Market Insight | 2026.6.25
그림: Midjourney(초지능 가리기 REDRED)

AI Hot Issue

Mythos/Fable 셧다운: Public API 단일 의존은 REDRED

한종목

chongmok.han@miraeasset.com

CONTENTS

Executive Summary 3
셧다운은 멈춤이 아니라, 수요의 강제 이전이다 3

I. Mythos/Fable 셧다운 4
1. 칩 통제에서 모델 접근통제로 4
2. 수요 붕괴가 아니라 아키텍처 단절 6
3. 자체 운영은 왜 경제적으로 불리한가 8
4. 밸류체인은 어디로 재편되는가 11
5. 투자 결론: 알파는 신뢰 경계에서 나온다 15

Executive Summary

셧다운은 멈춤이 아니라, 수요의 강제 이전이다

이번 달 Anthropic의 Fable/Mythos 셧다운을 두고 시장의 첫 반응은 둘로 갈렸습니다. 한쪽은 “프론티어 랩의 해외 매출은 0이 되고 핵심 인력이 연구에서 배제되니 AI TAM이 붕괴한다”고 말하고, 다른 한쪽은 “일시적 규제 잡음일 뿐”이라고 말합니다. 저희는 두 해석 모두 핵심을 비껴간다고 봅니다.

이번 보고서의 결론을 먼저 말하면 이렇습니다. 이 사건의 본질은 칩 수출통제에서 모델 접근통제로의 이동이며, AI 모델의 public serving의 중단이 capability race의 중단을 뜻하지는 않습니다.

더 중요한 투자적 함의는 따로 있습니다. 이 사태가 엔터프라이즈에게 “퍼블릭 API는 한 장의 서한으로 꺼질 수 있는 단일 장애점(SPOF)”이라는 트라우마를 각인시키면서, AI 워크로드의 수요 곡선을 관리형 프라이빗(managed-private) 한 칸으로 강제 이전시키고 있다는 점입니다.

즉 셧다운된 것은 특정 모델이지, 수요가 아닙니다. 수요는 사라진 게 아니라 자리를 옮겼습니다. 핵심 논지는 다섯 가지입니다.

첫째, 시장이 그다지 집중해서 보지 않은 이 사건에 대해 저희는 구조적 단절을 보고 있습니다. 모델 접근권 자체가 수출 통제와 국가 안보, 동맹국 접근권의 대상이 되었고, 이는 OpenAI와 구글, xAI의 최상위 모델에도 대동소이하게 적용될 수 있는 선례입니다.

둘째, 대부분의 기업들의 대응은 '오픈소스 자체 운영'이 아니라 '관리형 Private AI'가 될것입니다. 이 방향은 운영 역량, 책임 소재, 보안 책임자의 선호 같은 정성적 근거뿐 아니라, 추론 비용이라는 정량적 근거로도 뒷받침됩니다.

셋째, 물리적 온프레미스 인프라를 도입해야 하는 근거는 사실 '비용 우위'도 있지만, '주권 프리미엄'이라고 생각합니다. 싸기 때문에 들이는 것이 아니라, 데이터가 절대 외부로 나갈 수 없고, 서비스가 끊겨서는 안 되며, 지연 시간을 감내할 수 없기 때문에 비싼 줄 알면서도 들입니다. 비용 논리와 업무 연속성 논리는 분리해서 보아야 합니다.

이러한 전환에 따라, 수혜를 얻을 수 있는 곳들로 Dell(프라이빗 추론 설비 투자의 순증), Cisco(엔터프라이즈 보안 네트워크), Palantir(AI 운영·거버넌스 계층), 그리고 ‘신뢰 경계’를 장악하는 관리형 클라우드를 꼽았습니다.

또한 투자자로서 결국 관찰해야 할 것은 이제 AI 모델의 성능 점수 같은 것이 아닙니다. 각 기업이 AI를 실제 업무에 붙일 때 무엇을 좋은 답으로 볼지, 누가 승인할지, 어떤 모델로 바꿔 쓸지를 가장 자연스럽게 정하게 만드는 플랫폼이 최종 승자가 됩니다.

결국, 셧다운된 것은 AI 모델이지 ‘지능에 대한 수요’ 그 자체는 아니며, 그 수요는 더 안전하고 통제 가능한 ‘관리형 Private AI’의 길목으로 이동하고 있다는 것을 눈치채야 합니다.

I. Mythos/Fable 셧다운

1. 칩 통제에서 모델 접근통제로

(1) 무엇이 일어났는가

Anthropic은 Claude Fable 5를 일반 사용자용 Mythos급 모델로, Claude Mythos 5를 일부 사이버 방어 기관과 인프라 사업자에게만 제공하는 제한 접근 모델로 소개했다. 두 모델은 같은 기반 모델이지만, Mythos 5는 일부 사이버 안전장치가 완화된 형태로 Project Glasswing을 통해 제한 배포되는 구조였다.

그 직후인 6월 12일 오후 5시 21분(미국 동부시간), 미 상무부는 국가안보 수출통제 권한을 근거로 두 모델에 대한 모든 외국인 접근 중단을 지시했고, Anthropic은 국적별 선별 차단이 기술적으로 불가능하다는 이유로 전체 고객 접근을 전면 중단했다. 정부가 통상의 의견수렴·관보 절차까지 건너뛰고 '금요일 저녁'에 즉시 발동한 배경에는, NSA 국장이 상원 정보위 브리핑에서 "Mythos가 NSA 기밀 시스템의 거의 전부를 수 시간 내에 침투했다"고 증언한 사실이 있는 것으로 전해진다. 표면적으로는 특정 모델의 셧다운이지만, 그 방아쇠는 프론티어 모델의 공격적 사이버 능력 그 자체였다.

