

Deloitte Insights

August 2026



2028년 전망

딜로이트 기업 AI 인프라 조사

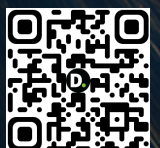
AI 네이티브 시대, 인프라 경쟁 본격화

Deloitte Global

Deloitte.

Download on the
App Store

GET IT ON
Google Play



'딜로이트 인사이트' 앱에서
경영·산업 트렌드를 만나보세요!

목차

리더메시지	03
개요	04
01 AI 운영 본격화... 토큰 수요도 폭증	05
02 토큰 사용량 급증, AI 확산의 신호이자 새로운 비용 리스크	08
03 AI 팩토리화 엮히, 2028년까지 두 배 확대	10
04 AI 인프라 투자 3배 급증 시대, 재무 관리 혁신이 필수	12
05 AI 스케일업의 병목은 기술이 아닌 인재	14
06 2028년 AI 대규모 확장 시대 경쟁력, 지금 결정된다	16
주석	17

리더메시지



김수연 파트너

엔지니어링, AI & 데이터 리더 |
한국 딜로이트 그룹

생성형 AI는 이제 개별 애플리케이션을 넘어 기업 운영 전반을 재설계하는 핵심 인프라로 진화하고 있습니다. 따라서 AI 경쟁의 중심도 더 이상 우수한 모델 확보에 그치지 않습니다. AI를 안정적으로 확장하고 지속적으로 운영할 수 있는 AI 팩토리, 하이브리드 인프라, 데이터 아키텍처, 전력과 반도체, 토큰 관리, 그리고 이를 뒷받침하는 거버넌스와 인재 확보가 새로운 경쟁력의 핵심 요소로 부상하고 있습니다.

이번 보고서는 미국 주요 산업의 비즈니스 및 기술 리더 515명을 대상으로 AI 인프라 투자와 운영 전략을 조사해, 기업들이 2028년까지 AI를 어떻게 대규모로 확장 도입할 것인지 분석했습니다. AI 애플리케이션의 확장 속도부터 AI 팩토리 구축, 하이브리드 인프라, 토큰 소비 증가, 예산 확대, 전력 확보, 인재 전략에 이르기까지 AI 시대를 뒷받침할 핵심 인프라의 변화 방향을 종합적으로 제시합니다.

딜로이트는 앞으로 3년이 AI 시대의 승자를 가르는 중요한 분기점이 될 것으로 전망합니다. AI를 단순한 기술 도입이 아닌 장기적인 운영 플랫폼으로 바라보고, 기술 투자와 재무, 조직, 운영 전략을 함께 재설계하는 기업만이 AI를 지속 가능한 성장 동력으로 전환할 수 있을 것입니다. 특히 AI 활용 규모가 커질수록 기술 성능뿐 아니라 비용 효율성, 운영 회복탄력성, 전력과 데이터 확보, 책임 있는 거버넌스가 기업 경쟁력을 좌우하는 핵심 요소가 될 것입니다.

이제 기업이 고민해야 할 질문은 'AI를 도입할 것인가'가 아니라, 'AI를 얼마나 효율적이고 책임 있게 확장할 것인가'입니다. AI를 안정적으로 운영할 수 있는 인프라와 조직 역량을 선제적으로 확보한 기업이 차세대 AI 경쟁의 주도권을 확보하게 될 것입니다.

AI

개요

딜로이트 서베이에 따르면, 2028년까지 AI 팩토리(AI factory)를 대규모로 운영할 계획이라고 응답한 기업이 70%를 넘었다. 이를 실현하기 위해서는 AI 모델, 운영 환경(호스팅), 투자 예산, 인재·역량 등과 관련한 중요한 전략적 의사결정이 필요하다.

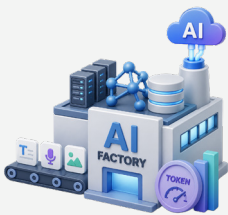
인공지능(AI) 전략은 이제 현실적인 전환점을 맞고 있다. 딜로이트가 기업 대상으로 처음 실시한 AI 인프라 서베이에서 AI 확산을 가로막는 가장 큰 장애물은 더 이상 기업의 의지가 아니라 인프라 부족일 수 있다는 중요한 결과가 도출됐다.

딜로이트는 2025년 12월, 연매출 5억 달러 이상인 미국 기업의 리더 515명을 대상으로 5개 산업에 걸쳐 AI 활용 현황과 엔지니어링, 인프라 확장 전략을 조사했다(보고서 하단 ‘리서치 방법론’ 참조). 응답자의 70% 이상은 2028년까지 AI 팩토리¹와 엣지 AI(AI at the edge, 중앙 클라우드 의존도를 최소화하고 데이터가 생성되는 현장 가까이에서 AI를 실행하는 방식)를 대규모로 구축할 것으로 전망했다. 이는 현재 수준의 약 두 배에 달하는 규모로, 향후 3년간 AI 도입이 급속히 확대될 것임을 시사한다.

이처럼 AI 도입이 급격히 확대됨에 따라 하이브리드 AI(hybrid AI, 서로 다른 인프라 환경에 AI를 분산 배치하는 방식)와 분산형 AI(distributed AI, 사용자와 데이터 가까이에서 AI 연산을 수행하도록 소규모 AI를 분산 배치하는 방식)가 각 기업의 현실적인 운영 모델로 부상하고 있다. 이에 따라 대규모 IT 인프라 전환도 불가피할 것으로 예상된다.

이러한 변화에 직면한 각 기업의 IT 리더에게 이번 조사 결과는 AI와 엔지니어링 분야의 핵심 변곡점과 현재 시장의 위치, 향후 3년간 AI 확장 방향을 가늠할 수 있는 중요한 시사점을 제공한다. 기업이 2028년을 향한 AI 도약을 실현하기 위해서는 애플리케이션 확산부터 하이브리드 인프라 전략, IT 예산의 운영 및 투자 확대, AI를 실질적인 운영 경쟁우위로 전환할 수 있는 인재와 역량 확보까지 다양한 분야에서 전략적이고 균형 잡힌 의사결정을 내려야 한다.

AI 팩토리란?



딜로이트는 AI 팩토리¹를 고성능 컴퓨팅 자원, 네트워크, 스토리지 등 전용 인프라와 AI에 최적화된 소프트웨어 및 서비스를 결합한 AI 전용운영 환경으로 정의한다. AI 팩토리는 모델 학습부터 추론에 이르는 AI 전(全)주기를 대규모로 운영할 수 있도록 설계되며, 텍스트·음성·이미지 등 다양한 멀티모달 데이터를 처리할 수 있다.

