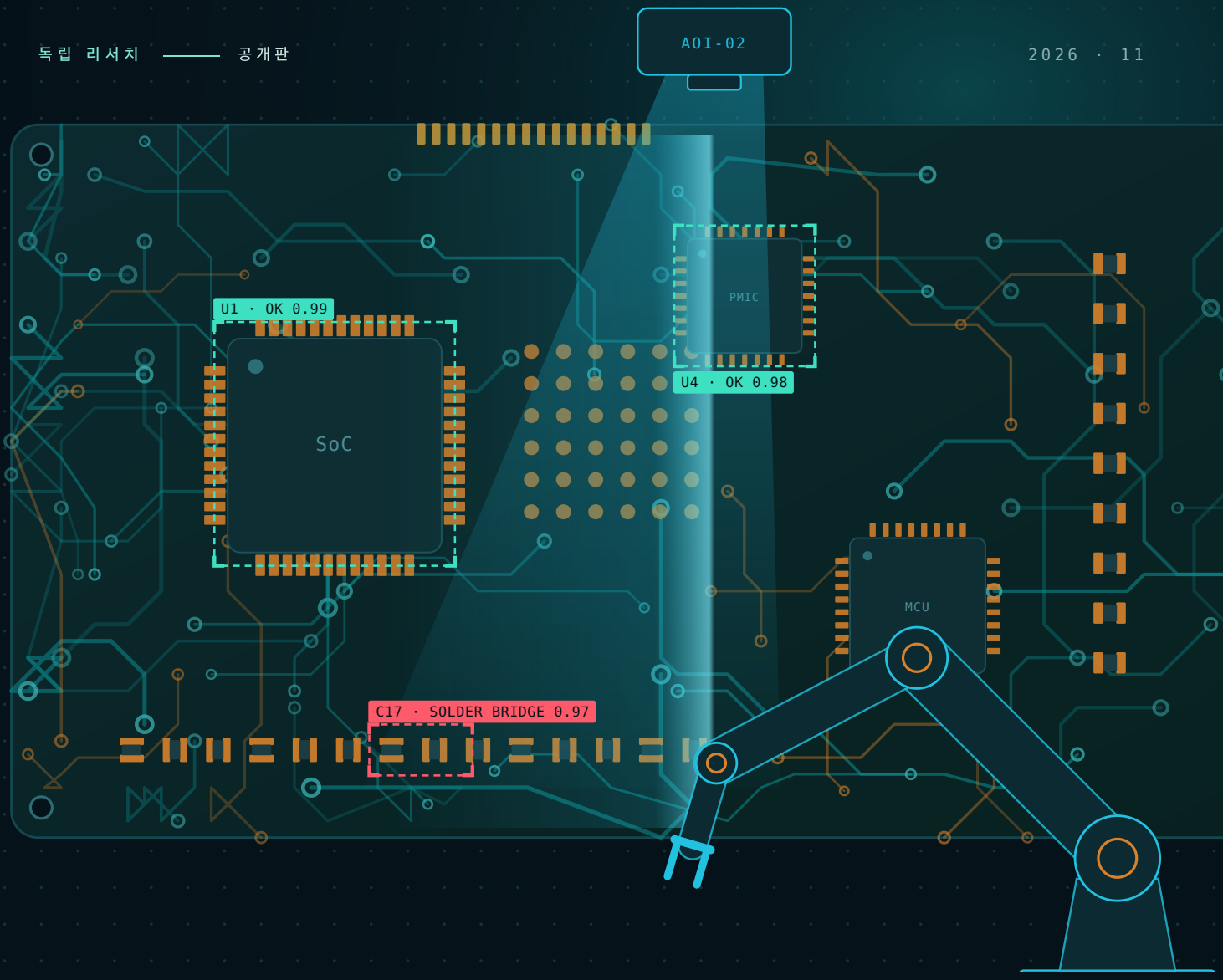


전자 제조 · AI 검사 × 로보틱스

전자 공장의 AI 비전 검사와 로봇 암

2026 평가

주류 구성 · 지역별 실현 가능성 · 온프레미스 배포 · 교사-학생 모델
투자 수익 · 경쟁 · 파일럿만이 답할 수 있는 것

핵심 발견 **AI 재판정은 좁은 범위에서 할 가치가 있다. 임바디드 로봇 암은 기다려도 된다.**

Tan Shuai
tanshuai.comi@tanshuai.com

데이터 기준 2026년 11월 11일
중국 본토 · 대만 · 동남아시아 · 미국

목 차

목차

◆ **요약**
할 가치가 있다, 다만 좁게, 지금, 파트너와 함께 — 4

01 **오늘날의 라인**
오늘날 전자 공장이 돌리는 것, 그리고 기술적으로 실현 가능한 것 — 7

02 **온프레미스**
공장 내부에서 AI 돌리기: 클린룸, 에어갭, 하드웨어, 양자화 — 13

03 **교사와 학생**
교사와 학생: 프런티어 모델은 검사 데이터를 사전 라벨할 수 있지만, 로봇 암의 진짜 교사는 여전히 사람이다 — 21

04 **비용과 회수**
비용과 회수: AI 재판정은 수개월, 머신 텐딩은 수년, 선별 전용 암은 결코 회수 안 됨 — 29

05 **시장과 경쟁**
시장, 경쟁, 그리고 구매자의 지불 방식 — 36

06 **임바디드 AI**
임바디드 AI: 먼저 로봇을 가르치는 방식이 바뀌고, 나중에 움직임이 바뀐다 — 44

07 **미해결 질문**
공개 기록으로 답할 수 없는 것, 그리고 파일럿이 해결하는 법 — 49

◆ **저자 소개**
Tan Shuai — 51

본문의 신뢰도 태그

- **verified** 둘 이상의 독립 출처, 또는 두 건의 독립 1차 문서가 일치함
- **single source** 1차 문서를 포함해 출처가 하나뿐임
- **vendor claim** 제품 또는 서비스의 판매자가 발표함
- **estimate** 인용 데이터에 대한 저자의 계산 또는 추론

이 보고서 읽는 법

핵심 수치는 신뢰도 태그를 달고 있다. **verified** 는 둘 이상의 독립 출처 또는 두 건의 독립 1차 문서가 일치함을 뜻한다. **single source** 는 출처가 하나뿐임(1차 문서 포함)을, **vendor claim** 은 제품·서비스 판매자가 발표함(널리 반복되더라도)을, **estimate** 는 인용 데이터에 대한 저자 자신의 계산 또는 추론(입력값은 출처 명시)을 뜻한다. "1차(primary)"는 출처 유형(연차보고서, 증권신고서, 규정, 표준, 공식 통계)을 나타낼 뿐 신뢰도 등급을 바꾸지 않는다. 출처가 상충하면 수치를 나란히 제시하고 평균내지 않는다. 환율은 2026년 11월 9일 People's Bank of China 기준 환율 $1 = RMB 6.7188(Xinhua)과, 같은 날 SMBC 고시환율 $1 = NT$32.12(SMBC)를 쓴다. 중국 본토 노동·보조금 데이터는 전자 제조가 가장 집중된 Pearl River Delta(공식 임금 데이터는 선전)를 쓰고, 비교용으로 Yangtze River Delta를 쓴다. 업계 인터뷰에서 나온 주장은 사실이 아니라 확인 대상 주장으로 취급한다. 공개 기록으로 답할 수 없는 질문은 7장에 모았다.

그 림

그림



01	오늘날 공장 현장에서 AI가 들어가는 자리	6
02	양산 SMT 라인: 사람은 AOI 뒤에 앉는다	7
03	AOI 알람의 약 97–99%는 오탐	8
04	범용 카메라는 AOI가 아니다	10
05	에어갭 배포	13
06	청정·ESD 구역이 컴퓨트 위치를 결정한다	14
07	용도별 속도 한계	15
08	온프레미스 컴퓨트: 메모리, 가격, 중국 본토에서 합법 구매 가능한 것	17
09	대형 워크스테이션 1대는 공유 어시스턴트가 되지 못한다	18
10	양자화: 8-bit는 안전, 4-bit는 레시피 의존, 3-bit 이하는 탈락	19
11	교사 후보: 범용 점수는 한 세대에 배가	22
12	"결함이 있는가?"는 범용 모델에 여전히 어렵다	23
13	두 개의 교사–학생 파이프라인	24
14	골든 세트는 얼마나 커야 하나	25
15	엣지 정책 지연	26
16	로봇 학생: 스테이션이 구조화될 때만 생산급에 근접	27
17	검사·이송 작업자 1인의 연간 비용	30
18	가격 레더	31
19	지역별 회수 기간	32
20	불량판 선별이 회수되지 않는 이유	33
21	로봇·AOI·엣지 컴퓨터의 미국 관세	34
22	시장 규모는 10배 이상 차이	37
23	보험업: 40,000라인 → 대상 16,000 → 일회성 약 RMB 320M	38
24	개조 윈도우는 약 5년	38
25	재판정 쐐기는 삼면에서 포위	39
26	마진은 제품화된 하드웨어에 있다	40
27	구매자의 지불 방식	41
28	매출채권 회전일수	42
29	임바디드 AI 스택	44
30	공장 내 로봇 배포가 실제로 공개하는 것	45
31	로봇당 순 대체 인력	46
32	표준·규제 타임라인	47

요약

할 가치가 있다, 다만 좁게, 지금, 파트너와 함께

핵심 발견

결론: 전자 공장에 있어 2026년에 성숙했고 공장이 실제로 지불할 것은 AOI 이미지의 AI 재판정과 틈새 검사, 그리고 프로그래밍된 로봇 머신 텐딩이지, "기판을 들여다보는 대형 모델 + 기판을 집어내는 임바디드 로봇"이 아니다.

0.46-0.65 년
라인당 RMB 50k 기준 미국, 대만, 선전의 AI 재판정 회수 기간

≥95% · <20 PPM
선두 EMS 자체 AI 재판정의 커버리지와 유출: 티어 1 EMS 판매 기준

100-700 백만 RMB
중국 본토의 서드파티 개조 대상 시장, 일회성, 약 5년 윈도우

0 점수
2026년 11월 플래그십의 공개 산업 결함 벤치마크 점수: 검사원이 아니라 교사

1.1-1.5 년
2교대 협동로봇 머신 텐딩 셀의 회수 기간 (미국, 대만, 선전)

0.003-0.14 FTE
SMT 라인에서 선별 전용 로봇 암이 교대당 절감하는 인력: 결코 회수되지 않음

검사
ROI

기존 AOI가 이미 만들어내는 이미지에 지도학습 딥러닝을 적용하면 오탐 또는 재판정 공수를 60–90% 줄인다(벤더 주장). 대형 EMS인 DBG는 연차보고서에서 자체 시스템이 플래그된 부품의 95% 이상을 커버하고 유출은 20 PPM 미만, 인력 재판정은 85% 이상 줄였다고 밝힌다. 이것이 티어 1 EMS에 판매하려면 서드파티 제품이 넘어야 할 기준이다. 라인당 RMB 50,000 기준으로 AI 재판정은 중국 본토, 대만, 미국에서 1년 안에 회수되며(선전 0.65년, 대만 0.53, 미국 0.46), 협동로봇 머신 텐딩 셀은 1.1–1.5년, 동남아시아는 2–6년이 걸린다. 반면 불량 기판을 골라내기 위해 SMT 라인에 추가된 로봇 암은 어디서도 결코 회수되지 않는데, 직통률 90% 이상이면 선별에 드는 인력이 교대당 0.003–0.14명에 불과하기 때문이다 estimate .