이번 사건에서 확인된 것은 모델 접근권 자체가 국가 안보 통제의 대상이 될 수 있다는 점이다. 과거 AI 통제의 중심은 GPU, HBM, 첨단 패키징, 네트워킹 장비 같은 물리적 병목이었다. 그러나 Fable/Mythos 사태는 통제축이 반도체에서 모델 접근권으로 확장되고 있음을 보여준다. "누가 칩을 살 수 있는가"의 문제에서 "누가 모델을 호출할 수 있는가"의 문제로 전장이 넓어진 것이다.

(2) 왜 이것이 구조적 선례인가

규제의 위력은 속도에서 나온다. 통상적인 규제는 의견 수렴과 관보 게재를 거치지만, 이번 사안처럼 정부가 특정 기업에 즉시 통보하는 방식은 기업의 대응 시간을 거의 남기지 않는다. 고객은 사전 예고 없이 모델 접근권을 잃을 수 있고, 기업은 계획된 마이그레이션 기간 없이 대체 모델이나 하위 모델로 이동해야 한다.

더 중요한 문제는 간주 수출이다. 기술이나 소스코드, 모델 가중치, 내부 논의에 미국 내 외국인이 접근하는 것까지 수출로 간주될 수 있다면, 단순히 외부 고객만 차단되는 것이 아니다. 시민권이나 영주권이 없는 핵심 엔지니어가 특정 코드베이스와 모델 개발 루프에서 격리될 수 있다. 이는 public serving을 넘어 다음 모델 개발 파이프라인 자체에 영향을 줄 수 있는 정밀 타격이다.

이 사건의 또 다른 특이점은 클라우드 사업자의 위치다. Amazon은 Anthropic의 투자자이자 AWS 인프라 파트너이고, 동시에 Bedrock과 자체 모델 Nova를 보유한 경쟁자다. 한 행위자가 자본 공급자, 클라우드 공급자, 모델 유통자, 플랫폼 경쟁자라는 역할을 동시에 쥐고 있다.

그리고 결정적으로, 이번 워싱턴이 날린 지시의 근거가 된 jailbreak 시연을 만든 주체가 Amazon 연구진이며, CEO Andy Jassy가 지시 발동 이전에 직접 정부에 모델 우려를 제기한 것으로 보도됐다. 즉 Amazon은 위험을 수동적으로 '전달한 통로'가 아니라, 통제를 촉발한 능동적 방아쇠였던 것이다.

따라서 이 사태는 단순한 "Anthropic 對 정부"의 양자 구도가 아니라 모델 기업, 클라우드 사업자, 정부, 경쟁사가 하나의 전략 게임에 묶인 사건이다.

투자자에게 필요한 결론은 명료하고 단순하다. 프론티어 모델은 더 이상 전통적인 SaaS처럼 안정적으로 배포되는 소프트웨어가 아니다. 사이버, 자동화, 지식 노동을 수행하는 이중용도 전략 인프라가 되었고, 그 접근권은 언제든 정책 변수의 '직접 대상'이 될 수 있다.

2. 수요 붕괴가 아니라 아키텍처 단절

(1) Public API의 SPOF화

이 사태를 바라볼 때 흔히 하는 첫 번째 오독은 "규제는 곧 연구 지연"이라는 등식이다. 그러나 정부가 막은 것은 외부 배포와 대중 접속이지, 데이터센터 내부의 연산 그 자체가 아니다. 외부 서비스가 중단되더라도 내부의 학습, 평가, 레드팀 점검, 합성 데이터 생성, 에이전트 연구 루프는 계속 돌아갈 수 있다. 오히려 일반 서비스 부담이 줄면 일부 자원을 내부 연구로 돌릴 유인까지 생길 수 있다.

따라서 사용자가 보는 모델만으로 프론티어 랩의 실제 능력을 평가하면 최전선을 과소평가할 수 있다. public serving과 capability race는 같은 것이 아니다. 공개된 제품은 빙산의 윗부분일 뿐이다.

두 번째 착각은 "해외 매출 훼손은 곧 시장 붕괴"라는 해석이다. 단기 매출 훼손은 사실이지만, 그 크기와 지속성은 시장의 공포보다는 작다. 셧다운된 것은 Fable 5·Mythos 5 두 모델뿐이고, 매출 베이스의 대부분을 떠받치는 Opus 4.8 등 나머지 모델은 내내 정상 서빙됐다(5월 기준 run-rate 매출 $47B는 그대로 가동 중이다).

게다가 복원 경로도 이미 형성되고 있다. 7월 8일 발효되는 정부 ID 검증 절차가 미국 우선 복원의 유력한 통로이고, 예측시장에서는 7월 1일 이전 복원 확률을 Kalshi 약 57%, Polymarket 34~41%로 매기고 있다.

그러나 이번 사건이 남긴 진짜 변수는 일시적 매출이 아니라 영구적으로 각인된 아키텍처다. 기업 보안 책임자에게 각인된 질문은 단순하다. "미국 상무부나 클라우드 사업자가 서한 한 장으로 우리 핵심 AI 인프라를 끌 수 있는가?" 이번 사건은 그 질문에 "그렇다"라고 답했다.

