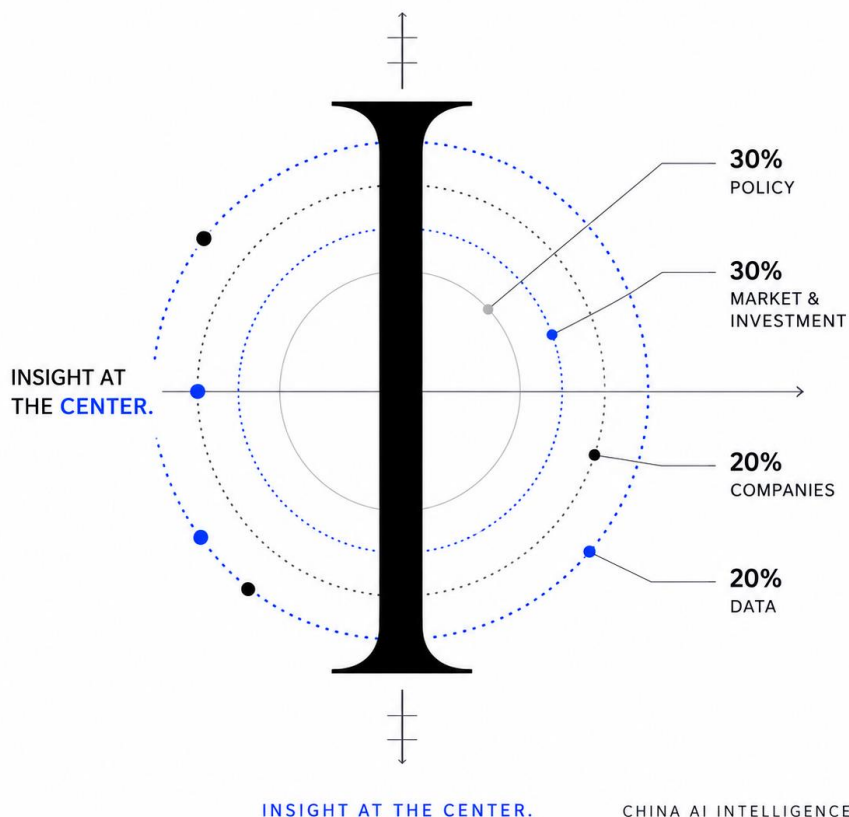


중국 AI·에이전트 심층 리포트

모델을 넘어 업무를 대체한다 — 모델은 상품이 되고, 가치는 하네스로 이동한다



INSIGHT AT THE CENTER.

CHINA AI INTELLIGENCE

에디션 월간(Monthly) — 창간호

발행일 2026년 09월 08일

발행인 차이나랩 CHINALAB

수신자 SI·SaaS·플랫폼 전략팀, 기업 IT·DX 담당, 증권사 리서치, 정책 입안자

환율 USD/KRW 1342.6 · CNY/KRW 200.6 · HKD/KRW 171.26 (2026-09-08 한국은행
매매기준율(ECOS))

목차

실행 요약	3
창간 지도 — 에이전트 스택 4 단은 지금 어디에 있는가	6
스택 1 단 — 모델이 상품이 되는 구간	9
헤드 목록의 혼재와 '충분히 좋은' 계층의 형성	9
스택 2 단 — 토큰 경제, 가격이 1 달러 아래로 내려간 뒤	11
스택 3 단 ① — 하네스, 가치가 이동한 계층	13
스택 3 단 ② — 플랫폼 전쟁, 왜 오피스가 전장이 되었나	15
스택 3 단 ③ — 프로토콜과 표준, 중국이 만드는 두 번째 규격	18
스택 4 단 ① — 금융, 입찰에서 결제까지	21
스택 4 단 ② — 제조와 정부, 결정성 요구와 조달 구조	24
정책·규제 — 보급률 목표와 행위 규제가 동시에 간다	26
한국 시사점·액션 플랜	27
부록 — 주요 지표와 출처	29

실행 요약

결론. 토큰 가격이 무너져 모델이 상품이 되었고, 가치는 모델을 실제 업무에 붙이는 계층 — 하네스(Harness)와 오케스트레이션 — 으로 이동했다. 중국은 이 계층을 오픈소스 공개, 국가 표준·인증, 정부·금융 조달이라는 세 갈래로 동시에 굳히고 있다. 한국의 과제는 모델 자국산화가 아니라 하네스 계층의 자산화다. 업무 온톨로지, 연결 규격, 행위 감사를 기업 자산으로 축적하는 것이 이번 호가 제시하는 핵심 방향이다.

1. 핵심 인사이트 5 개

- ① 모델은 상품이 됐다 — 백만 토큰당 가격이 2.04 달러(2026 년 5 월 말)에서 0.97 달러(2026 년 9 월 초)로 떨어져 사상 처음 1 달러 아래로 내렸고(실리콘데이터 토큰 지수, 21 세기경제보도 2026-09-03/04), 즈푸 AI 는 단위 토큰 추론 비용을 연초 대비 80% 낮추며 10 만 장급 자국산 칩 추론을 실현했다(경제참고망 2026-09-01). 같은 기간 화산엔진의 일평균 토큰 호출량은 2024 년 5 월 1,200 억에서 2026 년 6 월 180 조로 폭증했다(증권시보 2026-06-24). 가격 붕괴와 수요 폭증이 동시에 일어나는 것이 상품화의 정의다.
- ② 차이는 모델이 아니라 하네스에서 난다 — 같은 모델에 하네스만 바꿨을 때 평가 점수 평균 차이가 15.6%, 일부 영역은 48%에 달하고 "모델이 강할수록 하네스 기여 가치 비중은 오히려 커진다"는 것이 업계 분석이다(자쯔광넨 2026-08-20, 벤치마크 원자료 미공개).
- ③ 모델 회사가 하네스 회사로 확장한다 — 딥시크는 내부 정의를 "Model + Harness = Agent"로 못 박고 하네스 전담 팀을 꾸린 뒤, 딥시크 하네스(DeepSeek Harness) 개발자 프리뷰를 2026 년 8 월 13 일 공개했다(평파이신문 2026-08-15). 설계 특성은 '제어 가능·관측 가능·추적 가능'이다.
- ④ 중국은 하네스 계층을 표준·인증으로 굳힌다 — CAICT 의 「가신(可信) 인터넷 에이전트 운영체제 — 에이전트 실행 프레임워크」 1 차 평가 통과자가 국유 통신 2 사(렌통 데이터지능·차이나텔레콤 베이징연구원)였고(평파이신문 2026-07-30), 구이저우 정무 에이전트 「구이런즈반」은 생성형 AI 등록을 마친 국내 첫 정무 종합 에이전트로 조달에 들어갔다(양시망 2026-08-28). 국가가 표준을 만들고 국유 기업이 먼저 인증을 받는 구조다.
- ⑤ 조달이 수요를 만든다 — 금융 에이전트 입찰 유효 표본 97 건(2026-01-01~06-15, 링이즈쿠·치예위징통) 중 은행이 48.5%, 플랫폼 기반 수요가 28.9%이고, 공상은행은 상반기 대모델 600 여 개 시나리오에서 컴플라이언스 심사의견 36 만 건을 산출해 채택률 98.6%를 기록했다(중국경영보 2026-09-05). 국무원 국발〔2025〕11 호(2025-08-21)는 2027 년 차세대 지능형 단말·에이전트 등 보급률 70%를 국가 목표로 걸었다.

2. 한국 기업에 가장 중요한 변화 3 가지

- ① 모델 자체 개발의 경제성이 무너졌다. 즈푸 AI 의 GLM-5.3-Flash(총 320B·활성 18B)는 공개 즉시 오픈소스로 풀려 753B 규모 GLM-5.2 를 전면 상회했고, 62 조 토큰의 훈련이 자국산 카드에서 돌았다(양자위 2026-08-27). 한국 기업에게 필요한 것은 모델을 만들 이유가 아니라, 오픈소스 모델 위에 올릴 하네스를 소유할 이유다.

② 규격 종속 위험이 생겼다. 중국은 에이전트 통신 프로토콜(ANP·ACP·RVP), 에이전트 기초 기술·서비스 표준(계획번호 2024-1332T-YD 등 2 건), 결제 자율규약의 '당신의 에이전트를 알라(KYA)'

메커니즘(중국지급청산협회, 2026-08-24)까지 연결 규격을 국가 단위로 묶고 있다. 청량(淸朗) 정비 행동에서는 에이전트가 규제 집행 대상 품목으로 처음 명시 열거돼 1.4 만여 개 AI 상품이 처리됐다(신화망 2026-07-06). 규격을 누가 쥐는가가 모델 성능보다 시장 진입을 좌우한다.

③ 기업 사용자가 하네스로 이동하고 있다. 알리바바 천원반궁은 공개 베타 1 개월 만에 사용자 3,000 만 명을 넘었고 절반이 기업 사용자다(양자위 2026-09-04). 중국 데스크톱 AI 오피스 월 방문량은 2026 년 3 월 2,000 만 회에서 6 월 6,000 만 회로 3 배 늘었으나, 기업이 AI 의 핵심 업무 의사결정 참여를 신뢰하는 비율은 23%에 그친다(타이메이티 2026-09-05). 신뢰가 다음 관문이며, 신뢰는 감사 가능한 하네스에서 나온다.

3. 위험·기회 2×2 매트릭스

시장 규모 통계는 정의가 143 배 차이 날 만큼 혼란스럽다 — 즈옌컨설팅의 2025 년 중국 AI 에이전트 시장 182.34 억 위안(2026-05-07) 대 프로스트앤설리번 전망 2.6 조 위안(36 커 2026-09-06 인용). 어느 쪽도 단독 인용할 수 없으므로 아래는 방향성만 다룬다.

	단기(6 개월)	중기(1~2 년)
위험	중국산 오픈소스 모델+하네스 묶음의 저가 공세로 에이전트 도입 단가 붕괴(토큰 지수 최근 7 일 -8.6%) — 국내 사업자의 가격 방어 실패	중국 표준·인증 체계가 아시아 조달의 사실 규격으로 굳어져 한국 기업이 '준수자' 위치로 밀림; 인증 통과자의 국가 편중 고착화
기회	모델 조달 비용 붕괴로 하네스 실험 비용 급감 — 자산화를 시작할 적기	업무 온톨로지·행위 감사 로그가 축적되면 모델 교체가 자유로워져 벤더 협상력 확보; 창안자동차처럼 와이어하니스 선정 계산 2 일→5 분 사례(신화망 2026-09-03)의 재현 가능

4. 실행 액션 아이템 5 개(부서별·시기별)

① 전략기획실(즉시~3 개월): AI 예산에서 모델 라이선스 비중을 낮추고 하네스·오케스트레이션 비중을 올리는 재편을 승인하라. 토큰 단가가 5 월에서 9 월 사이 절반으로 떨어진 것이 근거다.