모델
로봇

멀티모달 대형 모델은 여전히 주 검사기가 될 수 없다. 2026년 11월 11일 현재 플래그십(Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 등) 중 어떤 산업 결함 또는 PCB 벤치마크에서도 공개 점수를 가진 모델은 없으며, ACM MM 2026 PCBA 챌린지에서 실제 라인 이미지 6,600장으로 학습된 상위 5개 시스템도 "결함이 있는가"에서 77–82점에 그쳤다. 프런티어 모델은 사전 라벨과 트리아지(분류)는 할 수 있다. 로봇 암의 진짜 교사는 여전히 사람 시연, 실기 강화학습, 시뮬레이션이다. 학생 모델은 지그가 있는 구조화 스테이션에서 90–99.4%에 도달하지만, 이동 컨베이어·미확인 부품·신제품에서는 30–80%에 그친다. 프런티어 에이전트는 준정적 실기 로봇 과제 108개 중 104개를 완수했지만 과제당 280초가 걸렸다.

온프레미스
배포

청정·통제된 공장에서의 에어갭 배포는 충분히 실용적이다. 일반 SMT 홀은 대개 등급 클린룸이 아니라 ESD 보호 구역이다. 라인에는 무팬 Jetson 또는 국산 NPU 박스를 두고, GPU 서버는 서버실에 둔다. NVIDIA 자체 수출 분류상 Jetson Thor와 DGX Spark는 미국 라이선스 없이 일반 중국 본토 공장에 판매할 수 있지만, DGX Station은 불가능하다. 정밀도 기본값은 8-bit이며, 벤더의 양자화 인지 4-bit는 조건부로 사용 가능하고, 3-bit 이하는 탈락이다.

시장
경쟁

시장은 실재하지만 제한적이다. 중국 본토에는 약 35,000–46,000개의 SMT 라인과 AI를 돌리지 않는 인라인 AOI 약 26,000–66,000대가 있다. 서드파티 개조의 대상 시장은 일회성 약 RMB 100–700 million이며, 윈도우

는 약 5년이고, AOI 제조사들은 신규 장비에 재판정 기능을 내장하기 시작했다 estimate. 경쟁은 이미 치열하다. 재판정 쐐기는 AOI 제조사, EMS 사내 팀, 플랫폼 벤더에게 포위되어 있고, 마진은 제품화된 하드웨어에 있다.

질문	판단	핵심 수치 (출처와 신뢰도는 각 장에)
① 오늘날 주류 전자 공장은 무엇을 돌리는가?	양산 SMT 라인은 "3D SPI + 리플로우 전·후 AOI + 스폿 X-ray + ICT/FCT + 수리 및 육안 검사"이며, 모든 AOI 뒤에 인력 재판정이 있다. AI는 세 가지 경로로 재판정에 들어온다: AOI 제조사 내장, 선두 EMS 사내 구축, 서드파티 추가. 조립과 핸들링은 학습된 인지와 프로그래밍된 모션을 갖춘 암, SCARA, AMR을 쓰며, 임바디드 AI는 여전히 파일럿 단계다	택트 기판당 ≈13–22초; 한 ODM이 18개 라인에 AOI 예산 RMB 30.6M을 책정, 라인당 ≈3–5대; AOI 알람의 ≈97–99%는 오탐
② 기술은 어디까지 가고, 돈으로 "인력 0명"을 살 수 있으며, 각 지역은 무엇을 쓸 수 있는가?	검사는 "전용 검출기 + AI 재판정 + 소수 인력"으로 산업급이다. 무제한의 돈으로도 인력 0명에 도달하지 못한다. 중국 본토는 국산 모델 또는 로컬 오픈 웨이트만 쓴다. 대만의 민간 공장은 거의 무엇이든 쓸 수 있지만, 팹과 중요 인프라는 사실상 중국 본토 브랜드를 배제한다. 동남아시아는 제약이 가장 적고, 미국은 풀스택에 관세와 인증이 더해진다	AI 재판정은 오탐 또는 재판정 공수를 60–90% 절감 vendor claims ; 기판 전체를 보는 4K 카메라는 픽셀당 ≈78 µm인 반면 AOI는 8–20 µm; 중국 본토산 암은 미국 종합 관세 ≈40%, AOI 계열 기기는 ≈37.5%
③ 클린룸과 에어갭 공장에서 온프레미스로, 어떤 하드웨어로, 충분히 빠르고 정확하게 돌릴 수 있는가?	그렇다. 제약은 컴퓨트가 아니라 절차다. 라인에는 ≤100 W 무팬 박스, GPU 서버는 서버실에 둔다. 스테이션에는 Thor, 공장 Q&A에는 Mac Studio 또는 국산 어플라이언스. 중국 본토에는 DGX Station 불가. 모델 크기가 과제에 맞으면 속도는 충분하다. 8-bit 유지, 벤더의 양자화 인지 4-bit는 조건부 사용, 3-bit는 절대 금지	Thor에서 8B 비전 모델은 크롭당 ≈0.5초 estimate ; Thor 개발 키트 China에서 RMB 59,104; DGX Station $123–140k, 743B 모델에서 32명 사용자 총 139 tokens/s
④ 최신 플래그십이 수십 라운드에 걸쳐 교사 역할을 하고 로컬 학생으로 증류될 수 있는가(유연 생산과 로봇 암 포함)?	검사: 사전 라벨과 트리아지. 골든 세트에서 "결함 ≥299개, 미검출 0"인 클래스에만 교사 라벨을 자동 채택. 로봇: 교사는 여전히 사람 시연, 실기 RL, 시뮬레이션이며, 학생은 구조화 스테이션에서만 생산급에 근접한다	11월 플래그십 중 산업 결함 점수를 가진 모델 없음; 학생은 학습된 커넥터에서 97.8%, 미확인 커넥터에서 29–67%, 이동 컨베이어에서 37–59%; 프런티어 모델이 관절을 직접 구동 시 0–5.5%
⑤ 비용은 얼마이고, 지역별 ROI는?	AI 재판정은 중국 본토, 대만, 미국에서 1년 내 회수. 협동로봇 텐딩 1.1–1.5년, 동남아시아 2–6년, SMT 라인의 선별 전용 암은 결코 회수되지 않는다	라인당 RMB 50k 재판정 박스: 선전 0.65년, 대만 0.53, 미국 0.46, 베트남 2.2 estimate ; 수용 가능한 서드파티 추가 가격 ≈라인당 RMB 10–30k 일회성 또는 연 RMB 5–10k
⑥ 경쟁자는 이미 충분히 강한가, 지금 시작하는 게 나은가 기다리는 게 나은가?	수요는 포화되지 않았지만 공급은 붐빈다. OEM AI가 개조 윈도우를 약 5년으로 제한한다. 플랫폼 벤더와 정면으로 싸우지 말고 채널 파트너를 통해 좁은 제품으로 지금 진입하라. 임바디드 AI에 베팅하기 전에 2027–2028년 생산 데이터를 기다려라	중국 본토에서 AI 미적용 인라인 AOI ≈26–66k대, 연 8–12% 축소 estimate ; 선두 AI 비전 벤더의 2025년 매출 RMB 1.086B, 순손실 RMB 991M; 가장 가까운 상장 비교 기업의 3C 매출은 2026년 1–5월 전년 동기 대비 79.9% 감소

FIG 01 **오늘날 공장 현장에서 AI가 들어가는 자리*AOI 오탐의 AI 재판정과 협동로봇 설비 로딩은 성숙해 빠르게 회수된다. 전용 선별 암과 LLM을 주 검사기로 쓰는 것은 할 가치가 없다.***



출처: 회수는 저자 모델(제4장); 성숙도 축은 저자 판단.

무엇을 만들고 무엇을 만들지 말아야 하나

	권고
만들 것	구형 또는 복수 브랜드 AOI를 갖고 AOI 제조사의 AI 업그레이드가 없는 중견 EMS를 위한 AOI 이미지 AI 재판정·결함 분류; SMT AOI가 잘 커버하지 못하는 검사(스루홀 및 수납땜, 컨포멀 코팅, 케이블·커넥터·FPC, 최종 외관·나사); 모든 부품을 이송하는 협동로봇 로딩·언로딩 셀(테스트 지그, 트레이, 포장)에 검사를 애드온으로; 모든 주장의 근거로 실제 라인에서의 골든 세트와 섀도 모드
만들지 말 것	범용 "AI + 로봇 암" 플랫폼; 주 검사기로서의 멀티모달 LLM; NG 기판 선별 전용 SMT 라인 로봇 암(NG 분류기나 버퍼로 충분); 휴머노이드와 로봇 서비스(RaaS); 제어 루프나 안전 기능 내부의 프런티어 모델; "풀사이즈" 모델을 돌리기 위한 스테이션의 GPU 메모리 더미; 팹, 대만 중요 인프라, 미국 방산 전자 공장에 대한 직접 판매
함께 일할 대상	AOI 제조사와 유통사(채널·이미지 인터페이스; 유통사가 Sinic-Tek 2025년 매출의 46.68%를 차지); 협동로봇 제조사(인증 암과 안전 기능); 10–40개 라인을 가진 EMS(단계적 확장과 예치금); Thailand 등의 현지 인테그레이터(BOI 로컬 연계 규정 충족용)

01 오늘날 전자 공장이 돌리는 것, 그리고 기술적으로 실현 가능한 것

오늘날의 라인

표준 양산 라인: AOI 여러 대, 스폿 X-ray, 중앙 재판정, 기판당 13–22초

상장 ODM의 1차 공시는 2025–2026년 양산 라인이 실제로 어떤 모습인지 보여준다. Longcheer의 홍콩 증권신고서(데이터 기준 2025년 9월 30일)는 SMT 품질 검사를 "X-ray 검사, 솔더 페이스트 검사(SPI), 자동 광학 검사(AOI), PCBA 기능 테스트, 육안 검사"로 나열한다. 그 공정 흐름은 리플로우 전 AOI, 리플로우 후 AOI, 스폿 X-ray, NG 기판의 SMT 수리 스테이션, 그다음 최종 육안 검사 순이다. Huizhou, Nanchang, 베트남 공장은 SMT 라인 34, 24, 8개와 최종 조립 라인 67, 60, 13개를 갖추고 있다(Longcheer H-share prospectus) single source, primary filing. Huaqin의 2023년 노트북 공장은 18개 SMT 라인에 AOI 예산 RMB 30.6 million, 라인당 약 RMB 1.7 million을 계획했는데, 국산 3D AOI 가격 기준으로 양면 기판의 리플로우 전·후 검사를 위해 라인당 3–5대에 해당한다(Huaqin A-share prospectus) primary filing + estimate. 따라서 양산 라인에는 보통 재판정 지점이 하나가 아니라 2–4개 있다.

장비는 이미 고도로 자동화되어 있다. Huaqin은 SMT 라인의 자동화율이 96%라고 보고한다(Huaqin H-share prospectus) single source. 남은 인력은 재판정, 수리, 육안 검사, 테스트, 최종 조립에 있고, 어떤 회사도 라인당 재판정자 수를 공개하지 않는다(7장).