이 순간 Private AI로의 전환은 ROI의 문제가 아니라 업무 연속성의 문제로 격상된다. 기업 입장에서 public API는 더 이상 단순 외부 서비스가 아니다. 핵심 업무 자동화, 코드 작성, 고객 응대, 내부 지식 검색, 보안 분석, 재무·법무 워크플로우에 연결된 운영 인프라다. 이 인프라가 정책 서한 한 장으로 꺼질 수 있다면, public API는 편리한 파트너이면서 동시에 단일 장애점이다.

(2) Capability race와 public serving의 분리

2023년 이후 엔터프라이즈 AI의 제1원칙은 "가장 뛰어난 프론티어 API를 호출해 애플리케이션을 만든다"였다. Fable/Mythos는 이 전제를 깨뜨렸다. 모델 제공자는 더 이상 단순한 파트너가 아니라 잠재적 SPOF이며, 경우에 따라 고객의 사용 패턴과 업무 데이터를 관찰하는 블랙박스로 간주된다.

이 변화는 구조적이다. 통제축이 반도체에서 모델 접근권으로 이동했다는 것은, 앞으로 어떤 기업의 최상위 모델이라도 국가 안보를 명분으로 제한될 수 있다는 뜻이다. 이 선례의 사정거리는 Anthropic 하나에 그치지 않는다.

이 점은, 통제 대상 기업 자신이 인정한다. Anthropic은 공식 성명에서 문제가 된 jailbreak 기법이 OpenAI의 GPT-5.5를 비롯한 다른 공개 모델에서도 동일하게 작동한다고 밝혔다. 가장 강력한 안전장치를 갖췄다고 주장해 온 회사조차 '이건 우리만의 취약점이 아니다'라고 말한 셈이며, 이는 같은 통제 논리가 OpenAI·구글·xAI의 최상위 모델로 번질 수 있음을 직접 시사한다.

시장이 이를 일회성 사고로 가격에 반영하는 동안, 기업 현장은 반복될 수 있는 위험으로 보고 아키텍처를 다시 짜기 시작한다. 이 인식의 시차가 투자 기회의 근원이다.

3. 자체 운영은 왜 경제적으로 불리한가

Private AI로 수요가 이동한다면, 다음 질문은 “프라이빗을 어떻게 구현할 것인가”다. 이경우 보통 “그러면 기업이 GPU를 직접 사서 돌리면 되지 않나”라는 직관이 떠오르지만, 그러나 이 직관은 운영 책임과 가동률 경제학이라는 두 벽에 부딪힌다.

(1) 정성적 측면: 자체 운영은 '운영'에서 진다

대기업 경영진은 "우리도 자체 AI 스택을 갖자"고 말할 수 있지만, 현장에서 모델 서빙, GPU 스케줄링, 권한 제어, 검색 증강(RAG) 보안, 평가셋 설계, 성능 변동 감시, 에이전트 도구 권한 관리, 감사 로그, 비용 통제를 모두 안정적으로 운영할 팀은 드물다. 에이전트는 단순한 챗봇이 아니라 사내 시스템에 행동 권한을 갖는 구조이기 때문에, 운영 난이도가 기존 SaaS보다 훨씬 높다.

게다가 오픈 가중치 모델을 직접 운영한다는 것은, 사고가 발생했을 때 책임 소재가 기업 자신에게 귀속된다는 뜻이다. 적지 않은 보안 책임자와 법무팀이, 차라리 벤더가 보안과 컴플라이언스, 감사 체계를 제공하는 관리형 환경을 선호하는 이유가 여기에 있다.

(2) 정량적 측면: 가동률 경제학

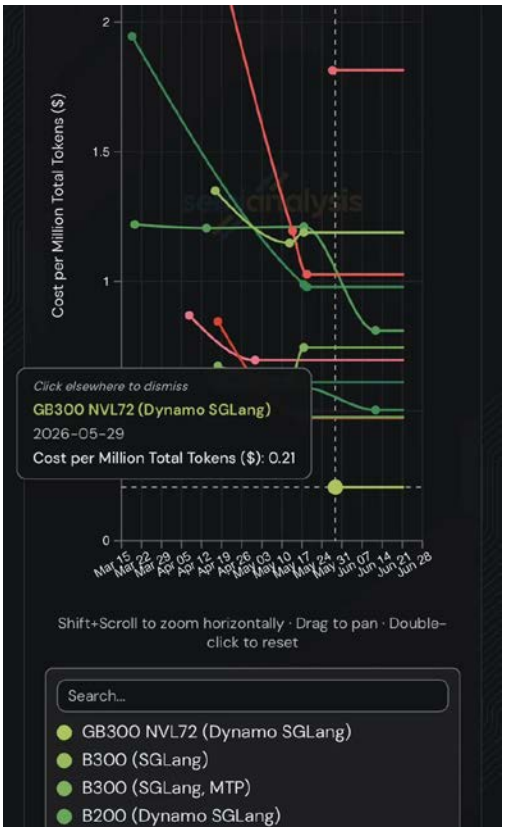
위와 같은 정성적 근거만으로는 약하다. 결정적인 근거는 결국 가격이다. SemiAnalysis의 연구원 Jordan Nanos가 공개한 추론 비용 곡선은, 기업이 자체 운영을 해야 한다는 주장을 숫자 하나로 무너뜨린다.