AI 팩토리의 핵심 산출물은 지능(intelligence)이며, 이는 토큰 처리량(token throughput)으로 측정된다. 이렇게 생성된 지능은 기업의 의사결정, 업무 자동화, 새로운 AI 서비스 개발을 지원하는 핵심 기반으로 활용되며, AI를 전사적 규모로 운영할 수 있도록 뒷받침한다.

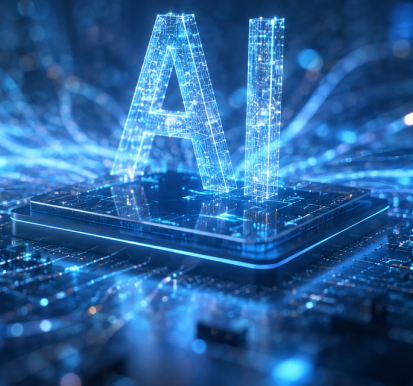
01 AI 운영 본격화... 토큰 수요도 폭증

경쟁사가 AI 파일럿 프로젝트를 실제 운영 환경으로 얼마나 빠르게 확대해 이를 경쟁력으로 전환하는지가 현재 기업들의 뜨거운 관심사다. 딜로이트의 이번 서베이는 이러한 전환 속도를 수치로 보여준다. 결론적으로 현재 AI 애플리케이션은 파일럿 단계를 빠르게 벗어나 실제 운영 단계로 확산되고 있다. 2025년 기준 AI 파일럿 프로젝트를 31건 이상 운영 중인 기업은 전체의 약 50%였으며, 2028년에는 약 70%까지 증가할 것으로 예상된다. 또한 31건 이상의 AI 운영 사례(production-ready use case)를 보유한 기업 비중도 2025년 44%에서 2028년 67%로 확대될 전망이다.

이러한 확산 추세는 AI 시장이 지속적으로 성장하고 있음을 보여준다. 동시에 많은 기업이 현재 운영 중인 AI 서비스와 앞으로 도입할 AI 서비스를 동시에 확장하는 전략을 추진하고 있음을 시사한다. 즉, 기업들은 '현재 운영 중인 AI'와 '차세대 AI'를 병행 확장하는 방향으로 움직이고 있다.

AI 활용 사례가 늘어날수록 AI 워크로드의 복잡성도 함께 증가한다. 응답자의 96%는 현재 AI 워크로드의 복잡성이 '중간 이상' 또는 '매우 높다'고 평가했다. 그러나 딜로이트의 다른 연구에 따르면 대부분의 기업은 아직 기본적인 업무 자동화나 단일 AI 에이전트(single-agent) 활용 단계를 벗어나는 과정에 있으며, 다중 AI 에이전트(multi-agent)를 동시에 운영하기 위한 시스템 구축과 업무 혁신에는 아직 본격적으로 착수하지 못한 상태인 것으로 나타났다.²

여기에 맥락 기반 추론(contextual reasoning), 다단계 추론(multi-turn logic), 멀티모달 생성형 AI(multimodal generative AI)와 같은 고도화된 기능이 추가되면서, 선도 기업조차 AI 운영의 복잡성이 더욱 커지고 있다. 그럼에도 불구하고 AI 확장 도입에 대한 기업들의 자신감은 매우 높다. 응답자의 97%는 향후 3년 내 AI 워크로드를 성공적으로 확장할 수 있다고 답했으며, 이 가운데 상당수는 '매우 자신 있다'고 응답했다.

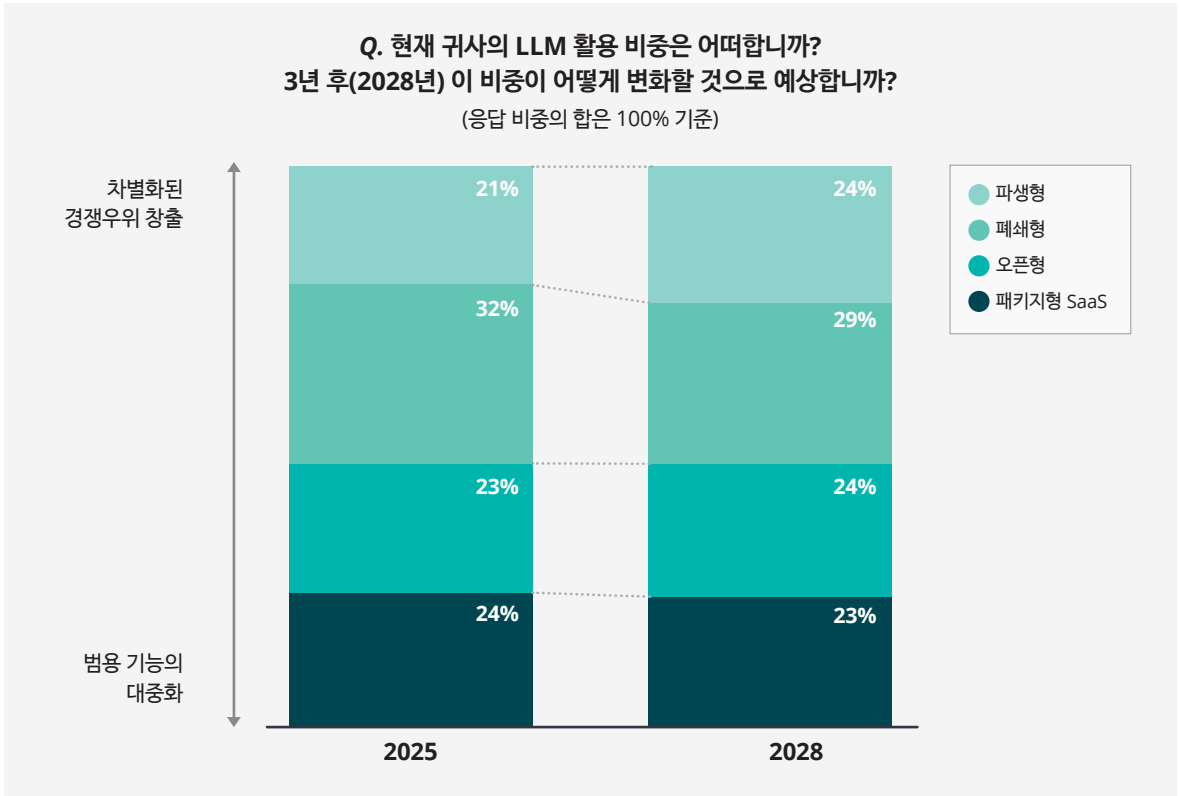


기업들은 AI 워크로드를 운영하기 위해 다음 네 가지 모델 가운데 하나 또는 여러 개를 조합해 활용하고 있다.