세 개의 독립 추정이 택트 **기판당 13–22초**(5개 Hebei 라인의 2026년 환경 공시(Hebei SMT project EIA), Huaqin의 노트북 라인, iFactory 사례)에 일치하며, 양면 기판은 면당 7–11초이다 estimate. 하이믹스 라인의 어려운 점은 물량이 아니라 신제품 도입(NPI)과 전환(changeover)이며, 모든 전환은 AI 모델에게 또 한 번의 콜드 스타트다.

소프트웨어 측면에서 AOI 제조사의 자체 재판정 스테이션은 이미 중앙화되어 있다. Sinic-Tek의 수리 스테이션 소프트웨어는 "호스트 한 대로 SMT 라인 8개의 장비를 제어"할 수 있게 한다(Sinic-Tek prospectus) single source, primary, 2022. Jabil이 2026년 ViTrox V-ONE 원격 AOI 제어를 도입한 뒤 엔지니어 한 명이 로컬 1–2대 대신 원격으로 AOI 5–8대를 관리할 수 있게 되었다(AtlasPCB) vendor claim. 서드파티 AI 재판정 제품은 바로 이 중앙 스테이션에, AOI의 이미지·결과 내보내기(IPC-CFX, IPC-HERMES-9852 또는 벤더 자체 포맷)를 통해 연결된다. AOI 자체도 빠르게 AI를 추가하고 있다. Yamaha는 자사 AI 판정이 호출당 약 1초가 걸리고, NG 이미지 30장이면 검사원을 돕기 시작하며, 1,000장이면 "매우 높은" 품질이 된다고 말한다(DirectIndustry/Yamaha) vendor claim.

FIG 02 **양산 SMT 라인: 사람은 AOI 뒤에 앉는다** ***자동 광학 검사 후 인력은 재판정 자리에 집중된다. AI가 들어가는 곳이 바로 그곳이다.***



출처: 공시 자료와 벤더 자료(Huaqin, DBG, Sinic-Tek 등); 저자 편집.

단계	주류 제품과 기술	대표 공급사	AI의 진척도
인쇄 후	3D SPI(구조광, 모아레, 일부 레이저 삼각측량)	Koh Young, Sinic-Tek, TRI	OEM AI 페이스트 판정이 양산 중; Jutze는 일부 모델이 "완전 AI"라고 주장
실장·리플로우 후	다각도 다색 LED와 구조광을 갖춘 2D/3D AOI, 픽셀당 8–20 µm, 8–12 MP 카메라	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek 외	OEM AI 재판정(Jutze 양산, ViTrox V510Ai, Sinic-Tek은 Tencent와 공동); 선두 EMS 자체 구축(DBG); 서드파티 애드온
바텀 터미네이션·히든 접합	스폿 또는 인라인 X-ray(AXI)	—	Hikvision은 Guanlan 모델이 X-ray 보이드를 1% 미만으로 오판한다고 주장 vendor claim
재판정·수리	중앙 재판정 스테이션(호스트 1대, 8라인), 원격 재판정, 수리 스테이션	AOI 제조사 소프트웨어, ViTrox V-ONE	AI가 오탐을 제거하고 나머지는 사람이 처리
육안 검사, IQC, FQC	대부분 수작업; 일부 로봇 로딩을 갖춘 AI 외관 장비	SmartMore 외	외관 장비는 양산 중; IQC/FQC 인원 공개 데이터 없음
핸들링·머신 텐딩	6축 암과 협동로봇, SCARA, AMR, 2D/3D 비전 가이던스	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	학습된 인지 + 프로그래밍된 모션이 양산 중; VLA와 휴머노이드는 여전히 파일럿
데이터·시스템	MES, IPC-CFX, IPC-HERMES-9852	—	결과를 MES에 기록; 구형 장비의 지원은 불균일

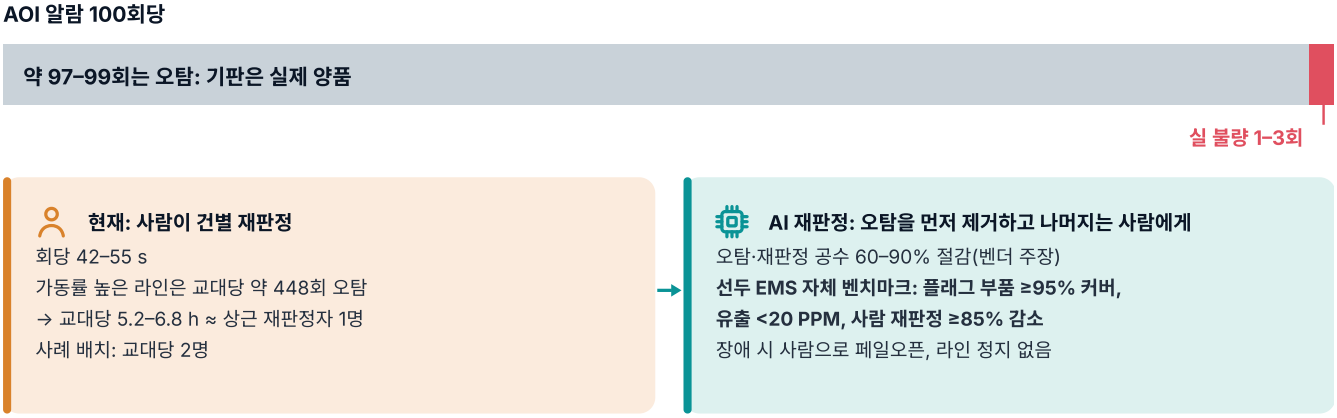
표의 출처: 위 공시 및 (Koh Young, Jutze 2025 annual report, Sinic-Tek 2025 annual report, ViTrox, Sohu on Hengtianyi, Xinhua on Hikvision).

사람은 AOI 뒤에 앉는다. 대부분의 알람이 오탐이기 때문이다

주류 SMT/PCBA 라인의 검사 체인은 다음과 같다. 인쇄 후 3D SPI로 누락·부족·과다 페이스트, 브리지, 오프셋을 보고, 실장 후(선택)와 리플로우 후 AOI로 부품 누락·이동·회전·반전·묘비(tombstone)와 솔더 볼륨을 본다. BGA 등 바텀 터미네이션 부품이 있으면 X-ray를 쓰고, 그다음 인서킷 또는 플라잉 프로브 테스트, 기능 테스트, 수동 육안 검사 순이다. 모든 AOI 뒤에는 재판정 자리와 수리 스테이션이 있다(Sinic-Tek prospectus, JLCPCB) verified. 리플로우 후 AOI가 가장 흔하고 가성비가 좋지만, 형태와 면적만 본다. BGA 아래의 접합, 보이드, 젖음성은 X-ray나 전기 테스트가 필요하다(JLCPCB).

인력이 재판정에 집중되는 이유는 대부분의 알람이 오탐이기 때문이다. 한 UK EMS는 부품 번호 917개에 걸쳐 **오탐 54,019건 대 실제 결함 544건, 알람의 약 99%**를 기록했다(arXiv 2505.03576) single source, preprint. Siemens의 Rastatt 공장에서는 **AOI가 불합격시킨 기판의 97.5%가 양품**이었다(Siemens case study) vendor claim. 가동률 높은 라인은 교대당 오탐 약 448건이 나오며 건당 42–55초, 교대당 5.2–6.8시간, 대략 상근 재판정자 1명에 해당한다(iFactory) vendor claim; hours are an estimate.

FIG 03 ***AOI 알람의 약 97–99%는 오탐 현재는 사람이 모든 알람을 확인한다. AI 재판정이 먼저 오탐을 제거하고 나머지를 사람에게 넘긴다.***



출처: Siemens 사례, iFactory(벤더 주장), DBG 2025 연차보고; 재판정 공수는 저자 추정.

"교대당 재판정 시간 = 기판 수 × 기판당 오탐 × 건당 초"로 계산하면, 규칙 기반 AOI에서 재판정자 1명이 커버할 수 있는 라인은 약 0.7–1개다. AI가 오탐의 60–90%를 제거하면 판정 시간만으로는 인당 2–14개 라인이 가능하지만, 이동, NG 버퍼 처리, 수리 인계가 로컬 커버리지를 인접 라인 2–4개로 제한한다. 중앙 또는 원격 재판정만이 이론적 상한에 가까워진다 estimate; seconds per call come from one source. 주류 AOI는 초당 60–80 cm²를 스캔하므로(Yunheng 2026 AOI guide) 20×15 cm 한 면은 4–5초가 걸린다. AI는 기판을 다시 스캔할 필요 없이 하나의 택트 안에서 플래그된 소수 영역만 처리하면 된다.

이미 양산에서 작동하는 것: AOI 자체 이미지의 지도학습 분류

검증된 패턴은 공장이 이미 보유한 AOI의 이미지에 대한 지도학습 딥러닝 분류다. Fujitsu가 대만에서 ADLINK의 AOI에 추가한 애드온은 오탐을 **90% 이상** 줄였다(Fujitsu case). OKI는 2026년 7월부터 자체 EMS 사업에서 AOI 이후 수동 확인을 약 **80%** 줄이는 데 AI를 쓰고 있다(Circuits Assembly). Siemens는 수동 분석 60% 감소, 첫해 €160,287 절감, 회수 8개월을 보고한다(Siemens case study) all vendor claims. 이 사례들에 공통된 두 가지가 있다. 거의 아무도 유출률을 공개하지 않고, 모두 남은 알람에 사람을 남겨둔다는 점이다. 유출 수치를 제공하는 유일한 사례(0.8% → 0.3%, 계약 기준 0.5% 대비)는 내부적으로 모순된다(iFactory). AI는 재판정 공수를 줄이지, 재판정이라는 일 자체를 없애지 않는다.

AI 재판정은 서드파티 애드온에서 OEM 기능과 사내 시스템으로 이동했고, 이는 신규 진입자가 무엇과 비교 평가되는지를 바꾼다. Jutze의 2025년 연차보고서는 "3D AOI 업그레이드"가 양산 중이며 "검사 단계에서 수동 재판정을 제거하는 AI 알고리즘을 쓴다"고 밝힌다(Jutze 2025 annual report). Sinic-Tek은 2019년부터 "멀티모달 AI 보조 수동 재판정 시스템"을 개발했고, Tencent와 상하 동시 3D AOI를 공동 개발 중이며 결함 커버리지 99.5%+, 인식 정확도 98%를 목표로 한다(Sinic-Tek 2025 annual report) vendor claims in primary filings. 대형 EMS는 자체 구축한다. DBG의 2025년 연차보고서는 "육안 검사용 AI 판정"을 완료 프로젝트로 기재하며, "AOI가 플래그한 부품의 95% 이상을 커버", "유출률 20 PPM 미만", "수동 재판정 85% 이상 감소", 완전 추적 가능이라고 밝힌다(DBG 2025 annual report) single source, primary. 이것이 커버리지, 유출, 절감을 함께 제공하는 전자 공장 수치 중 유일하게 찾은 것이다. **티어 1 EMS에 판매되는 서드파티 제품은 DBG의 95% 커버리지, 20 PPM 유출, 85% 절감을 맞춰야 한다.** 티어 1 미만에서는 벤더가 광고하는 50–85% 절감이 기준이다.

양품으로만 학습한 이상 탐지는 빠르지만(EfficientAD-S는 이미지당 2.2 ms(EfficientAD, WACV 2024)) 누락·오부품 같은 "논리적" 이상에는 약하고, 최고 방법론도 더 어려운 MVTec AD 2에서 평균 AU-PRO 60% 미만이다(MVTec AD 2) single source. 누락·오부품은 CAD와 BOM으로 확인하는 편이 낫다.