그림 1. 외부 추론 서비스(엔드포인트) 업체들의 토큰 판매 가격표
GLM-5.2 토큰 100만 개당 입력 약 1.40달러·출력 약 4.40달러
어디서 사도 비슷하게 비싸다, "차라리 직접 장비를 사서 돌릴까?”

Provider	Input /M	Output /M
Wafer	$ 1.20	$ 4.10
DeepInfra	$ 1.20	$ 4.20
Fireworks	$ 1.40	$ 4.40
Ambient	$ 1.40	$ 4.40
Z.ai	$ 1.40	$ 4.40
Friendli	$ 1.40	$ 4.40
NovitaAI	$ 1.40	$ 4.40
StreamLake	$ 1.40	$ 4.40
Cloudflare	$ 1.40	$ 4.40
AtlasCloud	$ 1.40	$ 4.40
Parasail	$ 1.40	$ 4.40
Phala	$ 1.40	$ 4.40
Together	$ 1.40	$ 4.40
io.net	$ 1.68	$ 5.28

자료: Reddit r/OpenModels, 미래에셋증권 리서치센터

그림 2. 엔비디아 대형 시스템에서 만든 토큰의 실제 원가 추이
GB300 NVL72에서 대규모로 찍어내면 100만 개당 0.21달러에 불과
그림 1의 판매가 1.73달러(평균)와 비교하면 약 8배 싼 셈



자료: Nvidia, SemiAnalysis, 미래에셋증권 리서치센터

위 그림에서 핵심 수치는 단순하다. 현재 최고의 오픈소스 모델인 GLM-5.2에서 같은 작업을 처리할 때, 엔비디아의 최신 대형 시스템인 GB300 NVL72에서 만들어내는 토큰의 원가는 “100만 개당 0.21달러”다. 반면 외부 추론 서비스(엔드포인트) 업체들이 받는 평균 판매가는 1.73달러로, 약 8배나 비싸다. 여기까지만 보면 "외부에서 사 쓰는 게 이렇게 비싸니, 차라리 우리가 직접 장비를 사서 돌리자"는 결론이 자연스러워 보인다.

그러나 이 차트의 진짜 메시지는 정반대다. 0.21달러라는 낮은 원가는 그냥 좋은 장비를 샀다고 나오는 숫자가 아니다. 세 가지 조건이 동시에 맞아떨어져야 나온다. 첫째, 72장의 GPU를 한 덩어리로 묶어 초고속으로 연결하는 랙 단위 구조다. 둘째, 수많은 고객의 작업을 한데 모아 장비를 거의 쉴 틈 없이 돌리는 높은 가동률이다. 셋째, 그 장비를 한계까지 쥐어짜는 추론 소프트웨어 최적화다. 이 세 가지는 모두 '대규모로 운영하는 사업자'에게서만 성립한다.

즉, 여기에 함정이 있다. 어느 기업이 NVL72 한두 랙을 직접 사서 자기 회사 업무만 돌린다고 해보자. 장비는 똑같지만, 한 회사의 업무량만으로는 그 거대한 장비를 가득 채우지 못한다. 낮에는 몰리고 밤에는 놀고, 바쁠 때는 모자라고 한가할 때는 텅 빈다. 가동률이 떨어지면 토큰 한 개를 만드는 데 드는 원가는 0.21달러가 아니라 1~2달러로 빠르게 치솟는다. 그 순간, 차라리 외부에서 1.73달러에 사 쓰는 편이 더 싸진다. 즉 장비를 직접 소유하는 순간, 0.21달러의 경제성은 사라진다.

그래서 기업이 거대한 인프라를 직접 사들이는 것이 아니라, 전용 가상 클라우드나 전용 용량, 소버린 AI 클라우드 같은 형태로 격리된 자기 구획만 빌려 쓰는 구조가 합리적이다. 인프라를 가득 채워 돌리는 힘든 일은 규모가 큰 클라우드 사업자와 GPU 클라우드에게 맡겨 0.21달러의 원가를 누리되, 기업은 그 위에서 자기 데이터 격리와 권한 통제, 평가 루프, 감사 체계만 따로 챙기는 것이다. 이것이 관리형 Private AI가 경제적으로도 가장 자연스러운 답인 이유다.

(3) 온프레미스는 비용이 아니라 보험이다

그렇다면 회사 안에 직접 장비를 들이는 온프레미스는 설 자리가 없는 걸까? 그렇지 않다. 다만 그 자리가 '비용'이 아니라는 점이 중요하다. 토큰 원가 0.21달러를 다투는 가격 경쟁에서 단독 온프레미스는 이길 수 없다. 온프레미스가 살아남는 이유는 싸기 때문이 아니라, 돈으로 환산할 수 없는 다른 이유 때문이다.