② AI 개발 조직(3 개월 내): 오픈소스 모델 2 개 이상을 같은 하네스에서 교체 가능한 모델 무관 실행 구조를 만들고, 딥시크의 '제어 가능·관측 가능·추적 가능' 3 요소를 내부 하네스 요구사항으로 채택하라.

③ 업무 혁신 부서(6 개월 내): 창신치즈의 '온톨로지 + 하네스 엔지니어링 + AI 코딩' 패러다임(인민망 2026-09-01)을 참조해 핵심 업무 10 개의 온톨로지를 문서화하라. 모델이 아니라 이 문서가 자산이다.

④ 정보보호·감사 부서(6 개월 내): 에이전트 신원 확인·권한 협약·종료 절차를 사내 규정으로 먼저 만들어 행위 감사 체계를 세우라. 중국이 KYA 를 결제 규약으로 굳히기 전에 내부 통제부터 확보하는 것이다.

⑤ 대관·표준 담당(연내): CAICT 차기 에이전트 실행 프레임워크 평가(접수 마감 2026-10-30)와 MCP·A2A 갱신을 추적하고 국내 표준 논의에 의견을 개진하라. 규격은 시장이 아니라 위원회에서 만들어진단다.

5. 다음 호 전망

2026 년 10 월 30 일 마감되는 CAICT 차기 에이전트 실행 프레임워크 평가의 통과자 명단이, 하네스 계층의 '국가 인증 조달 지도'가 민간으로 확산되는지를 가르는 첫 신호가 될 것이다.

창간 지도 — 에이전트 스택 4 단은 지금 어디에 있는가

중국 에이전트 산업을 읽는 첫 지도는 네 개의 수평 계층이다. 모델, 컴퓨팅·토큰, 에이전트 플랫폼·미들웨어(하네스), 산업 응용이다. 이 네 단은 같은 산업 안에 있지만 서로 다른 속도로 움직이고, 서로 다른 것을 팔고, 서로 다른 논리로 수익을 만든다. 2026 년 9 월 현재 각 단이 어디에 있는가를 먼저 표로 고정한다.

계층	무엇을 파는 계층인가	대표 사업자	2026 년 관측된 변화	수익 형태	한국 접점
① 모델	범용 추론 능력	즈푸 AI(Zhipu AI, 智谱)·딥시크(DeepSeek, 深度求索)·문샷 AI·미니맥스(MiniMax, 稀宇科技)·커다윈페이(iFLYTEK, 科大讯飞)	토큰 단가 붕괴, 오픈소스 확산, 홍콩 상장 개시	API 호출 과금·오픈소스 생태계	모델 조달의 선택지가 될 뿐 차별화 자산이 아니다
② 컴퓨팅·토큰	지능 컴퓨팅과 토큰 공급	화산엔진(Volcano Engine, 火山引擎)·알리바바 바이런(Bailian, 百炼)·무어스레드(Moore Threads, 摩尔线程)	자국산 칩 추론의 양산 단계 진입	토큰 과금·MaaS 구독·정부 컴퓨팅 구매	자국산 칩 검증 결과가 한국 추론 원가의 참조점
③ 에이전트 플랫폼·미들웨어(하네스)	모델을 일하는 에이전트로 바꾸는 실행 계층	딥시크 하네스·텐센트 워크버드·창신치즈 AlInnoGC·차이나텔레콤 AgentOS	통신 규격·국가표준 인증 경쟁	플랫폼 라이선스·에이전트 애저 서비스	한국 기업이 자산화 해야 할 계층
④ 산업 응용	금융·제조·정부의 업무 성과	핑안은행·공상은행(ICBC)·창안자동차·우시시·구이저우성	조달과 입찰이 시장을 만든다	프로젝트 수주·성과 기반 계약	수요 검증의 사례집, 직접 진출 시장은 아니다

네 단의 속도는 서로 다르다. 모델 단은 붕괴의 속도다. 실리콘데이터(Silicon Data)의 대언어모델 토큰 지출 지수는 백만 토큰당 가격이 2026 년 5 월 말 2.04 달러에서 9 월 초 0.97 달러로 떨어져 사상 처음 1 달러 아래로 내려갔고, 최근 7 일간 하락률만 8.6%였다(2026-09-03·21 세기경제보도). 즈푸 AI 는 연내 토큰 호출량이 40 배 늘었지만 단위 추론 원가는 연초 대비 80% 하락했다(2026-09-01·경제참고망). 모델은 팔수록 싸지는 상품이 됐다.

컴퓨팅 단은 대체의 속도다. 즈푸 AI 는 이미 10 만 장급 자국산 칩의 규모화 저원가 추론을 실현했다고 밝혔고, GLM-5.3-Flash 는 62T 토큰 전량을 자국산 카드에서 돌렸다(2026-08-27·양자위). 취징커지(趋境科技)와 무어스레드의 토큰 파드(Token Pod)는 프리필 단계를 독립 공급·독립 과금·독립 최적화되는 자원 풀로 분리해 양산에 들어갔다(2026-09-04·광명망). 토큰이 과금 단위를 넘어 인프라 설계 단위가 된 것이다. 미들웨어 단은 규격 경쟁의 속도다. 해외의 MCP·A2A 외에 중국은 ANP 프로토콜, AgentUnion 의 ACP, 허촨테크의 RVP 같은 자체 통신 규격을 병렬로 키우고 있다(CAICT 보고, 2026 년 1 월). 딥시크는 「모델 + 하네스 = 에이전트」라는 내부 정의 아래 딥시크 하네스 개발자 프리뷰를 2026-08-13 공개했고, CAICT 는 「가신(可信) 인터넷 에이전트 운영체제 — 에이전트 실행 프레임워크」 평가를 시작해 첫 통과자로 렐통 데이터지능과 차이나텔레콤 베이징연구원이라는 국유 통신 2 사를 올렸다(2026-07-30·평파이신문). 국가가 표준을 만들고 국유 기업이 먼저 인증을 받는 구조다.

응용 단은 조달의 속도다. 링이즈쿠(零壹智库)가 치예위징통(企业预警通) 원자료로 집계한 금융 에이전트 입찰은 2026 년 상반기 유효 표본 97 건이었고, 수요의 28.9%가 플랫폼 기반(底座) 구축이었다(2026-06-24·21 세기경제보도). 정부에서는 구이저우의 구이런즈반(贵人智办)이 중앙사이버판공실 생성형 AI 등록을 통과한 국내 첫 정부서비스 종합 에이전트가 됐다(2026-08-28·양시망). 등록이 조달 자격으로 작동한다는 뜻이다.

마지막으로 이 지도 위에 산업 규모라는 숫자를 얹으면 지도가 흔들린다. 출처마다 추정치가 수십 배씩 갈리기 때문이다.

지표	값	출처·기준시점	성격
AI 핵심 산업 규모 2024	9,000 억 위안(+24%)	CAICT, 2026 년 1 월 보고	1 차·공식 연구기관
AI 핵심 산업 규모 2025(전망)	1.2 조 위안 초과	CAICT, 2026 년 1 월 보고	1 차·전망치
AI 에이전트 산업 규모 2025	182.34 억 위안(+78.03%)	즈옌컨설팅, 2026-05-07	2 차·민간 시장조사
AI 에이전트 시장 2025 → 2030	2.6 조 위안 → 20.1 조 위안	프로스트앤설리번(36 커 인용, 2026-09-06)	민간 전망치

같은 2025 년의 중국 에이전트 시장이 182.34 억 위안이면서 동시에 2.6 조 위안이다. 약 143 배의 차이다. 어느 한쪽이 틀린 것이 아니라 정의가 다르다 — 전자는 에이전트 소프트웨어의 매출을 재고, 후자는 에이전트가 관여하는 경제활동 전체를 잔다. CAICT 의 핵심 산업 규모는 그 사이 어디쯤의 세 번째 정의다. 이

리포트는 이 숫자들을 하나로 합치지 않는다. 합쳐지지 않는다는 사실 자체가 결론이다 — 중국 에이전트 산업은 아직 통계로 잡히지 않는 단계에 있고, 그래서 계층별 구조와 조달 사실이라는 실측값이 시장 규모 추정치보다 신뢰할 만한 나침반이다. 이후 각 절은 이 네 단을 위에서부터 순서대로 내려간다.

스택 1 단 — 모델이 상품이 되는 구간

헤드 목록의 존재와 '충분히 좋은' 계층의 형성

중국 헤드 모델의 능력 곡선은 이미 국제 프런티어와 같은 표에 놓였고, 이것이 상품화의 첫 번째 조건이다. 중국정보통신연구원(CAICT, 中国信息通信研究院)이 2026 년 1 월 발표한 「인공지능 산업발전 연구보고(2025 년)」는 팡성(方升) 벤치마크 기준으로 2025 년 12 월 현재 헤드 언어 대모델의 종합 능력이 2024 년 말 대비 30% 향상됐고, 다중모달 이해 능력은 50%를 넘었다고 측정했다(CAICT, 2026 년 1 월). 주목할 것은 이 보고가 '헤드 모델'로 실명 열거한 목록이다 — GPT-5.2, Gemini 3 Pro, Grok-4, Claude 4.5 와 함께 딥시크(DeepSeek, 深度求索) V3.2, Qwen3, 더우바오(Doubao, 豆包)-Seed-1.6, Kimi K2 Thinking, GLM-4.5 가 아무런 등급 구분 없이 존재해 있다. 국가 연구기관의 공식 문서에서 자국 모델과 미국 모델이 동일 헤드 집합으로 취급된다는 사실은, 구매자 입장에서 '어느 나라 모델이냐'가 더 이상 1 차 선택 변수가 아니라는 뜻이다. 대체 가능성이 생기는 순간 가격 결정권은 판매자에서 시장으로 넘어간다.