어떤 비용에서도 천장: 장비의 겹겹, 극소수 인력, 그러나 0명은 아니다

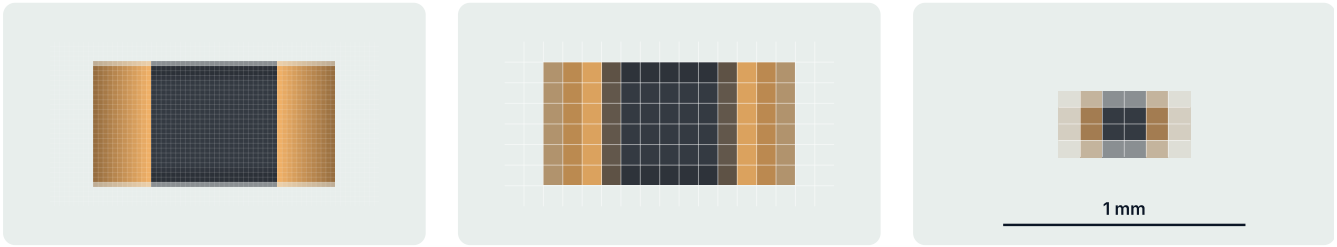
수동 재판정을 완전히 없애고 유출 0에 도달한 전자 공장은 찾지 못했다. IPC-A-610 Class 3(고신뢰성) 제품 관행은 자동 임계값이 분류할 수 없는 경계 판정을 숙련 검사원이 내리는 것이며, 한 PCB 조립 업체는 "자동 검사가 사람의 서명을 대체할 수 있다"는 것을 "흔한 오해"라고 부른다(PCBCart) single source. 사람 기준선도 보정이 필요하다. 2012년 Sandia 연구를 인용하면, 육안 검사원 1명은 최선의 경우 결함의 약 **80%**를, 2명이 직렬로 보면 약 96%를 잡는다(Instrumental) single source, cited by a vendor. 방어 가능한 주장은 "AI + 소규모 인력 재판정 루프가 전수동 검사보다 낫다"이지, "AI가 검사를 대체한다"가 아니다.

예산 제한이 없다면 천장은 3D SPI, 3D AOI, X-ray, AI 재판정, AI 외관 스테이션을 겹겹이 쌓고, 경계 판정·신제품·전환·신규 결함 유형·고객 서명에만 사람을 남기는 것이다. 플래그십 공장들은 수동 검사 85–92% 감소를 보고한다(I-Connect007, 2021, WiAdvance) vendor claims. 예산 무제한 천장에 대한 저자의 추정은 재판정 인력이 현재의 8–15%라는 것이다 estimate. 최종 조립은 여전히 노동 집약적이다. iPhone 16 버튼 자동화는 높은 결함률 때문에 중단되었다(9to5Mac) single source.

범용 고해상도 카메라는 AOI가 아니다

양산 AOI는 타일 단위로 픽셀당 8–20 µm, 8–12 MP 카메라, 프로그래머블 다각도 다색 조명, 구조광 3D로 촬영한다(Koh Young, Saki) verified. 업계는 오래전에 01005 부품이 25 µm/픽셀에서 약 8×16 픽셀에 불과해 신뢰성 있는 판정에 "충분하지 않을 수 있다"고 지적했다(Circuits Assembly, 2007). 같은 계산으로 estimate, 300 mm 기판 전체를 4K 카메라로 보면 픽셀당 약 **78 µm**가 된다. 인치계 0402 부품은 약 13×6 픽셀로 유무와 큰 오프셋은 가능하지만 접합 형태는 불가하고, 01005는 약 5×3 픽셀로 사실상 보이지 않는다. 기판 전체에서 15 µm/픽셀에 도달하려면 약 330메가픽셀이 필요한데, 그래서 AOI는 카메라를 기판 위로 움직인다. 누군가 "0402"라고 하면 어느 쪽을 말하는지 확인하라. 미터계 0402는 인치계 01005로, 어떤 광각 카메라로도 한참 벗어난다.

FIG 04 **범용 카메라는 AOI가 아니다** ***300 mm 기판을 4K 카메라로 보면 약 78 µm/px. 0402는 약 13×6 px로 줄고 01005는 사라진다.***



AOI: 8–20 µm/px
0402는 약 100×50 px

4K 카메라로 300 mm 기판 전체: 약 78 µm/px
0402는 약 13×6 px로 축소: 유무와 큰 오프셋만 판별 가능

같은 카메라로 01005 관찰
약 5×3 px: 사실상 안 보임

출처: Circuits Assembly; 300 mm 기판·4K 센서 저자 계산. 픽셀 그리드는 예시.

이미 AOI가 있는 라인에서 AI의 최적 위치는 AOI 이미지 위다. AOI가 없는 라인이나 라인 끝(EOL) 스테이션만이 제어된 조명의 밀폐 카메라 스테이션을 정당화한다. 라인 위의 개방형 고해상도 카메라는 주변광과 그림자에 가장 취약하다(CSDN).

조립과 핸들링: 학습된 인지 + 프로그래밍된 모션이 양산 중

IFR의 2026년 보고서는 "검사 및 로봇 가이던스용 머신 비전은 성숙했고 산업급"이지만, "범용 피지컬 AI와 시뮬레이션·경험 기반 학습은 아직 산업급이 아니다"라고 말한다(IFR World Robotics 2026 executive summary) single source, primary. 전자는 최대 로봇 구매처다(2025년 전 세계 설치 142,235대, 중국 설치의 59%(IFR, IFR China release)). 그러나 중국은 제조 노동자 1만 명당 로봇 166대인 반면 미국은 307대다(IFR, April 2026). 양산의 패턴은 학습된 인지 + 결정론적 모션이며, 엔드투엔드 학습 모션은 구조화 스테이션에서만 생산급에 근접한다(3장, 6장).

각 지역이 쓸 수 있는 것

중국 본토는 국산 모델과 로컬 웨이트에 의존하고, 대만은 고객의 정체성이 결정하며, 미국은 관세와 인증이 결정한다.

요소	중국 본토	대만 (민간 공장)	동남아시아	미국
Claude (모든 경로)	● 불가	● 조건부: 공급자가 중국 본토, 홍콩, Macau의 인물·법인이 과반 지분을 보유하거나 통제하지 않아야 함	● 동일	● 동일
GPT와 Gemini 클라우드	● 불가	● 가능	● 가능(국가별 목록 미확인)	● 가능
중국 클라우드 모델 (Qwen, DeepSeek, Doubao, GLM, Kimi)	● 가능(등록된 모델; 서비스 제공에는 등록 필요)	● 합법이나 중국 본토 호스팅은 고객 보안 심사를 통과하는 경우가 드물; 공공 부문·국영 기업·중요 인프라에서 금지	● 합법(국제 노드)	● 합법, 엄격한 심사
중국산 로컬 오픈 웨이트 (Qwen, GLM, DeepSeek)	● 1순위; 공장 내부 사설 배포에는 모델·알고리즘 등록 불필요(로펌 해석)	● 가능(공공 부문 제한)	● 가능	● 조건부: 원산지 심사
비중국산 로컬 오픈 웨이트(예: Gemma 4)	● 가능	● 가능	● 가능	● 가능
전통적 비생성형 비전 모델	● 가능(생성형 AI 규칙 밖)	● 가능	● 가능	● 가능

요소	중국 본토	대만 (민간 공장)	동남아시아	미국
Jetson Thor/T5000/T4000, DGX Spark	가능: NVIDIA가 5A992.c로 분류, 일반 민간 공장에는 미국 라이선스 불필요; Spark는 China SKU 존재; 채널 가격은 미국 리스트 대비 약 40–60% 높음	가능	가능	가능
RTX 5090, RTX PRO 6000, DGX Station	불가: 처음 둘은 4A090.a, DGX Station은 5A992.z; 합법 대안은 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20(4A994.l)	가능	가능하나 본사가 중국이거나 최종 중국 소유 구매자는 중국인으로 간주; 말레이시아는 수출·환적 허가 필요	가능
Mac Studio M5 Ultra 512 GB	가능(Apple China 스토어)	가능	가능으로 추정	가능
Huawei Ascend 및 기타 중국 가속기	가능	고위험	고위험	사실상 사용 불가
중국 본토산 고정 암, AOI, 산업용 카메라	가능	수입 가능; 2025년 12월 1일부터 당국이 중국 본토 브랜드에 직접 조치 가능, 기관·국영 기업·일부 재단·중요 인프라 공급자는 사용 제한 또는 금지 가능	가능	합법이나 암 종합 관세 약 40%, AOI 계열 기기·카메라 37.5%, 엣지 컴퓨터 25%
외국산 AMR과 모바일 매니퓰레이터	가능	가능	가능	신모델은 US Department of War 조건부 승인 필요
Hikvision·Dahua 카메라	가능	민간 사용 허용, 공공 부문 금지	가능	신규 FCC 인증 차단; 연방 계약자 금지
로봇 안전·인증	GB 11291.1-2011·GB 11291.2-2013 강제, 2026년 4월부터 개정 중; CR 인증은 자율; GB/T 39404-2026이 컨트롤러 사이버보안 커버	기본은 펜스; 펜스리스 협동은 CNS 14490 또는 ISO 10218 + 서면 평가 충족 필요	고객·수출 대상국별, 선적 전 서면 확인	전기 장비는 "승인"(실무상 NRTL 마크 또는 현장 평가) 필요; UL 1740은 2025년 8월 11일부터 로봇 셀이 ISO 10218-2 충족 요구
팹·통제 제품 공장	—	TSMC 구매 계약은 SEMI E187 요구; 2 nm 팹에서 중국 본토 장비 제거 중	—	ITAR 기술 데이터에 "사진" 포함; 중국에 대한 거부 정책; 중국 본토 통제 공급자는 사실상 방산 전자 공장에 서비스 불가

표의 출처: (Anthropic supported regions, OpenAI supported countries, Gemini availability, CAC interim measures on generative AI, NVIDIA export classification data, eCFR 15 CFR 774, NVIDIA China DGX Spark, Apple China, Covington on new BIS rules, Greenberg Traurig, June 2026, Fulcrum, Tom's Hardware on Ascend guidance, 대만 BOFT import list, 대만 Cyber Security Management Act history, 대만 review rules for products endangering national cybersecurity, Taipei Times, USITC HTS chapter 99, Latham & Watkins, IPVM, China national standards platform, GB 11291.2, Intertek UL 1740 update, eCFR 22 CFR 126.1).

중국 본토: 풀스택이지만, 폐쇄형 프런티어 모델도 가장 강한 단일 장비도 들어올 수 없다. 등록된 국산 클라우드 모델, 로컬 오픈 웨이트(Qwen은 Apache-2.0, GLM·DeepSeek은 MIT), 전통 비전 모델, 국산 AOI·카메라·협동로봇은 모두 사용 가능하다. NVIDIA의 자체 수출 분류는 Jetson AGX Thor 키트, T5000/T4000 모듈, DGX Spark를 5A992.c로 기재하므로 일반 EMS·PCBA 공장은 미국 라이선스가 필요 없지만, RTX 5090과 RTX PRO 6000(4A090.a), DGX Station(5A992.z, TPP 58,484)은 중국 본토에서 구매할 수 없다(NVIDIA export classifications, eCFR) primary; the legal reading is mine and not legal advice. RS China는 Thor 키트를 세금 포함 RMB 59,104, 미국 리스트 가격의 약 1.4배로 기재한다(RS China).