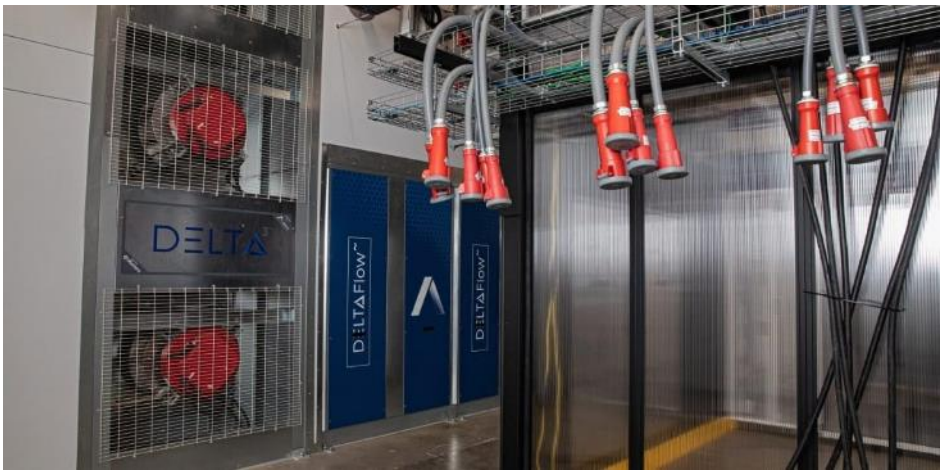
세 가지 경우다. 데이터가 법적으로 절대 회사 밖으로 나갈 수 없는 규제 산업, 응답이 0.01초만 늦어도 안 되거나, 데이터가 너무 무거워 외부로 보내기 어려운 현장, 그리고 외부 서비스가 어느 날 갑자기 끊겨서는 안 되는 업무 연속성이 생명인 곳이다.

국방, 1급 금융기관, 각 나라 정부, 의료기관 같은 고객이 여기에 해당한다. 이들은 토큰이 비싸다는 걸 알면서도 장비를 들인다. 싸게 쓰려고 사는 게 아니라, 통제권을 사는 것이기 때문이다. 이것은 비용 절감이 아니라 일종의 보험에 가깝다.

그림 3. 일라이릴리: 2025년 10월 제약회사로는 세계 최대 규모 온프레미스 AI 슈퍼컴퓨터를 구축
1,016개 Blackwell Ultra GPU (B300) 탑재, Lilly가 직접 소유·운영

자료: 일라이 릴리, Nvidia, 미래에셋증권 리서치센터

그림 4. 제인 스트리트: 100~200MW 규모의 자체 데이터센터 건설 계획 (자금 조달 포함)
기존 AI 랩을 소규모 Dell 서버 → 대규모 liquid-cooled GPU 클러스터로 업그레이드
퀀트 트레이딩 + AI 모델 학습/추론을 위한 온프레미스 인프라에 적극 투자



자료: Jane Street, 미래에셋증권 리서치센터

실제 제품을 봐도 이 점이 드러난다. 예를 들어, HPE의 Private AI 제품인 Private Cloud AI의 대형 구성은 H200 GPU 16장(확장해도 64장) 수준이다. 앞서 0.21달러를 만든 NVL72의 72장 랙과는 애초에 체급이 다르다. 즉 서버 제조사가 파는 이런 장비는 NVL72급의 원가 경쟁력을 흉내 내려는 물건이 아니라, 규제와 지연 시간, 데이터 주권이라는 프리미엄을 받고 파는 '격리된 안전 장비'인 것이다.

정리하면, 관리형 프라이빗은 규모를 빌려 그 경제성을 누리며, 물리적 온프레미스는 규제·주권이 필요한 곳에만 선별적으로 들어간다. 그렇다면 수요는 실제로 어느 기업의 매출로 흘러가는지를 생각해보면 된다.

4. 밸류체인은 어디로 재편되는가

모델 API 위험 증가 → 관리형 프라이빗·전용 클라우드 확산 → 기업별 추론 서버 수요의 순증 → 엔터프라이즈 보안 네트워크 수요 → 데이터 권한·평가·검수 소프트웨어 수요.

표 1. 수혜 지도

계층	수혜 후보	성격	핵심 논거
프라이빗 추론 서버	Dell, HPE, Supermicro·Lenovo·ODM	비용+주권	대형 학습 클러스터가 아닌 추론·검색 증강·미세조정 어플라이언스. 기존 IT 예산 위에 얹히는 설비 투자의 순증. Dell AI Factory 혹은 HPE Private Cloud AI가 제품화된 형태
엔터프라이즈 보안 네트워크	Cisco, HPE-Juniper	주권	대형 클라우드 패브릭이 아닌 기업 내부 AI 네트워크. 에이전트가 ERP·CRM에 접근하므로 대역폭이 아니라 신원 확인, 구간 분리, 제로 트러스트, 감시 가능성이 핵심. 기존 고객 기반과 영업 채널이 해자
AI 운영·거버넌스	Palantir, Microsoft, ServiceNow	주권	모델이 아니라, 모델을 기업의 데이터·권한·업무 객체·감사 체계에 묶는 계층. 온톨로지와 AIP가 인간 검수, 감사 추적, 모델 교체를 통합 관리
관리형 프라이빗 클라우드	Azure Foundry, AWS Bedrock, Vertex	비용+주권	'모델 제공자와 고객 사이의 신뢰 경계'를 장악. 0.21달러의 규모 경제를 공유 인프라로 임대. 컴퓨팅 종속의 완성
데이터 거버넌스·평가	Databricks, Snowflake	주권	기업별 평가, 성능 변동 감시, 감사. 관리형 프라이빗 위에서 더 빠르게 확산
대형 클라우드 AI 패브릭	Arista, NVIDIA, Broadcom	(별개 시장)	대형 AI 데이터센터 백엔드. Cisco와는 다른 전장이라는 점을 분리해서 인식해야 함

자료: 미래에셋증권 리서치센터

(1) 프라이빗 추론 서버 – Dell

Dell은 'Private 추론' 설비 투자의 순증 수혜다. 다만 이들을 0.21달러 비용 게임의 승자로 보면 안 된다. 이들의 핵심 포지션은 대형 공유 인프라 운영자가 아니라, 기업 내부에 들어가는 주권 프리미엄 장비 공급자다.