현재 가장 많이 활용되는 것은 폐쇄형 모델이다(그림 1). 특히 기술·미디어·통신(TMT)과 같은 혁신 중심 산업(34%)의 기업과 중견기업(35%)에서 활용 비중이 상대적으로 높게 나타났다.

그림 1. 현재 폐쇄형 모델이 LLM 활용 주도하지만, 3년 후 어떤 모델 또는 조합이 시장 주도할지는 미지수



참조: 응답자 수 = 515명

출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

다만 2028년에는 어떤 모델이 시장의 주류가 될 것인지에 대해서는 아직 명확한 방향성이 나타나지 않았다. AI 모델 전략이 매우 빠르게 변화하고 있기 때문에, 현재의 모델별 활용 비중은 설문조사가 실시된 시점의 시장 상황을 반영한 결과일 뿐, 향후 최종적인 시장 구도를 의미하는 것은 아니다.

설문조사 결과와 더불어 딜로이트의 고객 경험을 종합해 보면, AI를 핵심 전략으로 추진하는 선도 기업들은 폐쇄형 모델에 주력하는 것으로 나타났다. 시장 선도 기업의 폐쇄형 모델 활용 비중은 약 60~70% 수준으로 추정된다. 폐쇄형 모델은 고도화된 AI 워크로드를 처리할 수 있는 새로운 기능을 제공하고, 민감한 데이터를 보호하며, AI 서비스의 수익화까지 지원한다는 점에서 차별화된 경쟁력을 제공한다.³

최근 시장의 움직임도 이러한 흐름을 뒷받침하고 있다. 대표적으로 한 글로벌 테크기업이 기존 오픈소스 AI 모델을 독점 모델로 전환한 사례는 기업들의 AI 전략에 영향을 미치며, 폐쇄형 모델 중심의 전략을 더욱 강화하는 요인으로 작용하고 있다.⁴

반면 오픈형 모델과 파생형 모델 역시 낮은 비용으로 AI를 실험하고 혁신할 수 있는 중요한 기반으로 자리 잡고 있다. 강력한 개발자 및 데이터 과학 커뮤니티를 기반으로 지속적인 발전이 이루어지고 있기 때문이다. 예를 들어 엔비디아(NVIDIA)는 2026년 새로운 오픈 모델과 학습 도구, 데이터셋을 공개하며 오픈소스 AI 생태계를 더욱 확대했다.⁵ 또한 오픈소스 모델의 성능은 최고 수준 모델에 근접한 경우가 많으며, 규모가 작고 특정 목적에 특화된 AI 활용 사례에서는 오히려 더 적합할 수 있다.⁶

한편 이번 조사 결과는 에이전틱 SaaS가 앞으로도 지속적으로 확산될 가능성을 시사한다.⁷ 이에 따라 기업의 핵심 기록관리 시스템 역시 AI 에이전트가 내재된 형태로 진화할 것으로 예상된다. 동시에 AI 활용 수준을 반영한 새로운 가격 정책과 과금 모델도 함께 등장할 전망이다.

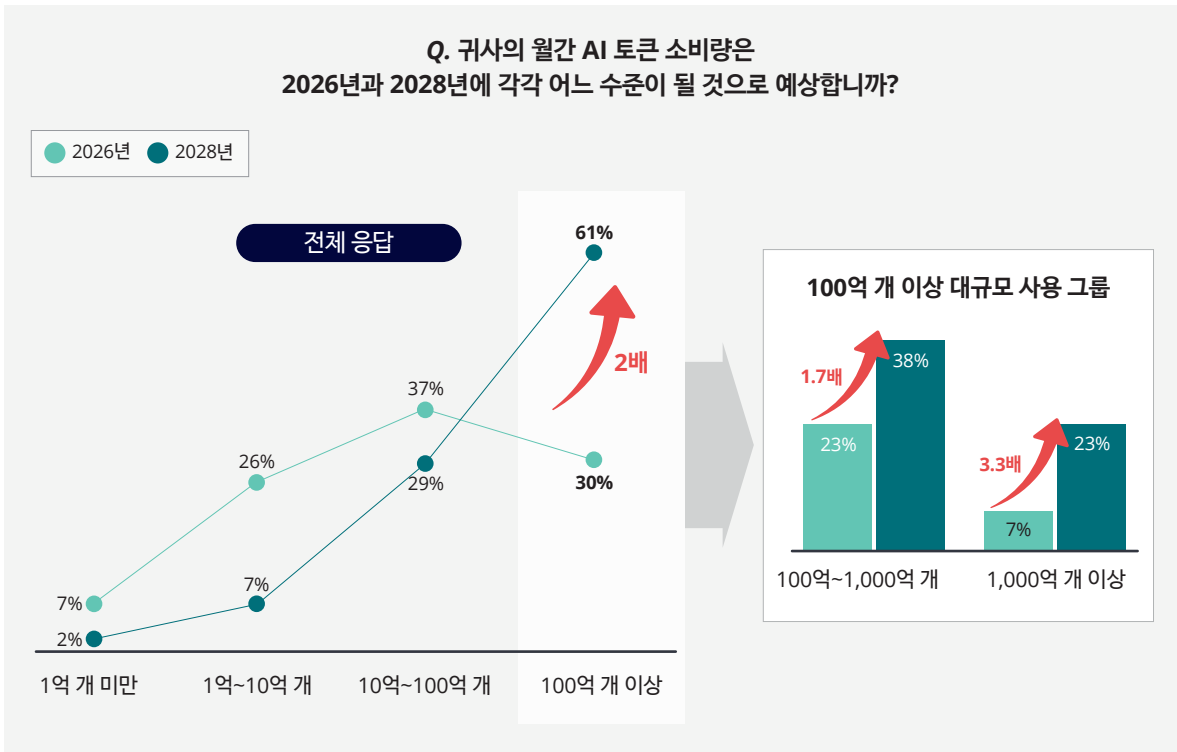
결국 네 가지 AI 모델은 각각 고유한 장점을 가지고 있으며, 어떤 모델을 선택할지는 크게 두 가지 요인에 의해 결정될 가능성이 크다. 첫째, 앞으로 어떤 새로운 AI 모델이 시장에 등장할지가 모델 선택에 영향을 미칠 것이다. 두 번째는 AI를 경쟁우위를 창출하기 위한 차별화 수단으로 활용할 것인지, 아니면 기존 업무를 더 빠르고 저렴하게 수행하기 위한 표준 역량의 확산 수단으로 활용할 것인지에 달려 있다.