2026 년 신형 모델의 공통 특징은 에이전트 능력이 선택 사양이 아니라 기본 사양이 되었다는 점이고, 학습 기반까지 자국산 컴퓨팅으로 옮겨갔다는 점이다. 즈푸 AI(Zhipu AI, 智谱)가 2026 년 8 월 27 일 공개와 동시에 오픈소스로 푼 GLM-5.3-Flash 는 총 파라미터 320B 에 활성 파라미터 18B, 45 층 구조다 — 직전 GLM-4.5 의 92 층에서 절반으로 줄였다(양자위, 2026-08-27). 30T 토큰의 다중모달 사전학습 코퍼스를 썼고, 누적 62T 토큰의 학습 전량을 자국산 카드에서 돌렸다. AA 순위 57 점으로 Claude Opus 4.8 과 동점이며 753B 규모의 GLM-5.2 를 전면 상회한다. 커다쉴페이(iFLYTEK, 科大讯飞)의 싱휘(Spark, 星火) X2.5 는 MoE 구조 293B-A30B 로, 코드와 에이전트 능력을 중점 강화했고 훈련·추론 전 과정을 전량 자국산 컴퓨팅으로 완료했다(신랑과기, 2026-09-07). 딥시크는 2026 년 8 월 13 일 저녁 V4 Pro 정식판을 내놓으면서 같은 날 딥시크 하네스(DeepSeek Harness) 개발자 프리뷰를 함께 공개했다(평파이신문, 2026-08-15) — 모델만 팔지 않고 실행 계층을 세트로 내놓는 행보 자체가 모델 계층의 마진 한계를 인정한 것이다.

모델	기업	구조·파라미터	학습 기반	특이 사항
GLM-5.3-Flash	즈푸 AI	총 320B·활성 18B·45 층, 선형+회소 어텐션 혼합	62T 토큰 전량 자국산 카드	공개 즉시 오픈소스, AA 57 점(Claude Opus 4.8 동점)
싱휘 X2.5	커다쉴페이	MoE 293B-A30B	훈련·추론 전 과정 자국산 컴퓨팅	코드·에이전트 능력 중점, 온디바이스 X2.5-4B·1.7B 오픈소스, 100 만 토큰 컨텍스트
딥시크 V4 Pro	딥시크	확인된 사양 수치 없음	확인된 사양 없음	정식판과 하네스 개발자 프리뷰

				동시 공개(2026-08-13)
--	--	--	--	-------------------

오픈소스는 중국 모델 계층의 철학이 아니라 정량적으로 검증된 시장 점유 전략이다. CAICT 집계로 자국산 오픈소스 대모델의 글로벌 누적 다운로드는 2025 년 12 월 기준 100 억 회를 돌파했고, 월 다운로드 점유율은 최고 17.1%에 달했다(CAICT, 2026 년 1 월). 본토 커뮤니티 다운로드 2024 년 12 월 5.8 억 건에서 2025 년 11 월 14.4 억 건으로 늘었다. 결정적 지표는 파생 생태계다 — 자국산 오픈소스 가중치를 기반으로 파인튜닝된 모델의 비중이 2024 년 2 월 10%에서 2025 년 8 월 60%로 상승했다. 절반 이상의 파인튜닝이 중국산 오픈 가중치 위에서 일어난다는 것은, 중국 모델이 최종 소비재가 아니라 전 세계 2 차 개발의 '원자재'로 들어간다는 뜻이다. 21 세기경제보도가 지목한 저가 실현 방식 — 혼합전문가(MoE) 구조·차등 가격·개방 가중치(동방재부 전재, 2026-09-03) — 중 개방 가중치가 바로 이 경로다. 무상 배포로 다운스트림을 잠그고, 수익은 위 계층에서 회수한다.

자본시장은 이 구간에 이미 판정을 내렸고, 상장은 두 회사의 내부 원가 구조를 공시 의무로 외부에 노출시켰다. 즈푸 AI는 2026 년 1 월 8 일 홍콩증권거래소에 상장해 43.48 억 홍콩달러를 조달했고, 청약 총액은 1,977 억 홍콩달러에 달했다. 시가총액은 첫날 약 579 억 홍콩달러에서 1 월 12 일 1,100 억 홍콩달러를 돌파한 뒤 1 월 13 일 917 억 홍콩달러였다. 미니맥스(MiniMax, 稀宇科技)는 다음 날인 1 월 9 일 상장해 55.40 억 홍콩달러를 조달했고, 청약 총액 2,533 억 홍콩달러, 시가총액은 첫날 1,050 억 홍콩달러를 넘어 고점 1,200 억 홍콩달러를 넘긴 뒤 1 월 13 일 약 1,128 억 홍콩달러였다(21 세기경제보도, 2026-01-14). 미니맥스의 2025 년 1~3 분기 매출이 5,343.7 만 달러였다는 점을 두고 보면, 시장은 모델 회사의 현재 매출이 아니라 토큰 공급 인프라로서의 위치에 돈을 댔다. 상장이 만든 실질적 변화는 공시다 — 즈푸 AI는 상장사가 된 뒤 실적 설명회에서 연내 토큰 호출량 40 배 증가, 단위 토큰 추론 원가 연초 대비 80% 하락, 10 만장급 자국산 칩의 규모화 저비용 추론 실현을 공개했다(경제참고망, 2026-09-01). 비상장이었다면 나오지 않았을 원가 곡선이 공시 의무를 통해 시장 전체의 가격 기대를 다시 썼고, 이것이 백만 토큰당 1 달러 붕괴를 가속했다.

반대 견해를 둔다. '모델이 상품이 된다'는 명제는 최상위 프런티어 구간에는 적용되지 않으며, 실제로 상품화되는 것은 '충분히 좋은' 구간이다. GLM-5.3-Flash의 AA 57 점은 Claude Opus 4.8 과 동점이지만, CAICT의 헤드 목록에 GPT-5.2·Gemini 3 Pro·Grok-4·Claude 4.5 가 여전히 공존한다는 사실 자체가 단일 모델로 수렴하지 않는 최상위 구간의 존재를 보여준다. 토큰 단가 2.04 달러에서 0.97 달러로의 붕괴(실리콘데이터(Silicon Data) 지수, 2026-05 말~2026-09 초)는 대체 가능한 성능 대역에서 벌어진 가격 전쟁이지, 프런티어 능력 자체의 가치 하락이 아니다. 업계 분석(자쯔광넨, 2026-08-20)이 관측한 바로도 같은 모델에 하네스를 달리 씌우면 평가 성적이 평균 15.6%, 일부 영역에서 48%까지 벌어지고 "모델이 강할수록 하네스 기여 가치 비율은 오히려 증가한다" — 프런티어 모델일수록 그 성능을 실행으로 바꾸는 계층의 기여가 커진다. 상품화는 모델 계층 전체의 몰락이 아니라, 헤드 아래 대역의 마진을 압착하고 차별화 이익을 위 계층으로 밀어 올리는 힘이다.

스택 2 단 — 토큰 경제, 가격이 1 달러 아래로 내려간 뒤

토큰 단가는 2026 년 9 월 처음으로 백만 토큰당 1 달러 아래로 내려갔다. 실리콘데이터(Silicon Data)의 대언어모델 토큰 지출 지수 기준, 2026 년 5 월 말 2.04 달러이던 지수는 8 월 초 1.16 달러, 9 월 초 0.97 달러로 떨어졌다(21 세기경제보도 2026-09-03). 넉 달 새 절반 이하이고, 최근 7 일 변화율만 -8.6%다. 하락이 멈춘 것이 아니라 가속되고 있다는 뜻이다.

기준시점	백만 토큰당 지출	비고
2026 년 5 월 말	2.04 달러	
2026 년 8 월 초	1.16 달러	두 달 새 -43%
2026 년 9 월 초	0.97 달러	사상 첫 1 달러 아래
최근 7 일	-8.6%	

(출처: 실리콘데이터 대언어모델 토큰 지출 지수, 21 세기경제보도 2026-09-03)

하락을 만든 요인은 세 가지다(21 세기경제보도 분석). 첫째, 혼합전문가(MoE) 구조다. 전체 파라미터의 일부만 활성화하므로 같은 품질을 더 적은 연산으로 공급한다. 2026 년 신형 모델의 기본 사양이 됐다 — 즈푸 AI의 GLM-5.3-Flash는 총 320B 중 활성 18B 만 쓰고(양자위 2026-08-27), 커다윈페이(iFLYTEK, 科大讯飞)의 싱훤(Spark, 星火) X2.5는 293B-A30B 구조다(신랑과기 2026-09-07). 둘째, 차등 가격이다. 모델 등급과 호출 조건에 따라 단가를 세분해 평균 지출을 구조적으로 끌어내린다. 셋째, 개방 가중치 전략이다. 오픈소스 가중치가 공개되면 기업은 자체 배포라는 대안을 갖게 되고, 이 대안이 API 가격 전체에 상한을 씌운다. 자국산 오픈소스 대모델의 글로벌 누적 다운로드가 100 억 회를 넘고 월 점유율이 최고 17.1%에 달한 것(2025 년 12 월 기준, CAICT)이 이 상한 압력의 크기를 보여준다.

가격이 반토막 나는 동안 수요는 세 자릿수 배로 불었다. 화산엔진(Volcano Engine, 火山引擎)의 일평균 토큰 호출량은 더우바오(Doubao, 豆包) 출시 시점인 2024 년 5 월 1,200 억 토큰에서 2026 년 6 월 180 조 토큰 초과로, 2 년 만에 약 1,500 배 증가했다.

기준시점	일평균 토큰 호출량	근거
2024 년 5 월	1,200 억	더우바오 출시 시점(증권시보 2026-06-24 회고)
2025 년 10 월	30 조	화산엔진 사장 강연(CAICT 보고 각주)
2025 년 12 월	50 조	2025 겨울 FORCE 대회, 사장 탄다이(InfoQ 2025-12-18)
2026 년 6 월	180 조 초과	2026 여름 FORCE 원동력 대회(증권시보 2026-06-24)

일평균 호출량 자체는 정의가 일관되므로 네 시점의 비교가 성립한다 — 반년 사이 50 조에서 180 조로 3.6 배다. 그러나 부속 고객 지표는 단순 비교가 불가능하다. 2025 년 12 월 공시된 "토큰 1 조 초과 기업 100 곳 초과"는 **누적 호출량** 기준이고, 2026 년 6 월의 "토큰 1 조 초과 기업 200 곳"은 **연간 호출량** 기준이다. 집계 창이 서비스 개시 이후 전 기간에서 1 년으로 좁혀졌으므로, 100 곳에서 200 곳으로의 증가는 '반년 새 고객 두 배'로 읽을 수 없다. 오히려 문턱이 높아졌는데도 수가 두 배가 됐다는 사실이 고객당 소비 강도의 상승을 시사할 뿐이다. 참고로 같은 달인 2025 년 10 월 구글의 일평균 토큰 호출량은 42 조로, 화산엔진의 30 조를 앞섰다(CAICT 각주) — 단일 클라우드 사업자 기준 격차는 이미 배 이내였다.