그림 5. 젠슨 황이 싸인하고 있는 랙은 엔터프라이즈(기업용) 온프레미스 데이터센터 겨냥한 시스템
Dell 행사의 주제가 "AI를 빅테크 전유물에서 벗어나, 모든 기업의 데이터센터로 가져오겠다"는 것
싸인하고 있는 GB200 NVL72 랙은 현존 최고의 'AI 학습(Training)' 머신이기도 하지만,
젠슨 황은 이 랙을 철저하게 'Agentic AI(자율 에이전트 AI)'의 구동 및 Inference 측면만을 강조

자료: DELL, 미래에셋증권 리서치센터

여기서 중요한 것은 이 시장이 기존에 없던 '순증' 수요라는 점이다. 모든 기업이 빅테크처럼 수만 장짜리 GPU 학습 클러스터를 사는 것이 아니다. 그 대신, 지금까지 존재하지 않던 새로운 종류의 장비 수요가 생긴다. 사내 문서를 검색해 답하는 추론·검색 서버, 회사 데이터로 모델을 다듬는 미세조정 서버, 공장이나 매장에 놓이는 현장용 AI 서버 같은 것들이다. 이는 빅테크의 클라우드 설비 투자를 대체하는 것이 아니라, 기업들이 원래 쓰던 IT 예산 위에 새로 얹히는 수요다. Dell이 'AI Factory'라는 이름으로 통째로 묶어 파는 제품이 바로 이것이다.

투자 논리는 "기업이 비용을 낮추려고 서버를 산다"가 아니라, "기업이 외부 API 차단, 데이터 반출, 지연 시간 등을 피하려고 일정 규모의 추론 설비를 회사 안에 들인다"가 맞다. 이 차이를 분리해서 봐야 Dell의 수혜를 과대평가하지 않으면서도, 구조적으로 새로 생기는 순증 수요는 놓치지 않을 수 있다.

(2) 엔터프라이즈 보안 네트워크 – Cisco

서버를 들였으면 그 서버가 사내 시스템과 안전하게 연결되어야 한다. 이 길목에 Cisco가 있다. 단, Cisco가 강한 시장은 빅테크의 초대형 AI 데이터센터 네트워크가 아니라, 기업 내부의 분산된 AI 네트워크다. 이 둘은 전혀 다른 시장이다.

왜 보안이 핵심이 되는지는 에이전트가 하는 일을 보면 분명해진다. AI 챗봇은 그저 질문에 답만 하지만 업무용 에이전트는 회사의 ERP(전사 자원 관리)와 CRM(고객 관리) 시스템에 직접 접속하고, 사내 데이터 창고와 코드 저장소, 문서 시스템을 뒤지며, 때로는 실제 업무를 대신 실행한다. 즉 에이전트는 사내 시스템을 자율적으로 돌아다니는 존재다. 이때 회사가 걱정해야 할 것은 '네트워크가 얼마나 빠른가'가 아니라, '이 AI가 누구 권한으로 어디까지 들어갈 수 있는가', '들어가면 안 되는 구역은 막혀 있는가', '무슨 행동을 했는지 나중에 추적할 수 있는가'다.

Cisco의 강점은 바로 여기서 나온다. 첫째, 금융사·병원·제조사·공공기관 같은 기존 고객사의 사내 네트워크에는 이미 Cisco 장비와 운영 체계가 깊이 들어가 있다. 회사가 AI 네트워크를 새로 깔 때, 보안팀과 네트워크팀은 가장 혁신적인 신생 장비보다 이미 다룰 줄 아는 장비를 택한다. 둘째, Cisco는 네트워크 장비 안에 보안 기능을 함께 묶어 판다. AI가 데이터를 주고받는 '길'과 그 길을 통제하는 '검문소'를 한 묶음으로 제공하는 것이다. 빠른 속도가 생명인 빅테크 백엔드 네트워크는 Arista·NVIDIA·Broadcom의 영역이지만, 신원 확인·구간 분리·제로 트러스트·행동 추적이 생명인 기업 내부 네트워크는 Cisco의 영역이라 할 수 있다.

(3) AI 운영·거버넌스 – Palantir

서버를 들이고 네트워크를 깔았어도, 가장 어려운 문제가 남는다. "이 AI가 회사의 어떤 데이터를 보고, 어떤 판단을 내리고, 그 판단을 누가 승인하며, 나중에 책임을 어떻게 추적하는가"다. 이 문제를 정면으로 겨냥하는 것이 Palantir다. Palantir는 모델을 만드는 회사가 아니라, 모델을 회사의 업무 체계 안에 안전하게 묶어 넣는 '운영·통제 계층'이다.

프론티어 모델의 신뢰가 흔들릴수록, 기업의 관심사는 "어떤 모델이 가장 똑똑한가"에서 " 모델을 어떻게 안전하게 교체하고, 감시하고, 승인하고, 업무 시스템에 연결하는가"로 옮겨 갈 수밖에 없다고 생각한다.