02 토큰 사용량 급증, AI 확산의 신호이자 새로운 비용 리스크

AI 활용이 고도화되고 활용 범위가 확대되면서 토큰(token, AI 애플리케이션과 인프라에서 연산량과 비용을 결정하는 기본 단위) 소비량이 폭발적으로 증가하고 있다.

응답 기업의 37%는 월평균 10억~100억 개의 토큰을 사용하고 있으며, 30%는 이미 100억 개 이상을 소비하고 있다. 나아가 2028년 월평균 토큰 사용량이 100억 개를 넘을 것이라 답한 비율은 61%에 달해, 2년 만에 두 배 증가할 것으로 집계됐다(그림 2). 월 1,000억 개 이상의 토큰을 사용하는 기업은 같은 기간 약 3배 증가할 것으로 예상된다.

그림 2. 토큰 월 100억 개 이상 사용하는 기업 비중 2026~2028년 약 두 배 증가 전망



참조: 응답자 수 = 515명

출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.



토큰 소비 급증의 배경에는 고도화된 추론 모델의 확산이 있다. 이러한 모델은 답변을 생성하기 전에 내부 추론 과정에서 이른바 '사고 토큰'(thinking tokens)을 추가로 생성하기 때문에 더 많은 연산량이 필요하다.⁸

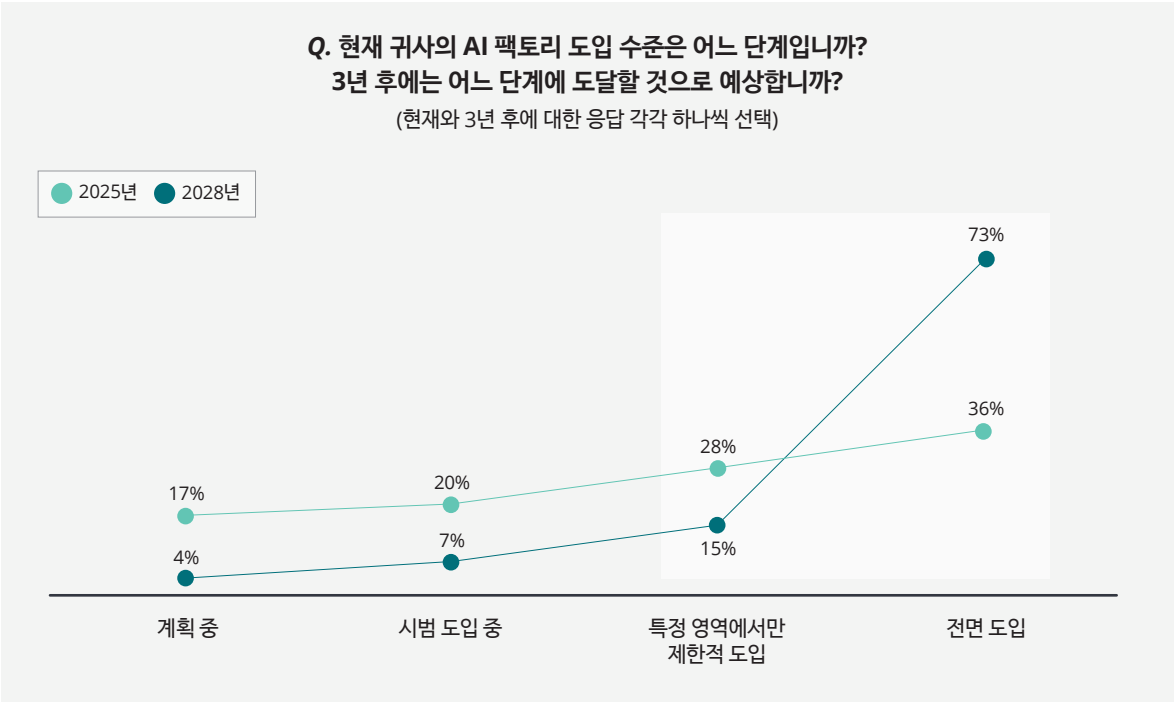
다만 토큰 사용량 증가가 반드시 AI를 성공적으로 활용한다는 의미는 아니다. 특히 초기 도입 단계에서는 과도하게 긴 프롬프트, 비효율적인 컨텍스트 관리, 재사용 부족 등 비효율적 설계로 인해 토큰만 과다하게 소비되는 경우도 적지 않다. 따라서 기업은 AI 애플리케이션의 전략적 가치와 함께 토큰 소비에 따른 비용 구조를 지속적으로 점검해야 한다. AI 활용이 확대될수록 우수한 솔루션 설계와 엔지니어링, 시스템 아키텍처를 통해 입출력 토큰을 효율적으로 관리하는 것이 중요하다.

향후 AI 애플리케이션의 수와 복잡성이 계속 증가하면 토큰 소비량은 인프라 수요와 공급 변화를 가장 먼저 보여주는 핵심 선행지표로 자리 잡을 가능성이 크다.

03 AI 팩토리와 엣지, 2028년까지 두 배 확대

기업들은 앞으로도 클라우드 등 다양한 환경에서 AI를 운영하겠지만, 자체 AI 팩토리를 구축하는 기업이 빠르게 늘어날 전망이다. 딜로이트 서베이에 따르면, 응답자의 64%가 현재 AI 팩토리를 제한적 또는 대규모로 구축하고 있으며, 2028년에는 해당 비율이 88%로 증가할 것으로 예상된다. 특히 73%는 2028년까지 AI 팩토리를 대규모로 운영할 것으로 전망해, 향후 3년이 AI 인프라 확장의 전환점이 될 것으로 보인다(그림 3).

그림 3. 대규모 AI 팩토리 구축하는 기업, 향후 3년간 두 배 가까이 증가 전망



참조: 응답자 수 = 515명

출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

옛지 AI 역시 같은 흐름을 보인다. 현재 옛지 AI를 대규모 운영하는 기업은 36%에 불과하지만, 2028년에는 72%까지 증가할 전망이다. 기업들이 가장 많이 선택한 방식은 클라우드에서 통합 관리하면서 실제 연산은 현장에 배치한 장비에서 수행하는 '클라우드 관리형 옛지'로, 응답자의 68%가 이를 도입했다고 답했다. 이는 별도의 운영체계를 새로 구축하기보다 클라우드의 장점과 옛지 컴퓨팅의 효율성을 동시에 활용하려는 전략으로 해석된다.