수요를 받치는 공급은 국가가 직접 깐다. 공업정보화부는 지능 컴퓨팅 규모가 1,590 EFLOPS 를 초과했고 만장(万卡)급 지능컴퓨팅 클러스터가 42 개라고 밝혔다(2026-01-21 기자회견, 대변인 세춘). 2023 년 416.7 EFLOPS, 2024 년 725.3 EFLOPS 에서 올라온 곡선이다. 2026 년 3 월 통과된 15 차 5 개년 계획 강요는 제 12 장에서 "초대규모 지능컴퓨팅 클러스터" 건설 논증과 "정부의 컴퓨팅 서비스 구매·컴퓨팅 임대"를 명시하고, "중소기업의 컴퓨팅 비용 절감"을 정책 목표로 적었다. 베이징은 별도로 2 년 내 10 만 장급 자국산 지능컴퓨팅 클러스터 건설 계획을 내놨다. 토큰 원가의 하방은 시장 경쟁만이 아니라 기획 문서에 적힌 국가 과제다.

비용 구조 개선은 기업 실적에서 실측된다. 즈푸 AI(Zhipu AI, 智谱)는 2026 년 9 월 1 일 실적 설명회에서 연내 토큰 호출량 40 배 증가, 단위 토큰 추론 원가의 연초 대비 80% 하락, 10 만 장급 자국산 칩의 규모화 저원가 추론 달성을 공개했다(경제참고망). 홍콩 상장이 만든 공시 의무가 원가 곡선을 외부에 노출시킨 첫 사례다. 수요자 측에서는 우정저축은행(PSBC, 邮储银行)이 화웨이와 공동 발표한 유즈(邮智) 금융 업종 대모델에서 캐시 점유 약 72% 감소, 추론 비용 최대 87% 초과 절감, 모델 온라인 주기 2 주에서 1 일로 단축을 확인했다(이동지불망 2026-09-07). 인프라 계층에서는 취징커지(趋境科技)의 토큰 파드(Token Pod)가 프리필 단계를 추론 경로에서 분리해 독립 공급·독립 과금·독립 최적화가 가능한 자원 풀로 만들고 KV 캐시 적중률 90% 이상을 달성했다(광명망 2026-09-04). 토큰이 과금 단위를 넘어 인프라 설계 단위가 된 것이다.

한 줄로 결론을 내리면 이렇다. 가격이 1 달러 아래로 내려가는 순간 토큰 자체에서 남는 마진은 사라지고, 이익은 토큰을 파는 모델 계층이 아니라 토큰 호출을 만들어내는 계층 — 하네스와 에이전트 응용 — 으로 옮겨간다.

스택 3 단 ① — 하네스, 가치가 이동한 계층

하네스는 모델 주위에 놓이는 완전한 제어·오케스트레이션 시스템이다 — 도구 호출을 조정하고, 작업 상태를 관리하고, 실행 결과를 검증해 다시 모델에 되먹이는 폐쇄 루프를 담당하는 계층이다. 딥시크(DeepSeek, 深度求索)는 이 정의를 조직 안에 "모델 + 하네스 = 에이전트"라는 공식으로 새겼다(자쯔광넨 2026-05-20, 딥시크에 가까운 소식통과 공식 채용 페이지 인용). 2026 년 5 월 확인된 채용 공고에서 딥시크는 하네스 프로덕트 매니저와 연구개발 엔지니어 두 직무를 베이징 하이덴구에 두고 앤트로픽의 클로드 코드에 대응하는 전담 하네스 팀을 꾸렸다. 원문 보도의 비유대로 모델이 뇌라면 하네스는 신체와 신경계다. 백만 토큰당 가격이 2.04 달러에서 0.97 달러로 무너진 시장(실리콘데이터, 2026 년 5 월 말~9 월 초)에서 뇌는 더는 차이를 만들지 못한다 — 차이는 신경계가 만든다.

그 차이는 수치로도 제시됐다 — 단, 출처 등급을 먼저 밝힌다. 아래 수치는 업계 미디어 자쯔광넨의 장이자(张一甲)가 2026 년 8 월 20 일에 낸 연례 판단에서 온 것으로, 벤치마크 원자료가 공개되지 않은 '업계 분석' 등급이다. 그 관측에 따르면 같은 모델에 다른 하네스를 씌웠을 때 평가 성적의 평균 차이는 15.6%이고, 일부 영역에서는 격차가 48%에 달한다. 더 중요한 것은 방향이다 — 모델이 강해질수록 하네스가 기여하는 가치의 비중은 낮아지지 않고 오히려 커진다. 모델 성능이 하네스의 성과를 대체하지 못한다는 뜻이다. 이 계층은 학계의 주제로도 올라섰다. 같은 분석의 집계에서 재귀적 자기개선(RSI) 관련 arXiv 논문은 2025 년 1 분기 30 편 미만에서 2026 년 2 분기 약 500 편으로 늘었고, 하네스가 스스로를 개선하는 Self-Harness·Continual Harness 라는 개념이 2026 년에 등장했다.

딥시크가 이 계층에서 내놓은 물건을 해부하면 설계 의도가 드러난다. 딥시크는 2026 년 8 월 13 일 저녁 V4 Pro 정식판과 함께 딥시크 하네스(DeepSeek Harness) 개발자 프리뷰를 공개했다(평파이신문 2026-08-15). 설계 원칙은 "모든 것이 플러그인"이다 — 도구 호출, 파일 읽기·쓰기, 샌드박스 환경이 모두 탈착 가능한 부품으로 제공된다. 라이선스 조건은 보도 시점에 공개되지 않았다. 이 제품이 내세운 세 성질 — 제어 가능, 관측 가능, 추적 가능 — 은 기업 도입의 전제다. 추적 가능하지 않으면 에이전트가 잘못된 조치를 했을 때 원인을 특정할 수 없고, 관측 가능하지 않으면 실행 중인 작업을 운영팀이 감독할 수 없으며, 제어 가능하지 않으면 권한 밖 행동을 막을 장치가 없다. 2026 년 상반기 금융권 에이전트 입찰 97 건 중 리스크 관리·컴플라이언스가 19.6%를 차지했고(링이즈쿠 정리, 표본 기간 2026-01-01~06-15), 중국지급청산협회가 "당신의 에이전트를 알라(KYA)" 자율규약까지 낸 시장(2026-08-24)에서 감사 가능성은 선택 사양이 아니라 구매 자격이다.

모델 회사가 이 계층을 오픈소스로 여는 것은 계산된 교환이다. 얻는 것은 생태다. 하네스를 공개하면 개발자들이 그 구조 위에 플러그인과 도구를 쌓고, 한번 쌓인 자산은 전환 비용이 된다. 자국산 오픈소스가 이 전략의 성공을 이미 증명했다 — 자국산 오픈소스 기반 파인튜닝 모델 비중은 2024 년 2 월 10%에서 2025 년 8 월 60%로 올랐고, 누적 다운로드는 100 억 회를 넘었다(CAICT, 2025 년 12 월 기준). 잃는 것은 차별점이다. 하네스가 열려 있고 통용되면 그 안의 모델은 교체 가능한 부품이 된다 — 하네스의 승리가 곧 모델 상품화의 가속이라는 자기모순이다. 딥시크가 토큰 가격 붕괴를 주도한 회사라는 사실은 이 교환을

우연이 아닌 전략으로 읽게 한다. 상품이 된 모델을 팔 수 없다면, 모든 모델이 반드시 거쳐야 하는 관문을 소유하는 쪽이 낫다.

관문의 실력은 실측에서도 확인됐지만, 실측의 한계를 함께 봐야 한다. 저장의 실재즈닝(实在智能)이 만든 실재 Agent(实在 Agent)는 2026 년 7 월 27 일 OSWorld 벤치마크에서 성공률 90.2%를 기록해, 2024 년 이 벤치마크 발표 이후 성공률 90%를 넘은 첫 컴퓨터 사용 에이전트가 됐고 글로벌 총괄 순위와 에이전트 부문 모두 1 위에 올랐다(신화망 2026-08-05). 비교 대상에는 메타·엔트로픽·오픈 AI 의 공개 최고 기록이 있다. 그러나 벤치마크 성적과 실무 성능은 다르다. OSWorld 는 정해진 과업을 정해진 환경에서 수행하는 시험이고, 실제 기업 현장은 SCADA·MES·ERP 가 서로 고립되고 데이터가 파편화된 환경이다(타이메이티 2026-09-02). 시험장의 90.2%가 업무장의 90.2%를 보증하지 않는다 — 기업이 AI 의 핵심 업무 의사결정 참여를 신뢰한다는 비율이 23%에 그친다는 조사(타이메이티 2026-09-05)가 그 간극의 크기다.

이 간극을 메우는 작업이 하네스 공학이며, 가치는 아래 단계에서 위 단계로 올라간다. 앞선 업계 분석이 제시한 경로는 프롬프트 → 컨텍스트 → 도구와 상태 → 피드백 검증 → 다중 에이전트 거버넌스다. 프롬프트는 지시문 작성 기술로 진입장벽이 거의 없다. 컨텍스트는 모델에 무엇을 먹일지의 층으로, 알리바바가 천원반궁(Qwen Office, 千问办公)과 함께 컨텍스트 인프라 마이컨텍스트(MyContext)를 오픈소스로 공개한 곳이 이 층이다. 도구와 상태는 창신치즈(AlInnovation, 创新奇智)가 AlInnoGC 플랫폼에서 온톨로지와 하네스 엔지니어링을 묶어 에이전트 구축 기간을 수 일에서 수 분으로 줄인 층이다(인민망 2026-09-01). 피드백 검증은 출력을 자동으로 검수해 되먹이는 폐쇄 루프고, 다중 에이전트 거버넌스는 수백 개 에이전트의 동시 실행을 통제하는 문제다 — 화산엔진이 2026 년 6 월 시연한 500 여 개 에이전트 동시 협업·수천 회 도구 호출이 그 전경이다. 이 경로는 이번 호 뒤쪽 절들의 뼈대다. 피드백 검증의 미비는 6 절이 다룰 신뢰 논쟁의 근거이고, 다중 에이전트 거버넌스는 7 절의 실행 프레임워크 평가·인증이 규격화하려는 대상이며, 경로 상단에 쌓이는 자산은 9 절이 볼 기업 간 격차의 원천이다.

스택 3 단 ② — 플랫폼 전쟁, 왜 오피스가 전장이 되었나

오피스가 전장이 된 이유는 단순하다. 토큰 단가가 백만 토큰당 0.97 달러(2026 년 9 월 초·실리콘데이터)까지 떨어져 모델이 상품이 된 지금, 가치가 이동한 하네스 계층을 가장 빨리 독점할 수 있는 곳이 사용자의 일상 업무 화면이기 때문이다. 시장 온도는 이미 뜨겁다. 타이메이티(2026-09-05) 정리 기준 국내 데스크톱 AI 오피스 합계 월 방문량은 2026 년 3 월 2,000 만 회에서 6 월 6,000 만 회 돌파로 3 개월 만에 3 배가 됐고, 잠재 화이트칼라 사용자는 약 1.2 억 명이다. 중국 빅테크 4 사는 이 여름에 일제히 진입했다.