여기서 결정적인 것은, 밑단의 모델은 6개월마다 갈아탈 수 있지만 그 위에 쌓인 업무 권한 구조, 감사 로그, 검수 절차, 승인 체계, 평가 기준은 계속 남는다는 점이다. 장기적으로 가치가 복리로 쌓이는 곳은 모델을 호출하는 부분이 아니라, 바로 이 운영 루프다.

Palantir의 온톨로지와 AIP가 겨냥하는 자리가 여기이며, Microsoft와 ServiceNow 등도 각자의 업무 영역에서 이와 같은 운영의 계층을 노린다. 최근 보안·소프트웨어 벤더들이 입을 모아 "진짜 AI의 가치는 모델의 기본 성능이 아니라 맥락·기억·안전장치·검수 계층에 있다"고 주장하는 것도 정확히 같은 방향을 가리킨다. 현재로서는 Palantir의 온톨로지가 그 계층에 가장 확률 높은 승자로 보인다.

(4) 관리형 프라이빗 클라우드 – CSP

마지막으로, 이 모든 흐름의 밑바탕에 클라우드 3사(AWS·Azure·구글)가 있다. 이들은 기업이 GPU를 직접 소유하지 않으면서도 데이터 격리, 전용 용량, 권한 통제, 감사 가능성, 모델 선택권을 확보하게 해주는 관리형 프라이빗 환경을 판다. 앞서 본 “0.21달러”의 규모 경제를 직접 복제하지 못하는 기업에게, 공유 인프라 위에서 그 경제성의 일부를 빌려주는 역할이다. 비용 논리와 소버린 AI 논리가 교차하는 지점이라 구조적으로 유리하지만, 이 사태가 동시에 드러낸 '서한 한 장에 의한 차단 위험'의 당사자이기도 하다. 이러한 점에서, 앞의 세 갈래(Dell·Cisco·Palantir)와는 결이 다르다.

(5) 최종 변수: 업무 지식 자산의 귀속

이 수혜 지도 전체의 운명은 결국 하나의 질문으로 수렴한다. 기업이 AI를 쓰면서 쌓이는 노하우가 최종적으로 누구 손에 남느냐다. 여기서 '노하우'란 단순한 데이터가 아니다.

예를 들어 어느 증권사가 리서치 업무에 AI를 도입했다고 하자. 1년을 쓰는 동안 "이런 보고서는 좋고 저런 보고서는 나쁘다"는 평가 기준이 쌓이고, 애널리스트가 AI 답변을 고친 수정 이력이 축적되며, "이 판단은 반드시 사람이 한 번 더 확인한다"는 검수 절차와 사내 승인 흐름이 모델 위에 얹힌다. 이렇게 쌓인 업무 노하우야말로 그 회사가 경쟁사 대비 AI를 더 잘 쓰게 만드는 진짜 자산이다. 모델 자체는 6개월이면 더 좋은 것으로 갈아탈 수 있지만, 이 노하우는 갈아탈 수 없다.

문제는 이 노하우가 어디에 저장되느냐에 따라 돈을 버는 주체가 완전히 달라진다는 점이다. 모델 기업(OpenAI, Anthropic 등)의 시스템 안에 쌓이면 모델 기업이 고객을 묶어두게 되고, 특정 소프트웨어 벤더의 폐쇄적인 제품 안에 갇히면 그 벤더가 협상력을 쥐며, 클라우드 사업자의 경계 안에서만 작동하면 클라우드 사업자가 주도권을 갖는다. 반대로 이 노하우가 기업 자신의 자산으로 남으면, 기업은 밑단 모델을 자유롭게 바꿔 끼우면서도 협상력을 유지한다.

우리의 판단은 마지막 쪽에 가깝다. 진짜 주도권은 '어떤 모델을 쓰느냐'가 아니라 '내 회사의 평가 기준과 수정 이력, 업무 기억을 누가 쥐고 있느냐'에서 나오기 때문이다.

이 판단이 맞다면 돈은 모델을 직접 만드는 회사보다, 모델을 갈아 끼워도 기업의 노하우를 그대로 지켜주는 '운영·통제 계층'으로 흐른다. 앞서 표에서 짚은 Palantir 같은 운영·거버넌스 사업자가 여기에 해당한다.

그래서 앞으로 추적해야 할 신호는 모델 성능 점수표가 아니다. "기업이 AI로 쌓은 업무 노하우를, 모델을 바꿔도 그대로 들고 나갈 수 있는가" 하나다. 기업이 그 노하우를 자기 것으로 소유하는 방향으로 제품과 계약이 진화하면 우리가 만든 수혜 지도가 강해지고, 벤더가 그것을 자기 플랫폼에 흡수하는 방향이면 반대로 기운다.

(6) 투자 결론

구조적 선호는 Dell, Cisco, Palantir, 관리형 클라우드다. Dell은 비용 우위 장비가 아니라 주권 프리미엄 추론 설비 공급자다. Cisco는 hyperscaler 백엔드가 아니라 엔터프라이즈 AI 보안 네트워크 수혜다. Palantir는 모델 선택이 아니라 모델 운영, 검수, 교체, 감사 계층의 수혜다. 관리형 클라우드는 기업이 규모의 경제와 소버린 AI 요구를 동시에 조달하는 사업자라 할 수 있다.