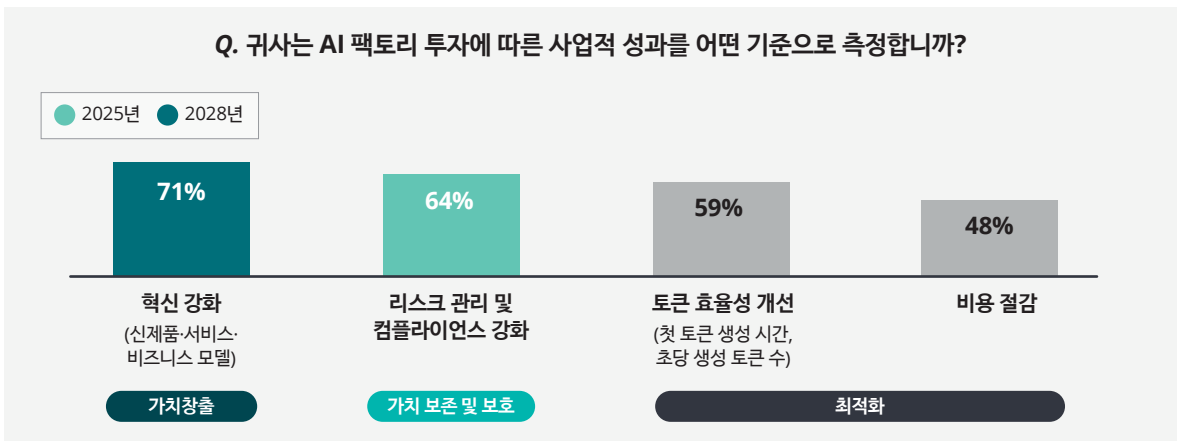
한편 서베이 결과 기업들이 텍스트·음성·이미지 등 다양한 데이터를 처리하는 멀티모달 AI와 더불어 독점 모델 및 오픈소스 모델을 함께 활용하는 멀티모달 AI를 지향하고 있는 것으로 나타났다. 이에 따라 멀티클라우드와 하이브리드 클라우드(퍼블릭 클라우드와 프라이빗·온프레미스 AI 인프라를 결합한 구조)가 성능, 비용, 데이터 주권, 통제력을 동시에 확보하기 위한 핵심 아키텍처로 자리 잡고 있다.

AI 팩토리 확대가 예상되지만 현실적인 제약도 적지 않다. 응답자의 51%는 경제 불확실성을 가장 큰 투자 제약 요인으로 꼽았다. AI 팩토리는 대규모 자본 투자와 장기간의 투자 회수가 필요한 만큼 거시경제 환경의 영향을 크게 받기 때문이다. 이 외 주요 장애요인으로 △ 조직 내 사업 과제(48%) △ 규제 부담(48%) △ AI 인재 및 역량 부족(40%) 등이 꼽혔다.

이러한 제약에도 불구하고 응답자들은 AI 팩토리의 핵심 가치로 △ 혁신 가속화(71%) △ 리스크 관리 강화(64%) △ 토큰 사용 최적화(59%) 등을 꼽았다.

또한 AI 인프라가 하이브리드 형태로 진화하면서 벤더별 AI 프레임워크(55%), 고성능 AI 네트워크 인프라(50%), 기반 모델(47%)이 그래픽처리장치(GPU) 활용도를 높이고 복합 인프라 운영을 지원하는 핵심 기술로 주목받고 있다. 결국 AI 인프라 경쟁의 초점은 어떤 아키텍처를 선택할 것인가에서 AI를 얼마나 경제적이고 효율적으로 대규모 운영할 것인가로 빠르게 이동하고 있다.

그림 4. 응답자 71%, 신제품·서비스·비즈니스모델 혁신을 AI 팩토리 투자의 핵심 성과 지표로 인식



참조: 응답자 수 = 515명

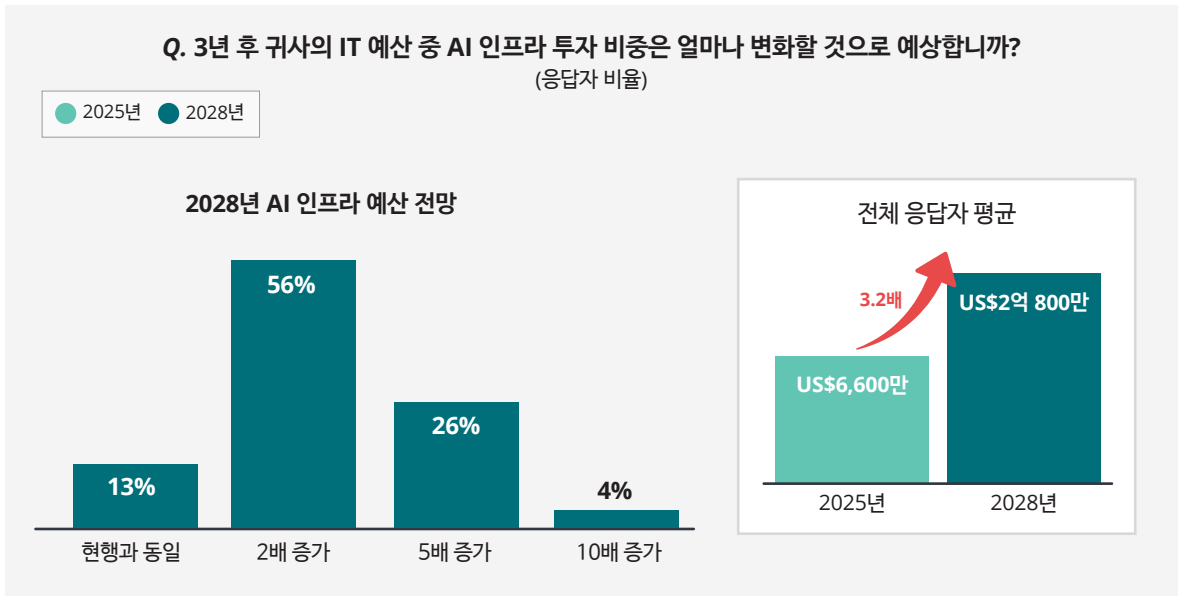
출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

04 AI 인프라 투자 3배 급증 시대, 재무 관리 혁신이 필수

AI가 기업 전반으로 빠르게 확산되면서 AI 애플리케이션과 인프라 투자 부담도 급격히 증가하고 있다. 딜로이트의 이전 조사⁹에 따르면, 2025년에는 이미 많은 기업에서 AI가 IT 예산의 절반 이상을 차지하고 있으며, 사업 부문과 IT 부문의 수요 증가에 힘입어 투자 규모도 매년 크게 확대되고 있다. 이제 AI 인프라 투자는 선택이 아닌 필수 요소가 됐지만, 비용 투명성, 예산 배분, 투자성과 관리 체계는 아직 충분히 성숙하지 않은 실정이다.