4개 로펌이 공장 자체 사용을 위한 사설 온프레미스 배포에는 모델·알고리즘 등록이 필요 없지만, 호스팅형 멀티테넌트 서비스는 회색지대라는 데 일치한다(Covington, FPF, JunHe) verified: consistent readings, no official ruling. GB 11291.1-2011과 GB 11291.2-2013은 여전히 강제 로봇 안전 기준선이며 2026년 4월 28일부터 개정 중이다. 컨트롤러 사이버보안에 관한 GB/T 39404-2026은 2026년 11월 1일 시행됐다(revision plan 20262542, GB/T 39404-2026) single source, primary. 새 GB 11291이 나오기 전까지(2028–2029년 estimate) ISO 10218:2025에 맞춰 설계하고 GB 11291 준수 서류를 유지하라.

대만: 민간 공장은 거의 무엇이든 쓸 수 있다. 고객의 정체성이 결정한다. 프런티어 API, 오픈 웨이트, 중국 본토 장비는 모두 합법이다(data.gov.tw). 그러나 개정된 Cyber Security Management Act가 **2025년 12월 1일** 시행된 뒤 당국은 "중국 본토 브랜드" 제품에 직접 조치할 수 있고, 정부 기관·국영 기업·중요 인프라 공급자에서 제한할 수 있다(Laws & Regulations Database, review rules) verified. 중요 인프라 공급자 목록은 비공개이고, TSMC는 SEMI E187을 구매 계약에 명시했으며(CNA) 2 nm 팹에서 중국 본토 장비를 제거 중이라고 보도된다(TrendForce) single source. 민간 EMS·조립·소형 PCBA 공장은 열려 있지만, 팹·단지 유틸리티·정부 관련 고객은 사실상 닫혀 있다. 중국 본토 통제 공급자는 E187에 맞춰 설계하고, 소프트웨어 BOM을 제공하며, 오프라인 업데이트를 지원하고, 중국 본토에서 원격 접속을 하지 말아야 한다.

동남아시아: 제약이 가장 적다. 문제는 느린 회수와 중국 모기업이다. 모든 것을 쓸 수 있지만, 2026년 5월 31일 미국 지침은 수취인이 제3국에 있더라도 중국에 본사를 두거나 궁극적으로 중국 소유인 법인에 첨단 컴퓨팅 품목을 수출하려면 라이선스가 필요하다(Greenberg Traurig) single source. 말레이시아는 첨단 칩에 자체 허가가 있다(Fulcrum). 수요는 이쪽으로 이동 중이다. TrendForce는 2026년 Thailand, 베트남, 말레이시아가 서버 ODM SMT 캐파의 50%에 근접할 것으로 예상했다(TrendForce) single source, a 2023 forecast.

미국: 관세·인증·통제 데이터라는 대가로 풀스택. 미국 관세표 제99장에 따르면 중국 본토산 로봇 암은 종합 관세 약 40%, AOI 계열 기기·카메라는 약 37.5%, 엣지 컴퓨터는 25%를 부담한다. 말레이시아·Indonesia산은 약 10–12.5%, 베트남·Thailand·필리핀산은 12.5–15%, 대만산은 10%, 컴퓨터는 중국 본토 외 지역에서 0%다(USITC HTS chapter 99, Great Lakes Customs Law) primary schedule + estimate. USTR의 11월 10일 공지는 기존 Section 301 예외 178건을 2027년 1월 9일까지 연장했으며, 로봇·검사 기기·카메라·컴퓨터를 다루는 예외는 없다(Federal Register 2026-22925) verified.

OSHA는 작업장 전기 장비가 "승인"(실무상 NRTL 마크 또는 셀당 약 $3,000–$20,000+의 현장 평가)을 받도록 요구하며, 2025년 8월 11일부터 UL 1740은 로봇 셀 전체를 ISO 10218-2에 따르게 한다(eCFR 1910.399, Intertek, Industrial Monitor Direct). 방산 전자는 사실상 중국 본토 통제 공급자에게 닫혀 있다. ITAR는 생산 사진을 기술 데이터로 취급하고 중국에 거부 정책을 적용한다(eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). 미국행 셀은 고정 암, 무제한 카메라, 로컬 또는 미국 클라우드 모델을 쓰고, 원산지를 실질적으로 바꾸는 경우 동남아시아에서 최종 조립해야 한다.

어떤 예산에서도 천장은 지역별로 다르지만 어디서도 인력 0명에 도달하지 않는다 estimate. 중국 본토는 국산 3D SPI, 3D AOI, X-ray, AI 재판정, 사내 AI 팀을 겹쳐 재판정 인력을 8–15%까지 낮출 수 있지만, 폐쇄형 프런티어 모델과 DGX Station이 없으므로 공장 전역 LLM Q&A는 로컬 오픈 웨이트와 국산 컴퓨트로 돌려야 한다(다중 사용자 풀사이즈 모델은 8카드 Ascend 서버 최소 2대 필요). 대만과 동남아시아는 프런티어 API와 DGX Station을 포함해 기술적으로 최고를 모두 쓸 수 있지만, 고객 정체성, 중국 본토 브랜드, 중국 모기업 라이선스에 제한된다. 미국은 기술적 제한이 없고, 관세·NRTL 인증·통제 데이터 인력 분리가 대가다. 어디서든 경계 판정, 신제품·전환, 신규 결함 유형, 고객 서명에는 사람이 남는다.

따라서 여러 지역에 판매되는 제품에는 두 개의 모델 스택이 필요하다. 하나는 등록된 국산 모델 또는 로컬 오픈 웨이트 기반의 중국 본토 스택, 다른 하나는 프런티어 API + 비중국산 로컬 모델 기반의 해외 스택이다. 2026년 11월 12일 시행되는 Anthropic의 정책은 소유와 통제로 판단하므로, 중국 본토 통제 공급자는 해외 자회사, 클라우드 플랫폼, 리셀러를 통해서도 Claude 위에 구축할 수 없다(Anthropic) single source, primary.

02 온프레미스
공장 내부에서 AI 돌리기: 클린룸, 에어갭, 하드웨어, 양자화

에어갭 공장에서 제약은 컴퓨트가 아니라 프로세스다

완전 오프라인 비전 검사는 일상적이다. 제약은 절차에 있다. 중국의 2024년 MIIT 산업제어시스템 보안 지침은 구역 분리, 방화벽 또는 데이터 다이오드, 감사되는 시한부 원격 유지보수, 스캔된 매체, 테스트된 패치를 요구한다(guide text) single source. 반도체 고객은 SEMI E187·E188을 추가하는데, TSMC가 2023년 구매 계약에 명시했다(SEMI, TXOne) verified. 결과는 IPC-CFX로 MES에 기록되고, 기판 신원은 IPC-HERMES-9852로 전달된다(Hermes standard). 흔한 함정은 네트워크를 전제한 라이선싱이다. Roboflow의 오프라인 모드는 가중치를 30일 리스로 캐시한다(Roboflow). 오프라인 제품에는 오프라인 활성화, 스캔된 매체의 서명된 업데이트, 소프트웨어 BOM, USB 제어, 로컬 감사 로그, CFX 어댑터가 필요하다.

FIG 05 ***에어갭 배포 라인의 소형 박스가 판정하고 서버실이 질문에 답한다. 장애 시 사람으로 폴백하며 데이터는 공장 밖으로 나가지 않는다.***



출처: 저자 종합(제2장).

폐쇄형 프런티어 모델은 실제로 공장 서버실에 들어가지 않는다. Gemini는 정부·대기업용 Google Distributed Cloud를 통해서만 에어갭으로 제공된다(Google Cloud). Microsoft의 Foundry Local은 "자격을 갖춘 고객"에게만 대형 모델을 열며 GPT 모델을 명시하지 않는다(Microsoft). Claude는 클라우드 정부 리전에서만 제공되며 온프레미스 배포가 없다(Claude for government). 셋 모두 중국 본토에서 제공되지 않는다. 공장 장비는 오픈 웨이트 중심으로 설계해야 하며, 프런티어 모델은 기껏해야 본사에서 오프라인으로, 고객이 익명화 이미지 반출을 허용할 때만 이미지에 라벨을 붙이거나 설명할 수 있다.

서버실 하드웨어는 사무용 등급으로 취급해야 한다. Mac Studio는 10–35 °C, DGX Spark는 5–30 °C 정격이며, 지속 추론 시 열 셧다운 보고가 있다(Apple, NVIDIA forum) verified. 라인에는 −25~60 °C 정격의 Jetson T4000 산업용 PC 같은 무팬 산업용 박스를 쓴다(AAEON) vendor claim. 대형 공장의 다운타임 1시간은 $36,000–$2.3 million의 손실이므로(Siemens True Cost of Downtime 2024) vendor survey, 라인의 모든 AI 단계는 반드시 **페일 오픈**이어야 한다. 어떤 장애에서도 기존 AOI+사람 프로세스로 폴백한다.

청정·ESD 구역: 일반 SMT 홀은 등급 클린룸이 아니라 ESD 보호 구역이다

컴퓨트 위치를 제한하는 것은 청정 등급이 아니라 정전기와 공기 흐름이다. 대부분의 SMT 조립 구역은 "공식적으로 입자 제어가 되지 않으며", 세계적 수준의 공장도 "ISO 8에 도달하거나 근접할 수 있다"(Circuitnet expert panel) single source. 의료 PCBA와 카메라 모듈은 보통 ISO 7, 베어 다이와 광학 정렬은 로컬 ISO 5–6에 있다(PCBSync) single source. 클린룸 시공사는 SMT에 일반적으로 ISO 7–8을 제시한다(Angstrom Technology) vendor claim.

보편적 요구는 ESD 제어다. ANSI/ESD S20.20-2021은 절연 도체를 35 V 미만으로 유지하고 강전계 절연체를 민감 소자에서 최소 30 cm 떨어뜨릴 것을 요구한다(ANSI, ESDA, Desco) verified. 팹은 컴퓨트를 서브팹 또는 데이터센터에, SEMI S2/S8/F47을 충족하는 액랭 랙에 둔다(Prodrive) vendor claim.

장비가 청정 구역에 들어갈 수 있는지는 ISO 14644-14의 입자 방출 시험으로 판정된다. 로봇은 이런 인증을 보유하지만(Fraunhofer IPA), 산업용 PC·엣지 AI 박스·GPU 서버 중 이를 가진 제품은 찾지 못했다. CE, FCC, IEC 60068은 박스가 환경을 견딘다는 것을 보여줄 뿐 ESD 보호 구역 진입 허용을 뜻하지 않으며, 고객의 ESD 코디네이터가 확인한다. GPU 서버는 라인에 두지 않는다. DGX Station은 테스트에서 GPU에서 1,292 W를 소비했다(StorageReview).

FIG 06 **청정·ESD 구역이 컴퓨트 위치를 결정한다** ***일반 SMT 홀은 클린룸이 아닌 ESD 보호 구역이다. GPU·NPU 서버는 서버실에 둔다.***



출처: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, SEMI 표준·벤더 자료; 저자 편집.

