저에 정착한 조직에게 주지만, 두 클라우드를 함께 쓰는 조직에게는 관리해야 할 제품군이 하나 더 늘어나는 일이다. 그럼에도 파운드리를 붙드는 힘은 에이전트 플랫폼과 토큰 상점이 한 콘솔에 있다는 데서 나온다.

맺으며: CIO와 CTO를 위한 제언

토큰 경제에서 구매자가 가져야 할 것은 특정 벤더가 아니라 구매 역량이다. 어느 벤더도 조직이 스스로 무엇을 얼마나 쓰는지 모르는 상태를 대신 해결해 주지는 않기에 다음의 네 가지를 권한다.

1. 무엇을 얼마나 쓰는지부터 계측하라

63) SAMexpert. "Microsoft Product Terms Update: January 2026."

64) Microsoft. "Foundry Models." Microsoft Azure.

65) Microsoft Learn. "Model router for Microsoft Foundry concepts." 2026.08.12.

66) Microsoft Learn. "Fireworks models on Microsoft Foundry."

디지털서비스 이슈리포트

AI 인프라 지출의 절반 이상이 추론인데 모델별, 팀별, 워크로드별 토큰 지출을 조직적인 차원에서 계측하기를 권한다. 어느 벤더가 싸가는 두 번째 질문이고, 첫 번째 질문은 우리가 어떤 질의를 얼마나 보내며 그중 최상위 모델이 꼭 필요한 비중이 얼마인가다. 첫 행동은 조직 안에서 발급된 LLM API 키가 몇 개이고 각각 어느 팀의 어떤 용도인지 목록을 만드는 일이다. 키가 팀과 용도별로 나뉘어 있지 않으면 어떤 대시보드도 팀별 지출을 보여 주지 못한다. 여기에 반복적인 에이전트에서 쓰는 토큰의 대부분은 같은 문맥을 반복해 읽는 입력 토큰이므로 캐시 적중 토큰을 정가의 10분의 1 안팎으로 받는 프롬프트 캐싱의 적중률이 단가표보다 큰 비용 변수로 작용한다.

한국의 구매자에게는 따로 셀 항목들이 더 있다. 첫째, 같은 내용을 한국어로 보내면 영어보다 토큰이 더 들고 그 배율이 모델의 토큰나이저마다 다르므로, 단가표에서 쓴 모델이 한국어 워크로드에서는 비싼 모델이 될 수 있다. 둘째, 단가가 달러로 매겨지니 원화 예산은 환율만큼 흔들린다. 셋째, 새 모델이 국내 리전에 등재되기까지의 시차와 국외 리전을 쓸 때의 데이터 반출 판단은 계약 전에 점검할 항목이며, AWS가 처리 지역을 전 세계로 열어 두는 요청에 특정 권역 안으로 묶는 요청보다 약 10% 낮은 단가를 매기듯⁶⁷⁾ 지역을 고정하는 선택에는 추가적인 비용을 고려해야 한다.

이 계측은 민간 기업만의 일이 아니다. 공공 부문에서는 토큰이 소모품인지 서비스 구독인지에 따라 예산 편성과 계약 방식이 달라지는데, 토큰당 과금은 소모품 예산의 논리에, 프로비저닝 처리량은 서비스 구독의 논리에 가깝고, 과업지시서를 어느 쪽으로 쓰느냐에 따라 입찰 단위와 정산 방식이 갈린다. 계측 없이는 어느 논리가 자기 워크로드에 맞는지도 알 수 없다.

2. 워크로드마다 구매 모드를 따로 정하라

자릿수로 보면 GPU 가동률 40~50% 부근이 서버리스와 전용의 경계이고, 셀프호스트는 GPU 원가에 인건비와 여유 용량을 얹어 계산해야 한다. 예측 가능한 대규모 워크로드는 아래층으로 내려보내고 탐색적 워크로드는 서버리스에 두되, 이 결정은 워크로드 단위로 내려야 한다. 첫 행동은 토큰 지출이 가장 큰 워크로드 하나를 골라 시간대별 호출량 그래프를 그리고 평균 대비 피크의 비율을 가늠해 보는 일이 좋고, 이 비율을 이용해서 우리 조직의 손익분기점을 숫자로 옮겨 보는 것이 필요하다. 그래프가 평평하면 약정 요금제나 전용 엔드포인트 견적을 한 건 받아 보는 것이 적절한 방법이고, 성숙한 조직은 세 모드를 동시에 이용한다.

67) AWS. "Global cross-Region inference." Amazon Bedrock User Guide.

디지털서비스 이슈리포트

3. 계측하고 평가한 다음에 라우팅하라

자사 워크로드 기준의 평가 체계 없이 라우팅부터 사면 품질이 떨어져도 무엇 때문인지 모르는 GPT-5 반발의 기업판이 된다. 계측을 통해 무엇을 쓰는지 알고, 자사 질의로 모델을 평가해 어느 등급이 어느 태스크에 충분한지 알아낸 다음에야 라우팅 정책이 의미를 갖는다. 첫 행동은 자사에서 실제로 쓰는 질의들을 뽑아 상위 모델과 하위 모델의 답을 나란히 놓고 담당자가 채점하는 평가 세트를 만드는 일이다. 모델 선택은 일회성 조달에서 계속 다시 정해야 하는 운영 변수가 되었고, 이 평가 세트의 관리는 최고 책임자들이 계속 챙겨야 할 일이다.

4. 벤더를 갈아탈 수 있는 상태를 유지하라

벤더를 갈아탈 수 있다는 사실 자체가 협상력이다. 오픈AI 호환 API 덕분에 엔드포인트 주소만 바꾸면 옮겨 갈 수 있다는 벤더들의 말은 들어오기 쉽다는 뜻인 동시에 나가기도 쉽다는 뜻이고, 오픈 모델의 자체 검증과 구매 데스크는 모두 전환 비용을 낮추는 도구다. 그리고 스트라이프의 오픈라우터 인수가 보여 주듯 구매 데스크도 남의 것이 될 수 있으므로, 게이트웨이 자체를 오픈소스로 두는 선택지를 함께 고려해야 한다. 첫 행동은 주력 워크로드 하나를 다른 사업자의 엔드포인트로 바꿔 하루 동안 돌려 보는 등의 전환 리허설이며, 그 기록이 계약 갱신 협상에서 우리는 갈아탈 수 있다는 증거가 된다. 다만 공장들도 같은 사실을 알고 있기에 갈아탈 수 없는 이유를 만드는 것이 그들의 일이고, 그 공장의 사정은 다음 보고서에서 이어진다.