표: 오피스 에이전트 4 사 비교

플랫폼	출시일	모기업 자산	개방 정책	취득 인증	확인된 수치
알리바바 첸원반궁(Qwen Office, 千问办公)	2026-08-03 공개 베타	큐오더워크·우궁·물런 3 개 제품 통합, Qwen3.8-Max 기반·전용 모델 Qwen3.8-Flash	컨텍스트 인프라 마이컨텍스트(MyContext) 오픈소스	CAICT 오피스 에이전트 능력 평가 국내 첫 통과(2026-08-06)	1 개월 사용자 3,000 만 명 돌파·기업 사용자 과반, 1 개월간 버전 업데이트 120 회
바이트댄스 더우바오워크(Doubao Work, 豆包工作)		더우바오 대모델 + 페이수 심층 통합, 트레이·커우즈 통합	30 일 무료 구독	「오피스 에이전트 능력·클라우드 벤치마크 이중 인증」 국내 첫 그룹	더우바오 월간 활성 3 억
텐센트 워크버디(Tencent WorkBuddy)	2026 년 3 월 출시, 2026-09-02 개방 플랫폼	기업 위챗·텐센트 문서·텐센트 회의·TAPD 전 경로 페루프	개방 등록·무문턱 인증	「자동화 의사결정 — 지능 어시스턴트 2 성+급 표식 증서」 (광둥성 완창·워크버디)	파트너 100 곳 초과·공동 하드웨어 9 종·업종 앱 30 여 개, 데스크톱 월 방문량 1 위
바이두 다쯔(百度搭子)	2026-09-07 통합 발표	코딩 어시스턴트 원신콰이마(文心快码)를 통합, 업종 스위트 15 개·전문 오피스 스킬 96 개			월간 활성 674.30 만, 전월 대비 +1,063.79%(증가율 1 위)

첸원반궁은 신규 개발이 아니라 통합의 산물이며, 통합 이유는 사내 방문 수 격차에서 읽힌다.

타이메이티(2026-09-07)에 따르면 이 제품은 큐오더워크(QoderWork)·우쿵(悟空)·물런(MuleRun) 세 제품을 합친 결과물이다. 2026 년 6 월 기준 큐오더워크 방문 수는 약 788 만 회, 우쿵은 약 131 만 회로 6 배 차이였다. 승자와 패자가 갈린 세 개의 독립 제품을 나란히 두는 것보다 하나의 기함으로 묶는 편이 낫다는 계산이다. 2026 년 7 월 초 큐오더워크 팀(10 여 명)이 Q-Team 에서 천위썬(陈宇森) 조직으로 이관됐고, 덩위(丁宇)는 큐오더(Qoder) 코딩 사업을 유지했다. 통합의 결과는 1 개월 만에 사용자 3,000 만 명 돌파, 그중 절반이 기업 사용자(36 커 2026-09-06)라는 숫자로 나왔다. 분산된 실험을 접고 승자에 조직을 엮는 것이 알리바바의 답이었다.

텐센트 워크버디는 포지셔닝부터 다르다. 스스로를 "에이전트 시대를 향한 운영체제"라 칭한다(타이메이티 2026-09-05). 구조 설명은 "기반모델부터 하네스, 컨텍스트, 도구 호출, 시나리오 연결까지"다 — 하네스가 소비자 제품의 정체성 문구에 들어간 첫 사례다. 2026-09-02 내놓은 개방 플랫폼의 1 차 생태 규모는 파트너 100 곳 초과·공동 하드웨어 9 종·업종 앱 30 여 개·20 여 개 영역 커버이며, 반년간 50 여 버전을 반복했다. 책임자 린쥘루(林佐露)는 "생태는 계획으로 만들어지지 않는다"며 개방 등록·무문턱 인증 모델을 택했고, 판판(帆帆)이 플랫폼에 올린 제품은 출시 후 호출량 1 만 회를 돌파했다. 운영체제라는 이름의 실체는 기업 위챗·텐센트 문서·텐센트 회의·TAPD 라는 기존 협업 자산의 전 경로 페루프다.

더우바오워크의 무기는페이수(Feishu, 飞书)다. 기능은 "사용자 목표를 중심으로 과업을 자율 분해하고 도구를 호출해 복잡한 업무 흐름을 지속 추진"하는 것(신화망 2026-08-25)이며, 문서·표·PPT·이미지·영상의 생성 편집과 브라우저·PC 직접 조작을 지원한다. 핵심은페이수와의 심층 통합으로 기업 컨텍스트와 지식을 호출하고페이수 권한 체계를 엄격히 상속한다는 점이다. 즈둥시(2026-08-27)의 48 시간 실측에서 이 제품은 약 2.5 만 자의페이수 문서에 추적 가능한 자료 47 건을 인용해 생성했고, 발표자 노트를 포함한 14 페이지 PPT 를 만들었다. 30 일 무료 구독은 월간 활성 3 억의 더우바오 사용자를 기업 시장으로 옮기는 통로다.

개발 도구 계층에서도 같은 패턴이 반복된다. CAICT 「인공지능 산업발전 연구보고(2025 년)」(2026 년 1 월, 보고번호 No.202516) 표 1 이 정리한 국내의 도구를 대조하면, 국내 4 종 — 바이트댄스 커우쯔(Coze, 扣子)·바이두 에이전트빌더(AgentBuilder)·텐센트 위안치(元器)·화웨이 Versatile-AI — 는 공통으로로우코드·드래그앤드롭 워크플로와 자사 생태 결합(위안치는 훈위안 기반에 위챗 생태 통합)을 특성으로 둔다. 반면 미국 측 — 오픈 AI AgentKit·Agent SDK·Responses API, CrewAI, 랭체인, 마이크로소프트 AutoGen — 은 프레임워크와 API 중심이다. 중국의 플랫폼 전쟁은 개발자 프레임워크 전쟁이 아니라 이미 존재하는 업무 생태를 에이전트로 잇는 전쟁이다.

사용자 수는 거품을 가리고, 소비 집중도가 진짜 지표다. 랍스터 AI(LobsterAI)의 국내 첫 오피스 에이전트 사용자 행동 보고서(양자위 2026-09-07)에 따르면 상위 5% 사용자가 토큰의 53.5%를 소비하고, 유료 사용자의 소비량은 무료의 6.2 배이며, 토큰 소비의 약 60%가 정규 근무시간 밖에서 발생하고, 모델 호출의 12.2%는 사람이 실시간으로 개시하지 않은 비대화형 호출이다. 1 회 태스크 규모는 5 개월간 3.1 배로 커졌다. 발표 주체가 자사 실사용 데이터를 공개한 이해당사자임을 밝혀두지만, 방향은 분명하다. 첸원반궁의

3,000 만이라는 등록자 수보다 상위 5%의 53.5%라는 집중도가 수익 구조를 결정한다. 시장 전체로도 월 방문 6,000 만 회 대비 일간 활성은 여전히 낮아 약 20 배의 성장 여지가 남아 있고(타이메이티 2026-09-05), 기업이 AI의 핵심 업무 의사결정 참여를 신뢰하는 비율은 23%에 그친다. 경쟁의 4 중문이 "가용도→진입점→신뢰→확산"인 이유다.

반대 견해가 필요하다. 텐센트는 워크버디가 "일부 전통 SaaS 기능을 완전히 대체"할 수 있다고 주장하지만 이는 벤더 주장이다. 실제 구조를 보면 오피스 에이전트는 기존 협업 소프트웨어의 대체물이 아니라 껍데기에 가깝다. 더우바오워크는 페이스 권한 체계를 "상속"하고, 워크버디는 기업 위챗·문서·회의·TAPD의 전 경로를 묶는다 — 대상을 없애는 것이 아니라 그 위에 얹는 것이다. 신뢰 23%의 벽 앞에서 오피스 에이전트의 현재 위치는 운영체제가 아니라 기존 소프트웨어 위의 새 진입점이며, 진입점이 운영체제가 되려면 다음 관문은 사용자 수가 아니라 신뢰다.

스택 3 단 ③ — 프로토콜과 표준, 중국이 만드는 두 번째 규격

모델이 상품이 되면 승부는 연결 규격으로 넘어간다. 에이전트가 도구를 부르고 다른 에이전트와 일하는 방식을 누가 정의하느냐가 미들웨어 단의 전쟁이며, 중국은 이 전쟁에 해외 규격의 채택과 자국 규격의 창안을 동시에 진행하고 있다. 현재 대결 구도는 다음 표로 압축된다.

프로토콜	무엇을 표준화하는가	발의 주체	중국 내 위치
MCP	모델과 외부 도구·데이터 소스의 연결 — "외부 도구·데이터 소스의 즉견즉용(即插即用)"을 밀어붙이는 통일 연결 방식	해외 오픈소스 진영	해외 발 규격이나 중국 개발 도구에 이미 흡수됨 — 화웨이 Versatile-AI의 기능 목록에 MCP가 명기돼 있다(CAICT 보고 표 1, 2026년 1월)
A2A	에이전트 간 협업 — 에이전트카드(AgentCard) 메타데이터 모델로 에이전트 능력을 기계 판독 가능한 구조화 서술로 추상화	해외 오픈소스 진영	중국 내 자국 규격 3종과 경쟁·병존
ANP	다중 에이전트 상호 연결·협업	ANP 오픈소스 기술 커뮤니티(중국)	CAICT가 "중국의 다중 에이전트 상호 연결을 지탱한다"고 공인한 자국 규격
ACP	다중 에이전트 상호 연결·협업	AgentUnion(중국)	CAICT 공인 자국 규격
RVP	허실융합(虚实融合) 통신 — 가상 에이전트와 실물 환경의 결합	해촨커지(氮川科技, 중국)	CAICT 공인 자국 규격

두 규격군의 분업을 정확히 봐야 한다. MCP는 '에이전트와 세계' 사이의 인터페이스이고, A2A는 '에이전트와 에이전트' 사이의 언어다. 전자는 도구 호출을, 후자는 협업 상대의 능력 발견을 표준화한다. CAICT 보고가 MCP·A2A를 "해외 오픈소스"로 명기하면서 ANP·ACP·RVP를 대등한 열거로 붙인 것은, 중국이 해외 규격을 수용하되 에이전트 간 통신 계층에서는 자국 규격을 병행 육성하겠다는 선언이다.

규격 창안은 민간 커뮤니티에만 맡겨져 있지 않다. 중국의 표준 경로는 3층 구조로 국가가 직접 주도한다.