구역	일반적 환경	라인 컴퓨트	GPU/NPU 서버
일반 SMT/PCBA 홀(소비재, 산업용)	S20.20 또는 IEC 61340-5-1의 ESD 보호 구역, 보통 등급 클린룸 아님(≈ISO 8 상당); RH 40–60%	AOI 또는 재판정 자리 옆 접지 금속 케이스의 무팬 박스(Jetson T4000/Thor, 중국 NPU 박스); 플라스틱 모니터·하우징은 구역 밖 또는 기판에서 ≥30 cm	서버실 또는 냉방 IT 랙
의료, 카메라 모듈, EV 전자(ISO 7–8 클린룸)	클린룸 + ESD 보호	무팬, IP65/66 밀봉, 매끄럽고 세척 용이; ISO 14644-14 데이터 또는 IPA 인증 요청	클린룸 밖, 네트워크로 연결
베어 다이, 와이어 본드, 광학 스테이션(로컬 ISO 5–6)	미니환경	내부에는 카메라·광원만; 컴퓨트 박스는 외부	서버실
팹 클린룸(ISO 5 이상)	팹 전용 ESD·정전 흡착 플랜, 장비는 SEMI E78; 사이버보안은 SEMI E187/E188	툴 통합·팹 인증 컨트롤러만; 팬 있는 애드온 PC 금지	서브팹 액랭 랙 또는 팹 데이터센터

이 표는 위 출처의 저자 종합이다 estimate. 제품으로 말하면, 약 100 W 미만의 무팬 라인 사이드 박스, 보호 접지에 본딩된 금속 케이스, 12–24 V DC 입력, 실제 캐비닛 공기 흐름에서의 광온도 동작, ISO 7 이상 고객을 위한 ISO 14644-14 입자 방출 보고서가 필요하다. "GPU 서버 1대가 4개 라인을 서비스"하는 설계는 서버를 서버실에 두고 라인에는 작은 박스나 카메라만 남긴다.

용도별 속도 한계: 정확도가 속도보다 어렵다

공장과 벤더는 AI 지연시간이나 가용성 목표를 거의 공개하지 않는다. 아래 임계값은 택트, 인력 재판정 시간, 로봇 제어 속도, 인터랙션 연구에서 도출했다 estimate .

용도	속도 한계	정확도 한계	가용성
모든 부품의 인라인 검사	½–1 택트 내: 10–30초 택트에 크롭 240–1,000개, 초당 8–100크롭	유출이 현재 "AOI+사람"보다 나쁘지 않게(계약 예 <0.5%)	페일 오픈, 라인 정지 금지; 목표 ≥99.9%(공개 SLA 없음)
AOI 플래그 크롭만의 인라인 세컨드 오피니언	플래그(보통 기판당 ≤5–20개)를 ½ 택트 내: 크롭당 0.5–3초	동일; 라인 자체 이미지로 학습	인력 재판정으로 폴백
오탐의 AI 재판정	호출당 2–5초, 사람의 42–55초보다 10배 이상 빠름	유출이 인력 재판정자보다 나쁘지 않게	성능 저하 시 인력 추가, 라인 정지 금지
로봇 제어(VLA)	실시간 청킹으로 액션 청크당 100–300 ms, 출력 10–50 Hz	프로그래밍된 모션을 이기려면 과제 성공률 ≥99%	모델과 독립된 안전 시스템
로봇 계획·스킬 선택	매 주기 필요 시 1–3초; 예외·전환 시에만이면 ≤1분	출력은 검증 가능한 코드 또는 스킬 ID여야 함	고정 프로그램으로 폴백
Q&A(SOP, MES 질의)	첫 토큰 이상적으로 ≤1–2초, 최대 10초; 사용자당 ≥15–20 tokens/s; 동시 사용자 5–20명	정확한 인용; 긴 컨텍스트에서 충실	사무 IT급(≈99–99.5%)
분석·근본 원인 보고	분 단위면 충분(배치)	최고 수준; 긴 컨텍스트에서 충실	사무 IT급

Physical Intelligence는 3B급 VLA의 추론 97 ms(네트워크·전처리 포함 약 108–139 ms)를 측정했고, 실시간 청킹이 최대 200 ms의 추가 지연까지 처리량을 유지함을 발견했으며, 성냥 켜기나 네트워크 케이블 연결 같은 정밀 과제를 300 ms 초과에서도 완수했다(Physical Intelligence) vendor research . 인터랙션 연구는 사고의 흐름을 유지하는 한계를 1초, 주의는 10초로 본다(NN/g). 정확도가 어려운 부분이다. 오픈 Qwen3-VL-8B는 5개 산업 이상 데이터셋에서 제로샷 평균 56.8에 그친다(IndusAgent) single source . 어떤 하드웨어도 라인 자체 이미지로 튜닝되지 않은 비전-언어 모델을 구제하지 못한다.

FIG 07 ***용도별 속도 한계인라인 검사는 밀리초, QA는 초, 분석은 분 단위. 사람은 건당 42–55 s.***



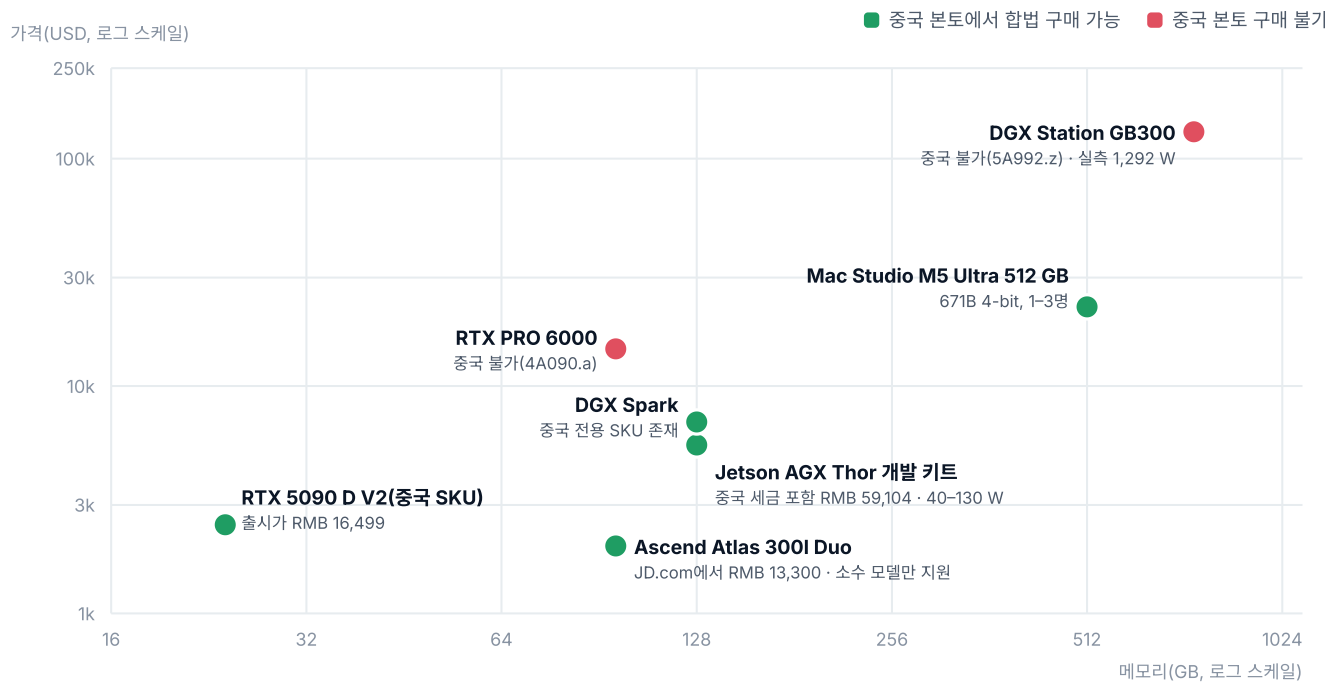
출처: 저자 종합(제1–2장).

하드웨어: 스테이션에는 Thor, 공장 Q&A에는 Mac Studio나 국산 어플라이언스, 중국 본토에는 DGX Station 금지

장치	메모리·대역폭	가격(2026년 11월)	적합한 용도	중국 본토
Jetson AGX Thor 개발 키트 / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	출시 $3,499 → $5,499(2026년 7월); T5000 $4,999, T4000 $2,999; RS China 개발 키트 세금 포함 RMB 59,104	크롭을 재확인하는 2–8B 비전 모델; 실시간 3B급 VLA	● 가능(5A992.c, 민간 공장 라이선스 불필요)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; 동일 칩 ASUS GX10 China 모델 JD에서 RMB 32,999(2025년 10월)	다중 사용자 8–35B MoE	● 가능(5A992.c, China 모델, 공인 채널)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250(2026년 6월) → 11월 eBay 평균 ≈$16,005	다중 사용자 27–35B FP8; 4-bit 120B급 MoE	● 불가(4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB 통합, 1.2 TB/s; 최대 지속 480 W	미국에서 $22,299부터; China 추정 ≈RMB 173–186k	1–3명용 DeepSeek 671B 4-bit	● 가능
DGX Station GB300	748 GB(252 GB HBM3e 7.1 TB/s + 496 GB LPDDR5X 396 GB/s); GPU 실측 1,292 W	11월 시중가 $123–140k(관세 전 ≈RMB 830–940k); 8월 $95–117k	HBM에 맞는 120–235B MoE의 고동시성; 433–743B 모델은 소수 사용자만	● 불가(5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	JD 판매자 RMB 13,300	저용량 배치 재판정; 지원 모델 소수	● 가능(국산)
Huawei Atlas 800I A2(8 × Ascend 910B 64 GB)	8 × 64 GB HBM	미확보; 풀 DeepSeek은 W8A8에서 2대, BF16에서 4대 필요	공장 전역 다중 사용자 Q&A·분석	● 가능(국산)
RTX 5090 D V2(China 모델)	24 GB	출시(2025년 8월) RMB 16,499, 이후 메모리 가격과 함께 상승	8B 비전 모델, 소형 VLA	● 가능

출처: (NVIDIA China Thor, NVIDIA Thor launch, CNX on price rises, NVIDIA export classification data, RS China, Implicator, IThome on ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, Apple specifications, TechRadar, IThome on M5 Ultra China pricing, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, vLLM-Ascend support matrix, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). 512 GB Mac Studio의 중국 본토 가격은 미국 구성 프리미엄에서 외삽했으며 신뢰도가 낮다 estimate .