향후 이러한 투자 확대는 더욱 가속화될 전망이다. 응답자의 86%는 향후 3년간 AI 인프라 예산이 증가할 것으로 예상했으며, 평균적으로 예산 규모는 현재의 3배 이상, 대기업은 최대 4배 가까이 확대될 것으로 전망했다(그림 5). 이는 AI가 단발성 프로젝트를 넘어 장기적인 핵심 투자 영역으로 자리 잡고 있음을 의미한다. 이에 따라 최고기술책임자(CTO)와 최고재무책임자(CFO)는 운영비와 자본투자의 최적 조합을 공동으로 설계하고, 토큰 경제(token economics), 총소유비용(TCO), 조직 전반의 예산 영향을 고려한 새로운 재무관리 체계를 구축해야 한다.

그림 5. AI 인프라 투자 규모, 2028년까지 평균 3.2배, 대기업은 3.9배 각각 증가 전망



참조: 응답자 수 = 515명

참조: AI 인프라 투자 규모는 IT 예산 구간을 기준으로 산출했으며, 1억 달러 이상 구간의 상한값은 2025년 11월 12일자 인터넷데이터센터(IDC) 기술 투자 데이터베이스를 바탕으로 16억 달러로 가정했다. 평균값의 평균을 적용해 산출했기 때문에 실제 금액과 일부 차이가 있을 수 있다.

출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

AI 인프라가 확대됨에 따라 비용뿐 아니라 물리적 인프라 확보 경쟁도 심화되고 있다. 고대역폭 메모리(HBM)와 기타 비휘발성 메모리 부품 가격이 빠르게 상승하면서¹⁰ AI 팩토리 구축 비용이 증가하고 있다. 게다가 수요가 공급을 초과하면서 조달 기간까지 길어져 비용 부담은 더욱 커질 전망이다.¹¹ 웨이퍼 가격도 약 20% 상승할 것으로 예상돼¹² GPU와 중앙처리장치(CPU) 가격 역시 추가 상승이 예상된다.

AI 팩토리가 유틸리티급(utility-grade) 규모로 확장될수록 전력 생산과 전력망 연계는 핵심적인 제약 요인이자 새로운 투자 영역으로 부상할 전망이다. 다만 AI 인프라의 확장 방식은 활용 환경에 따라 달라질 수 있다. 옛날 AI는 제한된 공간에서 CPU와 소형 언어모델(SLM)을 중심으로 운영되는 반면, 중앙집중형 AI 팩토리는 소규모 GPU 클러스터부터 전용 캠퍼스 규모의 대형 인프라까지 다양한 형태로 구축될 수 있다.

서베이에 참여한 리더의 47~48%는 전력 수요를 충족하기 위해 전력망 공급, AI 팩토리 내부 또는 인근 자가 발전, 외부 사업자가 구축한 인근 발전 설비를 병행하는 방식을 활용할 것으로 예상했다. 이러한 방식은 단일 전력원에 대한 의존도를 낮추고 운영 복원력과 계약 유연성을 높일 수 있다. 또한 냉각 방식 역시 전력 사용량과 투자 전략에 따라 기존 공랭식부터 액체 냉각까지 다양한 방식이 적용될 것으로 전망된다.

결국 AI 인프라 구축 비용과 전력 수요가 급증하면서, 누가 의사결정을 내리고 어떤 기준으로 비용과 성능 간 균형을 조정할 것인가가 핵심 과제로 떠오르고 있다. 특히 의사결정 권한과 책임, 예산이 조직 내 여러 부서에 분산된 기업일수록 이러한 거버넌스 체계의 중요성이 더욱 커질 것으로 예상된다.

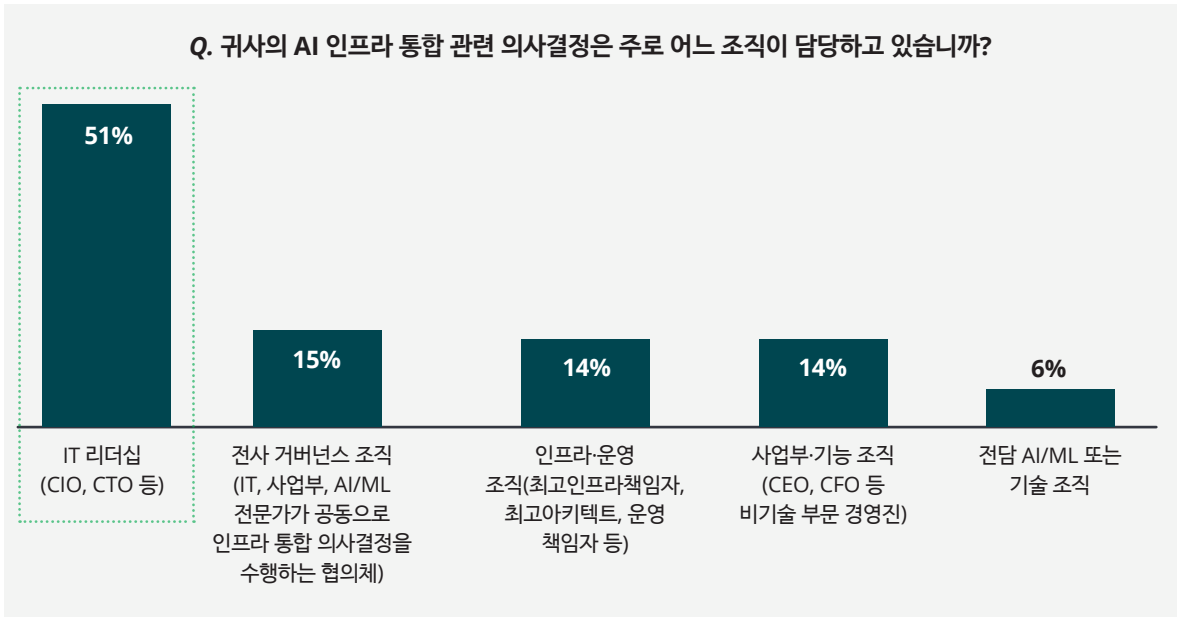


05 AI 스케일업의 병목은 기술이 아닌 인재

비즈니스와 재무 관리 측면에서 AI의 영향력이 커지고 있음에도, AI 관련 의사결정은 여전히 IT 부문이 주도하고 있다. 응답자의 절반 이상(51%)은 인프라 통합과 관련한 의사결정 권한이 최고정보책임자(CIO)와 최고기술책임자(CTO) 등 IT 리더십에 있다고 답했다. 나머지 응답은 거버넌스 조직(15%), 인프라 조직(14%), 사업·기능 조직(14%), AI 전문 조직(6%) 등으로 비교적 고르게 분산됐다(그림 6).