첫째 층은 공업정보화부 인공지능 표준화기술위원회다. 위원회는 10대 중점 표준 방향을 정했고 그중 에이전트 관련이 2건이다 — 「인공지능 관건기술 에이전트 기초 기술 능력 요구」(계획번호 2024-1332T-YD)와 「인공지능 관건기초기술 에이전트 서비스 응용 기술 요구」(계획번호 2025-0275T-YD)이다(CAICT 보고, 2026년 1월). 나머지 방향에는 산업 경계 확정·기업 인정, 소프트웨어·하드웨어 협동, 대모델 평가, AI

공학화, 피지컬 AI, 휴머노이드 로봇이 포함된다. 실적은 속도를 보여준다 — 중국통신표준화협회(CCSA)와 공동으로 표준 20 건을 발표했고, 연간 표준 입안 건의 176 건을 신규 접수해 11 건을 보고 승인, 25 건을 심사고까지 완료했다(CAICT 보고).

둘째 층은 CAICT 의 평가·인증으로, 두 갈래다. 한 갈래는 「가신(可信) 인터넷 에이전트 운영체제 — 에이전트 실행 프레임워크」 평가로, 1 차 통과자는 련통 데이터지능(联通数据智能)의 「위안징 완우(元景万悟)」

에이전트 플랫폼과 차이나텔레콤(China Telecom, 中国电信) 베이징연구원의 「AgentOS」 범용 에이전트 프레임워크 2 건이다(평파이신문, 2026-07-30). 평가는 자원 스케줄링·관리, 런타임 계층, 상호작용 서비스 계층의 3 대 모듈에서 이중 컴퓨팅 스케줄링·탄력 확장 등 10 개 핵심 지표를 본다. 첫 통과자가 국유 통신 2 사라는 배치는 이 절의 앞부분에서 본 구조 — 국가가 표준을 만들고 국유 기업이 먼저 인증받는다 — 의 그대로다. 차기 평가 접수는 2026-10-30 에 마감되고 증서는 2026 년 12 월 산업대회에서 수여된다. 다른 갈래는 오피스 에이전트 능력 평가로, 첸원반궁(Qwen Office, 千问办公)이 지능형 상호작용·멀티모달, 전 시나리오 오피스 작업 실행, 보안 방호, 스킬(Skills) 관리의 4 대 영역 16 개 항목을 충족해 국내 첫 통과를 공개했고(2026-08-06), 더우바오워크(Doubao Work, 豆包工作)도 오피스 에이전트 능력·클라우드 벤치마크 이중 인증을 국내 첫 그룹으로 통과했다(신화망, 2026-08-25).

셋째 층은 국제 표준이다. 중국은 ITU-T 에서 인공지능 표준을 47 건 발표하고 63 건을 연구 중이며, 대모델 평가·인공지능 플랫폼·디지털 휴먼 표준은 "국제 공백을 메웠다"고 CAICT 가 평가한다(CAICT 보고). 자국 규격을 먼저 만들고 국제 무대에서 기정사실화하는 순서다.

이 3 층 위에서 '에이전트 운영체제'라는 개념이 떠오르고 있다. CAICT 는 「에이전트 운영체제 기술 진화 연구보고(2026)」 편찬에 착수해 1 차 연구 세미나를 2026-09-10 에 연다(평파이신문, 2026-09-04).

보고가 구분하는 3 대 기술 방향은 에이전트 실행 프레임워크, 에이전트 임베디드 운영체제, 에이전트 네이티브 운영체제다. 이 구분은 플랫폼 락인의 형태를 결정한다. 실행 프레임워크에 머물면 락인은 프레임워크 호환성 수준이고, 기존 운영체제에 에이전트를 임베딩하면 기존 OS 벤더의 생태에 종속되며, 에이전트를 전제로 설계한 네이티브 운영체제가 이기면 그 OS 자체가 진입점이 된다. 텐센트

워크버디(Tencent WorkBuddy)가 "에이전트 시대를 향한 운영체제"를 표방하고 개방 플랫폼(파트너 100 곳 초과, 2026-09-02 출시)으로 생태를 먼저 쌓는 것은 이 세 번째 경로를 겨냥한 움직임이다.

CAICT 가 실행 프레임워크 평가 인증을 먼저 돌리고 운영체제 연구보고를 뒤이어 편찬하는 순서는, 국가가 어느 방향이 이기든 인증 제도가 그 위에 얹히도록 설계하고 있음을 보여준다.

마지막 고리는 평가·인증이 곧 조달 자격이라는 점이다. 구이저우의 정무 에이전트 「구이런즈반(贵人智办)」은 "중앙사이버안보정보화판공실 생성형 AI 대모델 등록을 통과한 국내 첫 정무서비스 종합 에이전트"라는 자격으로 조달됐다(양시망, 2026-08-28). 등록이 기능 명세가 아니라 입찰 문턱으로 작동한 것이다. 금융 입찰 97 건(2026-01-01~06-15, 링이즈쿠)과 정무 조달이 늘수록 CAICT 인증서는 품질 보증이 아니라 판매 허가가 된다.

한국 기업에 남는 질문은 기술 우열이 아니라 진로 선택이다. MCP·A2A 축 위에 설 것인가, ANP·ACP·RVP 와 CAICT 인증의 축 위에 설 것인가. 하네스·프레임워크·운영체제 인증이 조달 자격으로 굳어지는 속도를 보면, 어느 규격 위에 서느냐가 5 년 뒤의 전환 비용을 결정한다. 표준은 한번 자산으로

쌓이면 갈아타는 데 규격 전체를 다시 갈아야 하는 비용이 따른다. 모델은 언제든지 바꿀 수 있지만 규격은 그렇지 않다.

스택 4 단 ① — 금융, 입찰에서 결제까지

응용 단은 조달이 주도한다는 앞 절의 명제는 금융에서 가장 먼저 데이터로 확인된다. 링이즈쿠(零壹智库)가 입찰 정보 플랫폼 치예위징통(企业预警通)의 원자료를 토대로 2026 년 1 월 1 일~6 월 15 일 금융기관 에이전트 입찰 공고를 정리한 결과, 유효 표본은 97 건이다(21 세기경제보도 2026-06-24). 97 건이라는 작은 표본이 의미를 갖는 이유는, 그 안에 구매 주체와 구매 단위의 변화가 함께 드러나기 때문이다.

기관 유형으로 보면 은행이 절반을 차지한다.

기관 유형	입찰 건수	비중
은행(농촌신용사 포함)	47 건	48.5%
증권사	14 건	14.4%
보험사	12 건	12.4%
기타(금융리스·소비금융 등)	24 건	24.7%

은행 비중 48.5%에는 농촌신용사까지 포함된다. 기관별로는 Ningbo Bank·Wuzhou Bank가 각 6 건, Shengye Bank가 자회사 포함 6 건, Guangfa Securities 5 건, Guilian Asset Insurance 3 건으로, 대형 국유은행이 아니라 지방 중소은행과 증권사가 입찰을 주도했다.

수요 유형으로 나누면 구매 단위의 이동이 보인다.

수요 유형	비중
플랫폼 기반(底座)	28.9%
리스크 관리·컴플라이언스	19.6%
마케팅·고객서비스	11.3%
기타 시나리오	40.2%

에이전트 공통 기반을 통째로 사는 플랫폼 조달은 28.9%에 그치고, 개별 업무 시나리오 단위 조달이 40.2%로 이미 가장 크다. 2 위 단일 항목이 리스크 관리·컴플라이언스(19.6%)라는 점도 중요하다 — 사는 것 자체가 '통제 가능한 실행'으로 옮겨가고 있다는 뜻이다. 중국 금융기관의 에이전트 도입은 인프라를 까는 단계를 지나, 플랫폼 조달과 시나리오 조달이 공존하는 중간 지점에 있으며 무게중심은 시나리오 쪽으로 기울고 있다.

입찰 건수는 곧 성과가 아니다. 실측은 반기보고에서 나온다. 아래 수치는 중국경영보(2026-09-05)가 각 은행의 2026 년 반기보고를 인용해 보도한 것이다.

은행	에이전트·시나리오 규모	성과 지표
핑안은행	운영심사 에이전트 140 개(시나리오 55 개), 소비자보호 심사 요점 모델 누적	소비자보호 심사 지능화 커버율 90% 초과, 국경 간 대출 효율 약 30% 향상

	35 개	
난징은행	고부가 업무 에이전트 누적 156 개(반기 신규 30 개)	—
공상은행(ICBC)	대모델 600 여 개 시나리오 적용	내부통제·컴플라이언스 심사의견 36 만 건 산출, 채택률 98.6%, 심사 인원의 리스크 노출 발견 능력 130% 향상

우정저축은행(PSBC, 邮储银行)은 '어디까지 가려는가'를 수치로 공개한 사례다. 화웨이와 공동 발표한 「유즈(邮智)」 금융 업종 대모델은 행내 시스템 90 개 초과에 접속했고, 추론 비용을 최대 87% 절감하면서 모델 온라인 주기를 2 주에서 1 일로 줄였다(이동지불망 2026-09-07).

지표	현재(2026-09)	2026 년 말 목표
일평균 호출	600 만 회 초과	1,000 만 회 초과
일평균 토큰 처리량	100 억 초과	300 억 초과
착지 응용 시나리오	370 개 초과	1,000 개

시나리오를 연말까지 세 배 가까이 늘리겠다는 목표는, 은행이 에이전트를 실험이 아니라 확장하는 운영 자산으로 취급하기 시작했음을 보여준다.

한편 고객 접점은 은행 바깥으로 이동하고 있다. 차이렌서(2026-09-07) 보도에 따르면 알리바바 천원(Qwen, 千问) 앱에 금융 에이전트가 처음으로 대거 탑재됐다. 참여 기관은 10 여 곳으로, 상예증권·귀타이하이통·저상선물·이팡다펀드 등 투자 계열 9 종과 중안보험·샤오위산·후이쩌보험 등 보험 계열 3 종이다. 사용자는 천원 안에서 곧바로 시장 분석·투자 연구·펀드 선정·보험 상담을 호출한다. 금융기관은 에이전트를 공급하지만 고객이 머무는 화면은 플랫폼의 것이다. 고객 접점을 플랫폼이 흡수하는 구조이며, 은행과 증권사는 '에이전트 공급자'로 밀려날 수 있다.