표 2. 자본 배분 관점

스탠스	대상	논거
구조적 선호	Dell / Cisco / Palantir, 관리형 클라우드(CSP)	프라이빗 추론 설비 투자의 순증, 보안 네트워크, 운영·거버넌스 계층. 'API 위험 → 관리형 프라이빗'으로 이어지는 흐름의 직접 수혜
조건부 선호	주권 AI 하드웨어, 데이터 거버넌스·평가 소프트웨어	규제와 주권 수요의 확산 효과. 성공 시 탄력은 크지만 고객 채택의 검증이 더 필요
재평가 주의	인프라를 보유하지 못한 독립 프론티어 기업	밸류에이션 재조정 위험. 단 이 리스크는 양극단(bimodal) 하방은 kill-switch·접근통제·마진 압박, 상방은 정부가 capex 엔진이자 초대형 고객으로 전환되는 국유화 시나리오.
함정	'오픈소스 온프레미스 자체 운영'에 대한 순수 수혜 기대	가동률 경제학상 죽음의 계곡. 온프레미스는 비용이 아니라 주권 프리미엄으로만 정당화된다

자료: 미래에셋증권 리서치센터

재평가 주의 대상은 인프라를 직접 쥐지 못한 독립 프론티어 기업이다. 이번 사태는 그 위험을 압축해 보여줬다. Anthropic은 5월 28일 Series H에서 $965B 밸류로 OpenAI($852B)를 제치고 세계 최고가 AI 랩이 됐는데, 불과 2주 뒤 정부가 그 회사의 최상위 모델을 꺼버렸다. 밸류에이션이 정책 리스크에 재조정될 수 있다는 명제의 살아있는 사례다. 동시에 이 리스크는 양방향이며, 새로 열린 차별화 축은 '정부와의 관계'다. 워싱턴을 더 매끄럽게 다루는 랩이 규제 리스크 프리미엄을 덜 얹게 되는 구도다.

* Anthropic은 Fable 5 재출시 협상 창구를 Dario Amodei(CEO직은 유지)에서 공동창업자 Tom Brown으로 교체했고, 정부 측 기류가 누그러진 것으로 전해진다.

반대로 주의해야 할 함정은 “오픈소스 온프레미스 자체 운영”에 대한 순수 수혜 기대다. 기업이 모델 접근 위험을 느낀다고 해서 곧장 GPU를 직접 사서 운영하는 것은 아니다. 운영 책임과 가동률 경제학이 이를 어렵게 만든다. 물리적 온프레미스는 비용을 낮추는 수단이 아니라 통제권과 업무 연속성을 사는 보험이다.

5. 마무리

지금 벌어지는 일의 본질은 'AI가 위험해졌다'가 아니라, '지능의 향상 속도를 사회의 제도가 따라가지 못하기 시작했다'는 것이다.

Fable/Mythos 사태는 그 어긋남이 처음으로 표면에 드러난 사건이다. 모델의 능력은 분기 단위로 계단을 오르는데, 그것을 통제할 수출 제도, 국가 안보 체계, 기업의 거버넌스, 노동 시장의 적응 속도는 연(年) 단위로 움직인다. 이 속도 차이가 임계점을 넘는 순간, 국가는 가장 거친 방식으로 제동을 건다. 능력을 멈추는 것이 아니라, 능력이 사회에 도달하는 '시점'을 강제로 늦추는 것이다.

AI 모델의 대중 공개는 얼어붙지만 내부의 능력 경쟁은 계속된다. 우리가 본 것은 발전의 정지가 아니라, 발전과 공개 사이의 시차가 처음으로 벌어진 장면이다. 따라서 우리가 강조하고 싶은 것은 종목 몇 개 보다는, 상황이 바뀌고 있다는 신호 그 자체를 먼저 알아채야 한다는 것이다.

셧다운된 것은 모델이지 수요가 아니며, 멈춘 것은 '모델 공개'이지 '지능 발전'이 아니다. 지능이 사회의 소화 능력을 앞지르기 시작한 이 국면에서, 시장은 잠시 발전이 멈춘 것처럼 착각할 수 있다. 그러나 그 착각의 이면에서 수요는 사라지지 않고 조용히 자리를 옮기고 있다. 변화의 방향을 먼저 읽는 투자자가, 그 이동이 만들어낼 다음 국면의 과실을 얻게 되지 않을까 생각한다.

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트는 자료작성일 현재 조사분석 대상법인의 금융투자상품 및 권리를 보유하고 있지 않습니다.
- 본 자료는 외부의 부당한 압력이나 간섭없이 애널리스트의 의견이 정확하게 반영되었음을 확인합니다.

본 조사분석자료는 당사의 리서치센터가 신뢰할 수 있는 자료 및 정보로부터 얻은 것이나, 당사가 그 정확성이나 완전성을 보장할 수 없으므로 투자자 자신의 판단과 책임하에 종목 선택이나 투자시기에 대한 최종 결정을 하시기 바랍니다. 따라서 본 조사분석자료는 어떠한 경우에도 고객의 증권투자 결과에 대한 법적 책임소재의 증빙자료로 사용될 수 없습니다. 본 조사분석자료의 지적재산권은 당사에 있으므로 당사의 허락 없이 무단 복제 및 배포할 수 없습니다.