FIG 08 **온프레미스 컴퓨트: 메모리, 가격, 중국 본토에서 합법 구매 가능한 것*벤치마크보다 먼저 수출 규제 분류가 후보를 결정한다.***



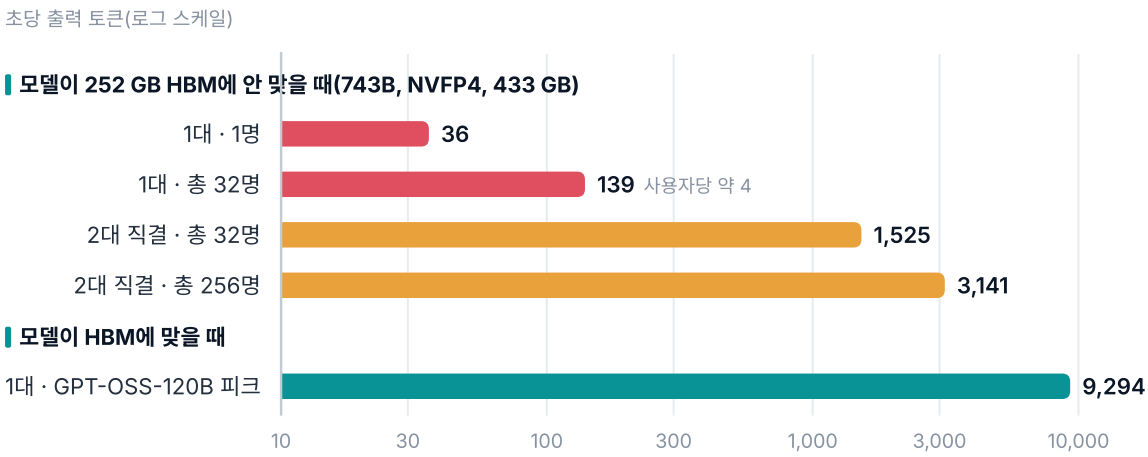
출처: NVIDIA 수출 분류, Apple, RS China, JD.com·리뷰 가격(2026년 11월).

Jetson Thor가 스테이션급 배포를 성립시킨다. NVIDIA의 TensorRT Edge-LLM 테스트는 Qwen3-VL-8B(NVFP4)가 Thor에서 프리필 2,048 토큰을 104.5 ms에, 디코딩을 44.4 tokens/s로 처리함을 보여준다(TensorRT Edge-LLM) vendor claim. AOI 크롭 하나와 20토큰 답변 기준으로 8B 모델에서 약 0.5초, 2B에서 0.2초, 배치 시 초당 8.5–21크롭이다 estimate. 지연시간을 결정하는 것은 파라미터 수가 아니라 출력 길이이므로, 클래스 토큰이나 짧은 JSON으로의 제약 디코딩이 설계의 일부다. GR00T N1.7(3B)은 Thor에서 25.1 Hz로 돈다(Jetson AI Lab) vendor claim. Orin NX는 크롭당 약 1.8초, Atlas 300I Duo는 p95가 23초를 넘어 배치 작업에만 적합하다(CSDN) single source. 중국 본토에서는 NVIDIA 제품 예산을 미국 리스트 가격의 40–60% 위로 잡아라 estimate.

512 GB Mac Studio는 한 대에서 풀사이즈 모델을 소수 사용자에게 가능하게 한다. M5 Ultra(2026년 8월)는 최대 512 GB를 1.2 TB/s로, 약 $22,299부터 제공한다(Apple, TechRadar). DeepSeek 671B 4-bit(≈380–404 GB)는 M3 Ultra에서 17–20 tokens/s로 디코딩하며(Pinggy) verified, M5 Ultra에서는 추정 25–29다. 긴 프롬프트가 약점이다. 10,000토큰 검색 프롬프트는 대략 17–70초 estimate로, 10초 인터랙션 한계를 넘는다.

DGX Station은 가장 강한 단일 장비지만 중국 본토에서 구할 수 없고 Mac의 5–6배 가격이다. 출하되는 GB300은 **748 GB**인데 그 중 252 GB만 고속 HBM이다(Wccftech, Tom's Hardware, Aug 2026) verified. 독립 테스트는 HBM에 맞는 GPT-OSS-120B에서 약 9,300 tokens/s를 보여주지만, LPDDR5X로 넘치는 743B 모델은 32명 사용자에게 총 139 tokens/s, 약 4 tokens/s씩만 주었고, 두 대를 연결하면 1,525에 도달했다(StorageReview, Aug 2026, StorageReview, Oct 2026) single source. 시중가는 11월까지 $123–140k로 올랐다(CyberPowerPC, monitors.com). 중국 본토에서 같은 작업의 현실적 선택은 Mac Studio, KTransformers 방식의 CPU+GPU 서버(KTransformers) vendor test, 또는 국산 Ascend 서버다.

FIG 09 **대형 워크스테이션 1대는 공유 어시스턴트가 되지 못한다** ***모델이 고대역폭 메모리에 안 맞으면 동시 사용자 32명이 각 약 4 tokens/s만 얻는다.***



출처: StorageReview(2026년 8·10월); ServeTheHome. 단일 소스 독립 테스트.

메모리 부족으로 같은 부품이 1년에 20–100% 움직였다(DGX Spark가 $3,999 → $6,950(TechRadar)). 따라서 장비 견적은 고정이 아니라 지수 연동이어야 한다.

국산 가속기는 풀사이즈 모델을 돌린다. 병목은 소프트웨어 유지보수다. Ascend 910B는 HBM 64 GB를 갖지만, vLLM-Ascend는 DeepSeek V3/R1을 "더 이상 유지보수하지 않음"으로 표시하고 작은 카드에서는 소수 모델만 지원한다(vLLM-Ascend) single source, primary. 새 모델마다 적응에 1–2개월이 걸린다(Caijing/Anue). 따라서 동결된 검증 모델 버전을 출하하라. 풀 DeepSeek 671B를 20–100명에게 서비스하려면 8카드 Atlas 800I A2 서버가 최소 2대 필요하다(China Telecom Cloud) verified. 그런데 중국의 LLM 어플라이언스 공공 입찰은 2025년 1–9월 총 RMB 27.5 million에 그쳤다(C114) single source. 공장은 컴퓨트 박스가 아니라 통합과 서비스에 돈을 낸다.

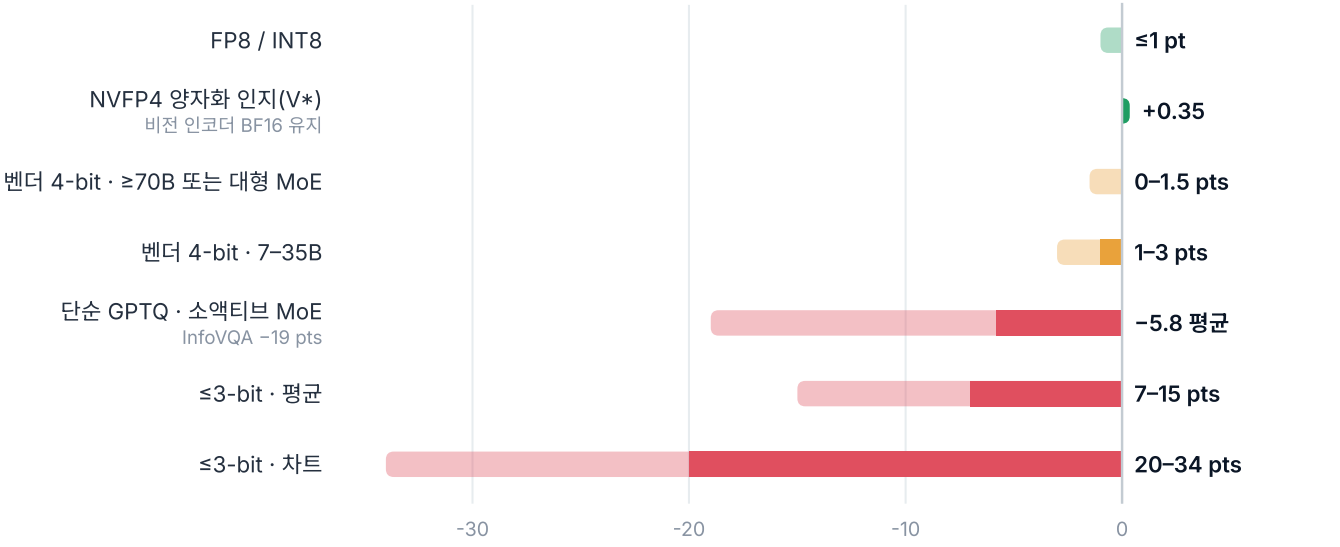
양자화: 8-bit는 안전, 4-bit는 레시피에 달림, 3-bit 이하는 탈락

8-bit(FP8 또는 INT8)는 비전-언어 모델에서 사실상 무손실이다(Red Hat, Hugging Face) verified. 4-bit에서는 손실이 크기와 레시피에 달린다. Qwen의 72B AWQ 빌드는 MMMU에서 1점 미만 손실인 반면 7B는 2.8점 손실이다(Qwen AWQ) verified. 세부가 많은 과제는 더 떨어진다(단순 4-bit GPTQ는 Qwen3-VL-30B-A3B를 InfoVQA에서 19점 깎는다(VEQ)). 3-bit 이하에서는 평균 손실이 7–15점, 차트는 20–34점 손실이다(VLMQ) verified. 64k 토큰을 넘으면 4-bit는 최대 59% 손실하며 비영어 입력에서 더 나쁘다(arXiv 2505.20276). 비전 인코더는 FP16/BF16으로 유지하라.

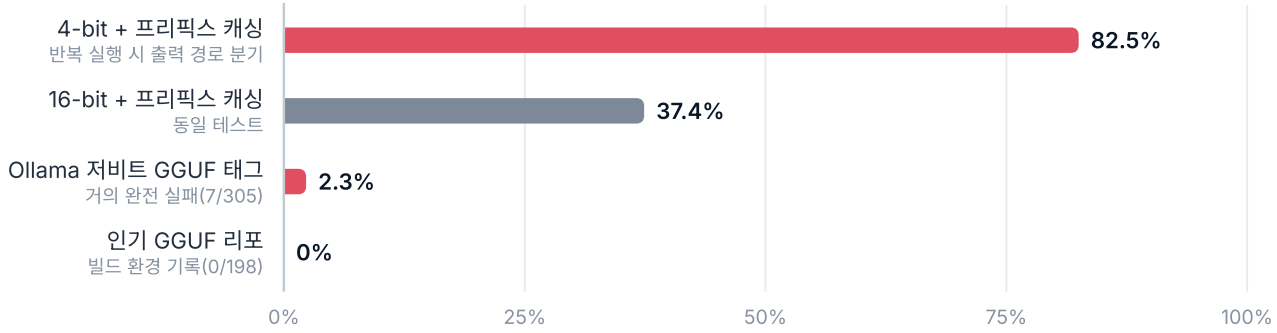
FIG 10 **양자화: 8-bit는 안전, 4-bit는 레시피 의존, 3-bit 이하는 탈락** ***서빙 스택의 엔지니어링 실수가 양자화 자체보다 정확도를 더 깎는다.***

정확도: 8-bit 안정, 4-bit 레시피 의존, ≤3-bit 탈락

전체 정밀도 대비 점수 변화(포인트)



양자화 오차보다 큰 것은 엔지니어링 사고



출처: NVIDIA Nemotron 모델 카드, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 등.

더 중요한 경고 두 가지가 있다. 첫째, **2026년 11월 11일 현재 어떤 연구나 모델 카드도 PCB/SMT 결함 이미지에서 양자화를 측정하지 않았으므로**, 골든 세트에서 재검증해야 한다. 둘째, 서빙 스택 실수가 양자화 오차보다 큰 경우가 많다. 같은 모델이 Ollama에서 참조 구현보다 약 30점 낮게 나왔고(SiliconBench), Ollama의 공식 코드 모델 태그 305개 중 7개는 파일이 온전함에도 사실상 고장이었으며(arXiv 2609.05881), 인기 GGUF 리포 198개 중 빌드를 기록한 것은 하나도 없었고(arXiv 2610.19498), 프리픽스 캐시 재사용 시 에피소드의 82.5%가 4-bit에서 분기했다(16-bit는 37.4%)(arXiv 2609.04748) single sources. 벤더의 양자화 인지 4-bit는 세부 과제에서 거의 무손실에 가깝지만(NVIDIA Nemotron 3.5) vendor claim, 4-bit VLA는 섭동에서 1.5–4.3점 손실하고 Thor에서 더 느리다(LoopQ) single source. 양자화된 모델은 새 계측기처럼 관리하라: 가중치 해시 고정, 빌드 기록, 프리픽스 캐시 리셋, 정밀도·엔진·템플릿 변경 시마다 재검증.