접점 이동의 종착지는 결제다. 중국지급청산협회(中国支付清算协会)는 2026 년 8 월 24 일 「에이전트 결제 응용 자율규약」(智能体支付应用自律公约)을 발표했다(인민일보 2026-08-27 제 8 면). 4 대 조항은 ① '당신의 에이전트를 알라'(KYA, Know Your Agent) 메커니즘 구축 ② 사용자와의 권한 협약 체결 ③ 등급별 분류 관리·자금 안전 강화 ④ 불만 접수·에이전트 결제 종료·정보 조회 채널 제공이다. 중국은련(中国银联)은 15 개 상업은행과 AI 스마트카드를 출시했고 평안은행·푸파은행·교통은행이 참여하며, 위챗페이(微信支付)와 알리페이(支付宝)도 합류했다. 자율규약은 강제력 없는 업계 협약이지만, 결제라는 영역에서 에이전트의 행위 규칙을 처음 문서화했다는 점에서 경성 규제 전주곡이다.

KYA 가 첫 조항이라는 사실이 이 규약의 성격을 말해준다. 에이전트가 돈을 움직이기 시작하면 규제는 모델이 무엇인지가 아니라 행위가 무엇인지를 본다. 누구의 에이전트인가, 어떤 권한을 위임받았는가, 거래를 어떻게 멈추고 조회하는가 — 질문의 주어가 모델에서 행위로 바뀐다. 4 대 조항 모두 모델 능력이 아니라

신원 확인·권한 위임·등급 분류·사후 규제라는 행위 감사의 문법으로 쓰였다. 입찰 표의 리스크·컴플라이언스 조달(19.6%)과 이 규약은 같은 수요의 양쪽이며, 10 절에서 다룰 지능정보서비스 규제와 청량(淸朗) 정비의 집행 논리와도 한 축 위에 있다.

스택 4 단 ② — 제조와 정부, 결정성 요구와 조달 구조

금융이 조달로 움직인다면, 제조와 정부는 그 조달의 규모를 국가가 직접 설계한다는 점에서 다르다. 2026 년 1 월 8 일 8 개 부처가 공동 발표한 「'인공지능+제조' 전문 행동 실시 의견」은 2027 년까지 통용 대모델 3~5 개를 제조업에 심층 응용하고, 고수준 공업 에이전트 1,000 개를 내놓으며, 공업 영역 고품질 데이터셋 100 개를 구축하고, 전형 응용 시나리오 500 개를 보급한다는 목표를 숫자로 못 박았다(신화망 2026-01-13). 응용 단의 수요가 시장이 아니라 문건에서 먼저 정의된다.

현장 성과는 이미 나오고 있다. 창안자동차(Changan Auto, 长安汽车)는 천원반공을 연구개발·생산제조·공급망·판매·서비스 다섯 영역에 배치했다(신화망 2026-09-03).

영역	사례	효율
연구개발	와이어하니스 선정 계산 도구	2 일 → 5 분
연구개발	조명 효과 시뮬레이터	수 주 → 회의 1 회
생산	차량 동봉 문서 검사 미니앱	15 분 → 1~2 분/대
공급망	포장 심사 자동화 스크립트	작업량 90% 이상 경감
판매·서비스	사용자 표본 분석	수 일 → 반나절 이내

와이어하니스 선정 계산이 2 일에서 5 분으로 줄었다는 수치는 인상적이지만, 창안의 사례가 모두 표준화된 반복 업무라는 점을 놓치면 안 된다. 창신치즈(Alnovation, 创新奇智)는 2026 년 8 월 27 일 AlnnoGC 공업 에이전트 플랫폼을 대폭 업그레이드하며 개발 패러다임을 '온톨로지(공업 본체) + 하네스 엔지니어링 + AI 코딩'의 3 위일체로 정의했다(인민망 2026-09-01). 설비관리·에너지관리·제조데이터분석 제품군의 구축 기간이 수 일에서 수 분으로 압축됐다. 이 3 요소가 5 절의 하네스 논의와 정확히 겹친다는 점이 핵심이다 — 제조 현장의 에이전트도 결국 모델이 아니라 온톨로지와 하네스에서 성능이 갈린다.

그러나 제조는 가장 어려운 응용지다. 타이메יתי(2026-09-02) 보도에서 바오강(Baosteel, 宝钢)의 공학 도면 설계 효율이 10 배 향상됐고 산둥징공전자의 리튬전지 라인 설비 경보가 페루프로 돌아가지만, 동시에 세 가지 장애가 지적됐다 — 산업 데이터의 파편화, SCADA·MES·ERP 시스템의 고립, 그리고 모델의 확률적 생성 메커니즘과 공업 생산의 결정성 요구 사이의 불일치다. 세 번째가 본질이다. 잘못된 답을 그럴듯하게 내놓는 모델의 성질은 도면 초안에서는 감당할 수 있어도 생산 라인 제어에서는 허용되지 않는다. 이 불일치는 모델을 더 키운다고 풀리지 않고, 검증·롤백·감사를 담당하는 하네스 계층에서만 통제된다.

그래서 도입은 쉬운 곳에서 어려운 곳으로 이동 중이다. CAICT 의 공업 대모델 응용 분포를 보면 생산제조 단계의 비중이 18.8%에서 25.9%로 올랐고, 후단 운영관리 단계가 여전히 최고이며 소폭 상승한 반면, 전단 연구개발설계 단계 비중은 소폭 하락했다. 중국과학원 다론회화학물리연구소의 지능화공 대모델 2.0 Pro 가 실험 효율을 10 배 이상 끌어올리고, 중국강연(中国钢研)의 야금 공정 감지 대모델이 금속조직 분석과 표면 결함 검출 정확도를 각각 95% 초과로 끌어올린 사례가 그 이동의 앞머리다.

정부는 조달의 또 다른 얼굴이다. 장쑤성 우시시(无锡市)에서는 이미 100 개의 에이전트가 체계적으로 가동 중이다(신화망 장쑤 채널 2026-08-24). 「즈후이(知惠)」 정부 대모델 위에 딥시크를 포함한 18 개 주류 대모델을 배포하고 정부 데이터 1,900 만 건 이상을 표준 입고했다.

에이전트	성과
지능 안내	기업 신청 서류 제출 생략, 신청 효율 30% 향상
지능 보조 심사	심사보고서 자동 생성, 심사자 효율 70% 향상
지능 교통영향평가	검토의견 작성 2 영업일 → 1.5 시간, 정확도 90% 이상
시공도면 심사	업무 효율 400% 이상 향상
상사 등기	심사 1 회 통과율 약 70% 향상

정부 에이전트의 도입에는 민간에 없는 관문이 하나 더 붙는다. 구이저우성의 「구이런즈반(贵人智办)」은 중앙사이버안보정보화판공실의 생성형 AI 대모델 등록(备案)을 통과한 국내 첫 정부서비스 종합 에이전트로 조달에 들어갔다(양시망 2026-08-28). 지능 문답으로 상담 255 만여 건을 응대하고, 지능 순찰 효율을 인력의 50 배로 끌어올렸다. 광둥성은 텐센트와 「완칭·워크버디(湾擎·WorkBuddy)」를 2026 년 8 월 18 일 발표해 정부 서비스·정부 협동·보조 의사결정·사회 거버넌스 네 영역에 시범 적용하며 '자동화 의사결정 지능 어시스턴트 2 성+급 표식 증서'를 취득했다(다양망 2026-08-18). 등록과 등급 인증이 정부 조달의 자격 요건으로 작동한다는 사실 — 이것이 7 절에서 본 '평가·인증이 곧 조달 자격'이라는 구조의 정부판이며, 한국이 볼 지점은 중국의 정부 조달이 플랫폼 벤더에게 대규모 레퍼런스 공장 역할을 한다는 데 있다.

정책·규제 — 보급률 목표와 행위 규제가 동시에 간다

중국의 에이전트 정책은 촉진과 규율이 한 손에서 나온다. 두 축을 나눠 읽어야 진짜 조건이 보인다.

촉진 축의 상단에는 국무원 「'인공지능+' 행동 심층 실시에 관한 의견」(국발〔2025〕11 호, 발문 2025-08-21)이 있다. 2027 년까지 차세대 지능형 단말·에이전트 등의 응용 보급률 70% 초과, 2030 년 90% 초과, 2035 년 지능경제·지능사회 발전 신단계 진입을 국가 목표로 걸었고, '에이전트 애즈 어 서비스'라는 발전 모델을 명시했다(국가인터넷정보판공실 게재 2025-08-27). 여기서 흐리기 쉬운 지점을 정확히 해야 한다 — 70%·90%는 에이전트 단독 목표가 아니라 '차세대 지능형 단말·에이전트 등'을 묶은 보급률이다. 이 구분을 지우면 목표를 과대 해석하게 된다. 2026 년 3 월 통과된 15 차 5 개년 계획 강요가 그 아래를 받친다. 계획은 '모심운용(模芯云用)' 협동 혁신을 강화하고 다모달·에이전트·피지컬 AI·군체지능 기술 혁신을 장려하며, 정부 구매 컴퓨팅 서비스와 컴퓨팅 임대료 수요를 충족해 중소기업의 컴퓨팅 비용을 낮춘다고 적었다. 소프트웨어·정보, 금융·상무, 교통·물류 서비스업에서 지능 단말·에이전트를 광범위하게 응용한다는 조항(전란 10)이 응용 단 확산의 정책 근거다.

규율 축은 세 단계로 조여든다. 첫째, 등록과 등기라는 사전 관문이다. 생성형 AI 서비스 등록(备案)은 2026 년 6 월 30 일 기준 누적 988 건, 등기(登记)는 598 건에 이르렀다(국가인터넷정보판공실 2026-07-10 공고). 두 달 전 기준(备案 868 건·登记 530 건)과 비교하면 5~6 월에만 등록 120 건이 새로 늘었다.

등록은 이미 정무 조달의 자격으로 작동하기 시작했다(9 절 구이런즈반). 둘째, 집행이다. 「청량(淸朗)·AI 애플리케이션 무질서 정비」 전문 행동은 2026 년 4 월 개시돼 1 단계에서 위규 웹사이트·앱·에이전트 등 AI 상품 1.4 만여 개를 처치하고, 위법 정보 600 여만 건을 정리했으며 계정 2.6 만여 개를 처분했다(신화망 2026-07-06). 여기서 에이전트가 규제 집행의 명시적 대상 품목으로 처음 열거됐다는 사실이 이번 호의 핵심 규제 신호다. 셋째, 사전 규범의 신설이다. 「인공지능 의인화 상호작용 서비스 관리 잠행 방법」이 5 개 부처 공동으로 2026 년 4 월 10 일 공포돼 7 월 15 일 시행됐고, 안전 평가·알고리즘 등록과 함께 인공지능 샌드박스 안전 서비스 플랫폼 건설, 미성년자·고령자 권익 보호를 제도로 담았다(국가인터넷정보판공실). 이어 「인터넷 정보 서비스 관리 방법」 개정안이 2026 년 8 월 2 일까지 의견을 수렴하며 '지능정보서비스' 장을 신설했다. 이 장은 기술의 기본 원리·주요 운영 메커니즘·훈련 데이터 출처를 공시하고, 생성·합성 콘텐츠를 표시하며, 사용자에게 지능정보서비스 사용을 강제하지 못하게 한다.