딜로이트는 이러한 구조가 반드시 문제를 의미하는 것은 아니라고 본다. 실제로 IT 리더들은 AI 확장 도입에 대비해 기존 데이터센터를 전환하고 신규 인프라 부지를 선제적으로 확보하는 등 기반 마련을 주도하고 있다. 또한 현재의 애플리케이션 환경이 CPU 중심으로 운영되고 있지만, 앞으로의 운영 모델은 CPU와 GPU를 모두 활용할 수 있어야 한다는 점을 인식하고 있다. 이에 따라 CIO와 CTO는 플랫폼 준비 수준을 사업 수요와 맞추기 위해 관련 의사결정을 선제적으로 이끌며 AI 인프라 전환을 주도하는 것으로 나타났다.

그림 6. 기업 절반 이상이 AI 인프라 의사결정을 CIO·CTO 등 IT 리더십이 주도



참조: 응답자 수 = 515명

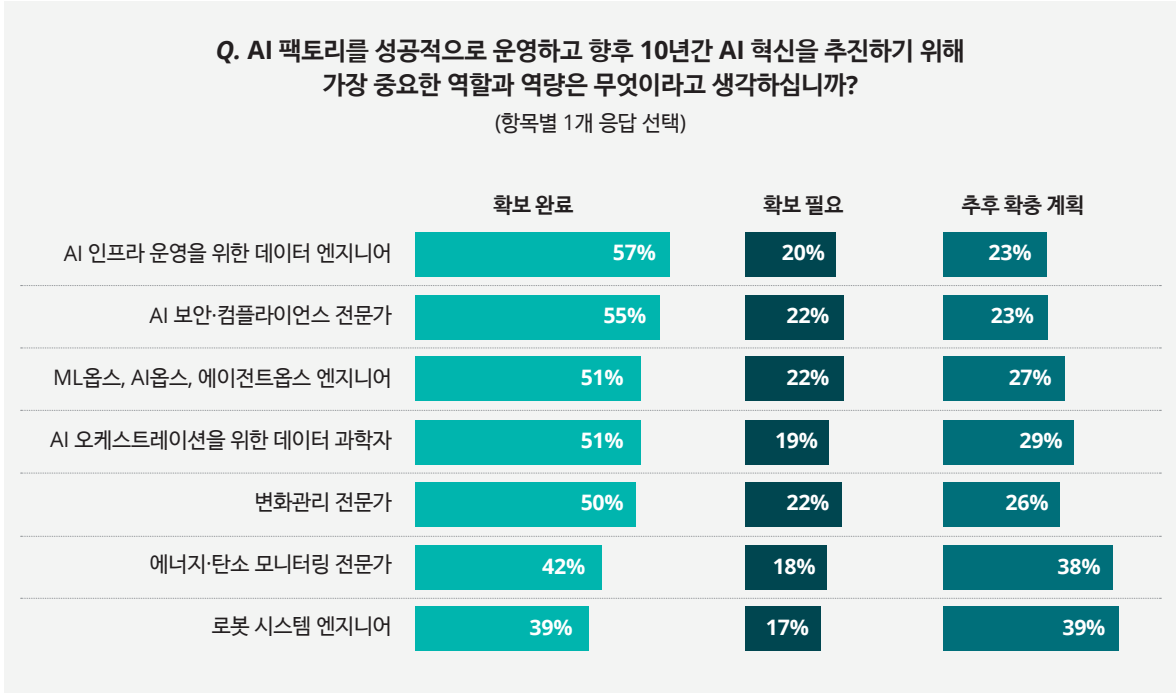
출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

대다수 기업이 AI 확장 도입에 필요한 예산을 확보하고 있지만, 인재 확보는 여전히 가장 큰 과제로 남아 있다. 첫 번째 과제는 테크 인력 부족이다. 응답자의 절반 이상은 데이터 엔지니어, 보안 및 컴플라이언스 전문가, AI·머신러닝(ML) 및 에이전트 운영 엔지니어, 데이터 과학자를 이미 보유하고 있다고 답했지만, AI를 대규모로 확산하고 혁신을 이끌기에는 전문 인력이 여전히 충분하지 않은 것으로 나타났다(그림 7).

특히 AI의 성공은 단순한 시스템 구축이 아니라 일상 업무에 안정적으로 정착시키는 데 달려 있는 만큼, 보안·컴플라이언스 전문가와 AI/ML 및 에이전트 운영 엔지니어, 변화관리 전문가에 대한 수요가 더욱 커질 것으로 예상된다. 장기적으로는 로봇틱스 시스템 엔지니어링과 탄소·에너지 모니터링 역량에 대한 수요도 확대되면서, 물리적 자동화와 에너지 효율 중심의 AI 활용이 본격화될 것으로 전망된다.

두 번째 과제는 부서 간 역량 격차다. 응답자의 81%는 IT 부서가 AI 확산을 추진하는 데 필요한 기술적·재무적 역량을 갖추고 있다고 평가한 반면, 사업 및 제품 부서에 대해서는 65%만이 같은 평가를 내렸다. 두 부서 간에는 16% 포인트의 역량 격차가 존재하는 셈이다. 딜로이트는 향후 AI 전략을 성공적으로 실행하기 위해서는 이러한 핵심 분야의 인재 육성과 역량 개발에 대한 지속적인 투자가 필요하다고 분석했다.

그림 7. 기업이 현재 가장 필요로 하는 핵심 인재는 AI/ML 운영, 에이전트 운영, 변화관리, 보안·컴플라이언스 전문가



참조: 응답자 수 = 515명

출처: Deloitte Center for Integrated Research analysis of data collected from November to December 2025 from US business and technology leaders.

06 2028년 AI 대규모 확장 시대 경쟁력, 지금 결정된다

AI 수요는 빠르게 증가하고 있지만, 대다수 기업에게 AI를 대규모로 확장 도입하는 과정은 여전히 불확실한 의사결정의 연속이다. 어떤 AI 모델을 선택할 것인지부터 토큰 사용량을 어떻게 관리할 것인지, 워크로드를 클라우드와 AI 팩토리 중 어디에서 운영할 것인지까지 지금의 선택이 앞으로 기업의 기술 경쟁력과 재무 구조를 좌우할 것으로 예상된다. 향후 3년간 기업들이 목표로 하는 AI 확장 수준은 매우 야심차지만 충분히 실현 가능한 수준이다.

IT 리더는 지금부터 2028년을 준비해야 한다. AI 사용량을 재무 시스템 수준으로 정밀하게 측정 및 감사하고, 인프라 투자를 실질적인 사업 가치와 연계하며, 전력·인재·거버넌스를 핵심 아키텍처 요소로 관리해야 한다. 앞으로 경쟁력을 확보하는 기업은 단순히 AI를 더 빠르게 확장하는 기업이 아니라, 경제성과 운영 회복력, 명확한 책임 체계를 기반으로 AI를 확장하는 기업이 될 것이다. 향후 3년은 AI가 지속 가능한 경쟁우위로 자리 잡을지, 아니면 구조적 비용 부담으로 남을지를 결정하는 중요한 분기점이 될 전망이다.

리서치 방법론

본 조사는 2025년 11~12월 딜로이트가 실시한 AI 인프라 설문조사를 기반으로 한다. 조사 대상은 미국 소비재, 에너지·자원·산업재, 금융서비스, 생명과학·헬스케어, 기술·미디어·통신(TMT) 등 5개 산업 기업의 비즈니스 및 테크 부문 의사결정권자 515명이다.

응답자는 연매출 5억 달러 이상의 상장·비상장 기업에 재직 중이며, 이사급 이상 임원과 최고경영진(C레벨), 이사회 구성원 등 조직의 AI 전략과 투자 의사결정에 관여하는 리더로 구성됐다.

주석

1. Nicholas Merizzi et al., "[Navigate the economics of AI](#)," Deloitte, Jan. 12, 2026.
2. Tim Smith, Gregory Dost, Garima Dhasmana, Parth Patwari, Diana Kearns-Manolatos, and Iram Parveen, "[AI is capturing the digital dollar. What's left for the rest of the tech estate?](#)" Deloitte Insights, Oct. 16, 2025.
3. Brian Eastwood, "[AI open models have benefits. So why aren't they more widely used?](#)" MIT Sloan School of Management, Jan. 20, 2026.
4. Jonathan Vanian, "[From llamas to avocados: Meta's shifting AI strategy is causing internal confusion](#)," CNBC, Dec. 9, 2025.
5. Kari Briski, "[NVIDIA unveils new open models, data and tools to advance AI across every industry](#)," NVIDIA, Jan. 5, 2026.
6. Tim Lu, "[Frontier models explained: What defines the cutting edge of AI](#)," Datacamp, Jan. 13, 2026.
7. David Jarvis, Sayantani Mazumder, Girija Krishnamurthy, Gopal Srinivasan, China Widener, and Gillian Crossan, "SaaS meets AI agents: Transforming budgets, customer experience, and workforce dynamics," Deloitte Insights, Nov. 18, 2025.
8. Eamon Phelan, "[Think fast: Reasoning at 3ms a token](#)," Fin, July 18, 2025.
9. Smith, Dost, Dhasmana, Patwari, Kearns-Manolatos, and Parveen, "[AI is capturing the digital dollar. What's left for the rest of the tech estate?](#)"
10. Kif Leswing, "[AI memory is sold out, causing an unprecedented surge in prices](#)," CNBC, Jan. 10, 2026.
11. Getty, Doug, "[AI is not a cycle, it is a structural reset of the global hardware economy](#)," Rand Technology, Jan. 28, 2026.
12. Lisa Wang, "[Prices of 8-inch wafers to rise 20 percent, TrendForce says](#)," Taipei Times, Jan. 14, 2026.
13. Martin Stansbury, Kelly Marchese, Kate Hardin, and Carolyn Amon, "[Can US infrastructure keep up with the AI economy?](#)" Deloitte Insights, June 24, 2025.

한국 딜로이트 그룹 산업 전문가

인공지능 및 데이터(Artificial Intelligence & Data) 서비스

딜로이트의 인공지능 및 데이터(Artificial Intelligence & Data) 서비스는 비즈니스 인사이트와 기업 인텔리전스 간의 유기적 연계를 입증하여 지속 가능한 성장을 견인합니다. 데이터 분석, 로보틱스 처리 기술, 그리고 고급 인공지능을 융합하여 기업의 성장과 효율성을 개선하는 핵심 통찰력을 확보하도록 돕습니다. 이를 통해 신속하고 안정적인 의사 결정 체계를 구축하고, 업무 효율성을 개선하며, 잠재적 리스크를 선제적으로 방어합니다.

김수연 파트너

EA&D(Engineering, AI & Data) |
컨설팅 부문

☎ 02 6676 3434
@ skim86@deloitte.com

김정민 파트너

Transformation Strategy | 컨설팅 부문

☎ 02 6676 3379
@ jungminkim@deloitte.com

정창모 수석위원

AI Agent(생성형 AI), Data Analytics |
컨설팅 부문

☎ 02 6676 3288
@ changjung@deloitte.com

Deloitte Insights

세일즈&마케팅 대표

권지원 Partner
jekwon@deloitte.com

성장전략부문 부대표

서정욱 Partner
juseo@deloitte.com

딜로이트 인사이트 편집장

박경은 Director
kyungepark@deloitte.com

연구원

김선미 Senior Manager
seonmikim@deloitte.com

디자이너

박근령 Senior Consultant
keunrpark@deloitte.com

Contact us

krinsightsend@deloitte.com



앱

Download on the
App StoreGET IT ON
Google Play

카카오톡 채널



'딜로이트 인사이트' 앱과 카카오톡 채널에서
경영·산업 트렌드를 만나보세요!

Deloitte.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited (“DTTL”), its global network of member firms, and their related entities (collectively, the “Deloitte organization”). DTTL (also referred to as “Deloitte Global”) and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

Deloitte Asia Pacific Limited is a company limited by guarantee and a member firm of DTTL. Members of Deloitte Asia Pacific Limited and their related entities, each of which are separate and independent legal entities, provide services from more than 100 cities across the region, including Auckland, Bangkok, Beijing, Hanoi, Hong Kong, Jakarta, Kuala Lumpur, Manila, Melbourne, Osaka, Seoul, Shanghai, Singapore, Sydney, Taipei and Tokyo.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited (“DTTL”), its global network of member firms or their related entities (collectively, the “Deloitte organization”) is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of DTTL, its member firms, related entities, employees or agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities.

본 보고서는 저작권법에 따라 보호받는 저작물로서 저작권은 딜로이트 안진회계법인(“저작권자”)에 있습니다. 본 보고서의 내용은 비영리 목적으로만 이용이 가능하고, 내용의 전부 또는 일부에 대한 상업적 활용 기타 영리목적 이용시 저작권자의 사전 허락이 필요합니다. 또한 본 보고서의 이용시, 출처를 저작권자로 명시해야 하고 저작권자의 사전 허락없이 그 내용을 변경할 수 없습니다.