세 단계를 관통하는 방향은 하나다. 규제의 주어가 모델 등록(무엇을 만들었나)에서 에이전트 행위(무엇을 했나)로 이동하고 있다. '지능정보서비스'라는 넓은 명칭은 에이전트가 사용자를 대신해 조작하는 행위를 사정권에 넣고, 청량 정비는 그 행위를 사후에 처분하며, 결계 자율규약의 KYA(8 절)는 행위 이전에 신원과 권한을 확인한다. 한국 기업의 대중 진출과 협력에 이 이동이 요구하는 것은 분명하다 — 모델을 등록할 준비가 아니라, 에이전트가 무엇을 했는지 공시하고 표시하고 멈출 수 있는 행위 감사 체계를 갖출 준비다.

한국 시사점·액션 플랜

이번 호의 결론을 한 문장으로 줄이면, 한국의 과제는 모델 국산화 논쟁이 아니라 하네스 계층의 자산화다. 토큰 가격이 백만 토큰당 1 달러 아래로 무너진 시장(4 절)에서 모델을 직접 만들 경제적 이유는 사라졌고, 성능을 가르는 지점은 모델을 업무에 붙이는 계층으로 이미 이동했다(5 절). 중국은 그 계층을 오픈소스·국가표준·정부금융 조달의 세 갈래로 굳히는 중이다. 한국 기업이 지금 축적해야 할 자산은 세 가지다 — 업무 온톨로지의 정의 체계, 도구·데이터 연결 규격의 선택(MCP·A2A 대 중국 발 ANP·ACP·RVP), 에이전트 행위 로그·권한·감사 체계(중국의 KYA 에 상응하는 내부 규율)다. 이 셋을 소유하면 모델은 언제든지 교체할 수 있고, 소유하지 못하면 어느 벤더에 붙든 종속된다.

주체	0~3 개월	3~6 개월	6~12 개월
SI·SaaS	모델 무관 실행 구조 설계 — 오픈소스 모델 2 개 이상을 같은 하네스에서 교체 가능하게	고객사 핵심 업무 10 개의 온톨로지 문서화(창신치즈 3 위일체 참조)	온톨로지·하네스를 재사용 가능한 제품 자산으로 패키징
협업 소프트웨어	첸원반궁·더우바오워크·워크버디의 실제 대체 범위 벤치마크	페이수식 '컨텍스트 통합'에 대응할 자사 데이터 권한·지식 연결 설계	오피스 에이전트의 계약 잠식에 대비한 기능·가격 재편
클라우드·MSP	국산 가속기+오픈소스 하네스 조합의 추론 단가 실측	KYA 에 상응하는 에이전트 신원·권한·종료 절차를 매니지드 서비스로 상품화	행위 감사 로그를 규제 대응 자산으로 제공
정책 입안자	CAICT 평가·MCP/A2A 갱신·중국 발 규격 추적 체계 구축	국내 에이전트 표준·행위 감사 규율 논의 착수	조달 자격과 연동되는 평가·인증 프레임 검토

전략기획 차원에서 가장 먼저 승인할 것은 예산 재편이다. AI 예산에서 모델 라이선스 비중을 낮추고 하네스·오케스트레이션 비중을 올리는 것 — 토큰 단가가 2026 년 5 월에서 9 월 사이 절반으로 떨어진 것이 그 근거다. 개발 조직은 딥시크가 내부 하네스 요건으로 삼은 '제어 가능·관측 가능·추적 가능' 세 성질을 그대로 자사 요구사항으로 채택할 만하다. 정보보호·감사 부서는 중국이 KYA 를 결제 규약으로 굳히기 전에 에이전트 신원 확인·권한 위임·종료 절차를 사내 규정으로 먼저 세워야 한다.

기회만큼 위험도 분명하다. 첫째, 중국산 오픈소스 모델과 하네스를 묶어 채택할 때 공급망·라이선스·데이터 잔류 리스크가 따른다 — 딥시크 하네스의 라이선스 조건은 원문에 명시되지 않았고, 자국산 칩에서 훈련된 모델의 가중치를 그대로 들여올 때 검증 부담이 남는다. 둘째, 어느 규격 위에 설지를 지금 성급히 고착하면 5 년 뒤 전환 비용을 스스로 만든다. 셋째, 첸원반궁 같은 오피스 에이전트가 기존 협업 소프트웨어 계약을

잠식할 가능성이 있다. 다만 그 잠식은 아직 확정되지 않았다 — 기업이 AI의 핵심 업무 의사결정 참여를 신뢰하는 비율은 23%에 그치고(타이메이티 2026-09-05), 신뢰야말로 다음 관문이다.

무엇이 관측되면 판단을 바꿔야 하는지를 남긴다. 한국의 구체 도입 통계는 확인된 값이 없으므로 창작하지 않고, 대신 다음 지표를 감시 대상으로 둔다.

모니터링 지표	판단을 바꾸는 신호
토큰 가격 방향(실리콘데이터 지수)	하락이 멈추고 반등하면 모델 상품화 명제가 흔들린다
하네스 기여도(평균 15.6%·최대 48%)	차세대 모델이 하네스 없이 장기 태스크를 처리해 기여도가 축소되면 자산화 우선순위가 내려간다
오피스 에이전트 유료 전환	사용자 수가 유료·기업 계약으로 이어지지 않으면(상위 5%가 토큰 53.5% 소비) 거품 신호다
중국 발 규격 채택(ANP·ACP·RVP)	실채택 없이 목록으로만 남으면 규격 종속 위험이 완화된다
규제 집행 강도(청량·지능정보서비스)	집행이 국내 배포 속도를 실제로 늦추면 진출 경로를 재설계한다

규격은 시장이 아니라 위원회에서 만들어진다. CAICT 차기 에이전트 실행 프레임워크 평가의 접수 마감인 2026년 10월 30일이라는 사실은, 한국이 국내 표준 논의에 의견을 개진할 시간이 이미 카운트다운에 들어갔음을 뜻한다.

부록 — 주요 지표와 출처

이번 호에서 쓴 핵심 수치를 출처 등급과 함께 모은다. 등급은 1 차(원문·공시)·보도(언론이 1 차 자료를 인용)·업계 분석(민간 미디어의 판단)·시장조사(민간 리서치 추정)로 구분했다. 등급이 낮을수록 단독 인용을 피하고 방향성으로만 읽는다.

지표	값	기준시점	출처	등급
토큰 지출 지수	백만 토큰당 0.97 달러(첫 1 달러 미만)	2026-09 초	실리콘데이터·21 세기경제보도	1 차(지수)
화산엔진 일평균 토큰 호출	180 조 초과	2026-06	증권시보	보도
하네스 평가 성적차	평균 15.6%·최대 48%	2026-08	자쯔광넨	업계 분석
자국산 오픈소스 누적 다운로드	100 억 회, 월 점유율 최고 17.1%	2025-12	CAICT	1 차
금융 에이전트 입찰	유효 97 건(은행 48.5%)	2026 상반기	링이즈쿠·치예위징통	2 차
공상은행 컴플라이언스 심사의견	36 만 건, 채택률 98.6%	2026 상반기	중국경영보	보도
지능 컴퓨팅 규모	1,590 EFLOPS 초과, 만장급 클러스터 42 개	2026-01	공업정보화부	1 차
생성형 AI 등록·등기 누적	등록 988 건·등기 598 건	2026-06-30	국가인터넷정보판공실	1 차
청량 정비 AI 상품 처치	1.4 만여 개(에이전트 명시 열거)	2026-07	신화망	보도
GLM-5.3-Flash	총 320B·활성 18B, AA 순위 57 점	2026-08	양자위	보도

스택 4 단을 사업자로 옮기면 다음과 같다. 한 사업자가 여러 단에 걸치는 것이 이 산업의 특징이다.

단	대표 사업자·제품
모델	즈푸 AI(GLM-5.3-Flash)·미니맥스·딥시크(V4 Pro)·통이첸원(Qwen)·커다윈페이(싱훤 X2.5)
컴퓨팅·토큰	화산엔진·바이렌(MaaS)·토큰 파드(취징커지+무어스레드)
하네스·플랫폼	딥시크 하네스·첸원반궁·더우바오위크·텐센트 워크버디·바이두 다쑤·커우쑤
산업 응용	금융(핑안·공상·우정저축)·제조(창안·바오강·창신치즈)·정부(우시·구이저우·광둥)

정책·규제의 근거 문건을 한자리에 둔다. 발문일과 시행일을 구분해 적었다.

문건	문서번호	발문·발표	시행·비고
국무원 「인공지능+」 행동	국발(2025)11 호	2025-08-21	2027 년 보급률 70%↑ 목표

의견			
15 차 5 개년 계획 강요	—	2026-03 통과	에이전트·피지컬 AI 조항
「인공지능+제조」 실시 의견(8 부처)	—	2026-01-08	2027 년 공업 에이전트 1,000 개
의인화 상호작용 관리 잠행 방법(5 부처)	—	2026-04-10 공포	2026-07-15 시행
인터넷 정보 서비스 관리 방법 개정안	—	의견수렴 ~2026-08-02	'자능정보서비스' 장 신설
에이전트 결제 응용 자율규약	—	2026-08-24	KYA 등 4 대 조항
청량·AI 애플리케이션 정비	—	2026-04 개시	1 단계 성과 2026-07-06 공개

끝으로 이번 호에서 처음 정의한 용어를 정리한다.

용어	뜻
하네스(Harness)	모델 주위에서 도구 조정·작업 상태 관리·폐쇄 루프 실행을 담당하는 제어·오케스트레이션 계층. 딥시크의 'Model + Harness = Agent' 정의가 대표적이다
에이전트 애즈 어 서비스	에이전트를 서비스 형태로 제공·소비하는 발전 모델. 국무원 의견 원문 개념
에이전트 운영체제	CAICT 가 편찬 착수한 연구보고의 3 대 방향 — 실행 프레임워크·임베디드 OS·네이티브 OS, 플랫폼 락인의 형태를 규정한다
KYA(Know Your Agent)	결제 자율규약 제 1 조. 에이전트의 신원과 위임 권한을 확인하는 행위 감사 메커니즘
토큰 파드(Token Pod)	프리필 단계를 독립 공급·과금·최적화 가능한 자원 풀로 분리한 자국산 추론 인프라