정밀도	≥70B dense 또는 ≥100B MoE	7–35B	≤4B 또는 세부 과제	판정
FP8 / INT8	안전(≤1점)	안전	안전; 검사 기본값	사용
4-bit AWQ/GPTQ(벤더 레시피, 비전 인코더 FP16)	0–1.5점	1–3점; 소액티브 MoE의 단순 GPTQ 평균 −5.8, InfoVQA −19	2–3점; 세부 과제 우선 검증	조건부
NVFP4(벤더 레시피)	FP8 대비 ≈0.3–1점 아래(벤더 데이터)	양자화 인지 빌드가 V*, OCRBench에서 거의 무손실	공개 데이터 없음	조건부(양자화 인지 빌드 우선)

정밀도	≥70B dense 또는 ≥100B MoE	7–35B	≤4B 또는 세부 과제	판정
GGUF Q4_K_M	미검증	도메인 과제에서 분산 큼	소형 텍스트·결함에 부적합	세부 판정에 부적합
GGUF Q5_K_M 이상	안전으로 추정	두 테스트에서 거의 무손실	llama.cpp에서 세부의 하한	사용
≤3-bit(한 장비에 1.8–2 bits로 압축된 1T 모델 포함)	결정에 부적합	불가	불가	불가

어떤 작업에 어떤 구성

구성	인라인 검사	오탐 AI 재판정	로봇 계획·제어	Q&A·분석
전용 검출기 또는 이상 모델 (YOLO, RF-DETR, PatchCore 계열), INT8/FP16, 모든 NPU·GPU 엣지 박스	가능: 주 판정기, 크롭당 2–12 ms	가능(라인에 튜닝된 분류기)	인지만	—
2–4B 비전 모델, NVFP4/AWQ/FP8 + FP16 비전 인코더, Thor/T4000 또는 DGX Spark	조건부: AOI 플래그 크롭을 크롭당 0.2–0.3초에 재확인	조건부, 튜닝 후 1초 미만	VLA 백본(청크당 ≈40–130 ms의 2–3B VLA)	약함
7–9B 비전 모델(Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 또는 벤더 4-bit, Thor·Spark 또는 24–32 GB 카드	조건부: 플래그 크롭당 ≈0.5초	가능: 호출당 0.5–1초, 최고 가성비	계획·스킬 선택(1–3초), 제어 루프는 아님	1명~소수 사용자 기본 SOP Q&A
26–35B MoE(Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	불가(너무 무거움)	가능, 설명자 또는 어려운 케이스용	계획	가능(Spark에서 8명, 첫 토큰 ≈0.1초)
Mac Studio, RTX PRO 6000 또는 Spark의 4-bit 106–241B MoE	불가	불가	오프라인 계획	가능, 1명~소수 사용자
512 GB Mac, KTransformers 서버 또는 DGX Station의 4-bit DeepSeek 671B 또는 Kimi K2.x 1T	불가	불가	오프라인 프로그램 생성만	조건부: 1–3명·짧은 프롬프트; 긴 프롬프트는 16–170초
≤3-bit 무엇이든; 소형 모델 세부 작업의 GGUF Q4_K_M; 다중 사용자 Ollama	불가	불가	불가	불가

출처: 위 TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research, KTransformers 데이터 및 (DGX Spark tests, RTX PRO 6000 tests, Hardware Corner on Atlas, Ultralytics YOLO26, RF-DETR) **community and vendor tests, mostly single source**.

그래서 흔한 두 걱정에 답하면: **느릴 필요가 없다**(소형·중형 비전 모델을 Thor나 Spark에 두면 모든 인라인·재판정 예산을 충족한다. 671B–1T 모델은 라인에서 빼라), 그리고 **양자화가 틀리게 할 필요가 없다**(기본 FP8/INT8, 또는 골든 세트로 검증한 7B 이상의 벤더 레시피 4-bit). 더 큰 위험은 범용 모델의 제로샷 산업 정확도 약 57점과 서빙 스택 실수다. 검사 요청에는 고정 배치 크기와 전용 큐를 주고(온도 0도 배치에서는 비트 재현이 안 된다(vLLM documentation)), N+1 핫 스탠바이를 유지하며, 라인의 모든 단계를 페일 오픈으로 만들어라.

교사와 학생

03 교사와 학생: 프런티어 모델은 검사 데이터를 사전 라벨할 수 있지만, 로봇 암의 진짜 교사는 여전히 사람이다

2026년 11월 교사 후보: 범용 비전은 두 배가 됐고, 산업 결함 점수는 없다

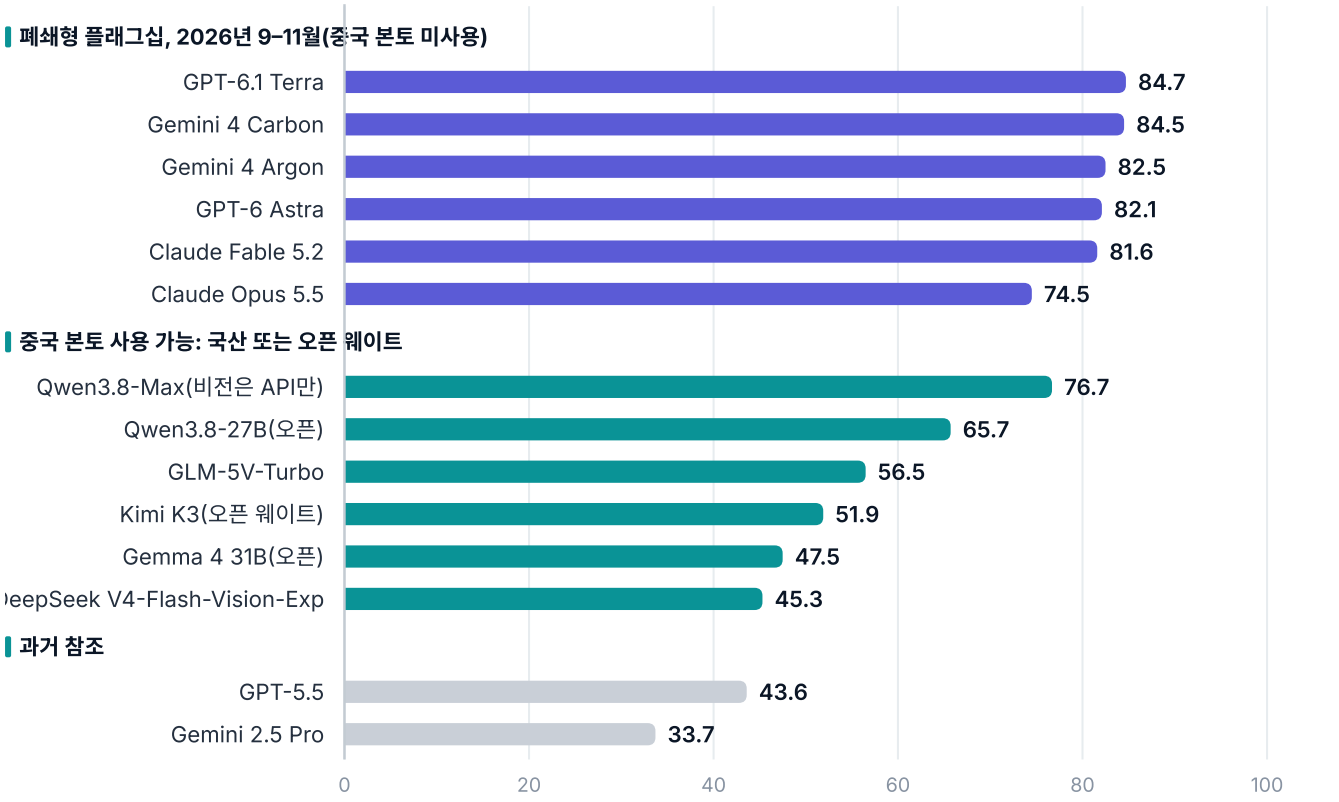
2026년 11월에 교사 역할을 할 수 있는 모든 폐쇄형 플래그십은 이미지를 받는다. Claude Fable 5.2(11월 9일 정식 출시, 백만 토큰당 $8/$40)와 Opus 5.5($4/$20)(Anthropic, Claude models overview) verified; GPT-6 Astra($10/$50)와 GPT-6.1 Terra($1/$5)(OpenAI pricing); 단계적으로 개방된 Gemini 4 Argon과 Carbon(Google Carbon). 중국 본토에서 쓸 수 있는 것은 Qwen3.8-Max(비전은 API 경유)와 Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp, Gemma 4 같은 오픈 모델이다(Roboflow blog, Hugging Face).

핵심 격차를 분명히 말해야 한다. 이 모델들 중 어떤 것도 MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB 또는 어떤 PCB/SMT 데이터셋에서도 공개 점수가 없다. MMAD에는 여전히 리더보드가 없고(MMAD GitHub), 가장 최근 산업 이상 논문 두 편(2026년 8–9월)도 최신 범용 기준선으로 Gemini 2.5 Pro와 Claude Sonnet 4를 쓴다(InspectorGPT, Anomaly-LR). 11월 11일 arXiv 전문 검색에서 이 맥락의 "Fable 5.2"나 "Gemini 4 Carbon"은 0건이었다(arXiv API) single source, absence of evidence. 가장 가까운 대체재:

- MMAD 점수가 있는 최신 프런티어 모델은 GPT-5.5다. 500장 부분집합에서 "결함 유무" 69.95%, 7개 과제 평균 76.20%인 반면 사람 전문가는 95.24%, 86.65%다(AnomalyClaw) single source. GPT-4o는 2024년에 68.63%, 74.92%였다(MMAD, ICLR 2025). 2년에 약 1.3점의 진전이다.
- IC 리소그래피의 실제 SEM 재판정 이미지이자 2026년 모델이 있는 유일한 전자 벤치마크인 LDU-Bench에서, 양불 트리아지는 괜찮다(매크로 F1: GPT-5.4 0.932, GLM-5V-Turbo 0.884, Claude Opus 4.6 0.876, Gemma 4 31B 0.504). 그러나 11개 결함 형태 분류는 0.20–0.41에 그친다(LDU-Bench) single source.
- Roboflow의 독립 비전 평가(2026년 11월 10일 갱신)에서 객체 검출 mAP@50은 Gemini 4 Carbon 84.5, GPT-6.1 Terra 84.7, Fable 5.2 81.6인 반면 Gemini 2.5 Pro 33.7, GPT-5.5 43.6, Claude Opus 4.8 38.6이다(Roboflow Vision Evals) single source, independent evaluator, private datasets.

FIG 11 **교사 후보: 범용 점수는 한 세대에 배가*그러나 2026년 플래그십 중 산업 결함 벤치마크 공개 점수는 하나도 없다.***

Roboflow 범용 검출 mAP@50



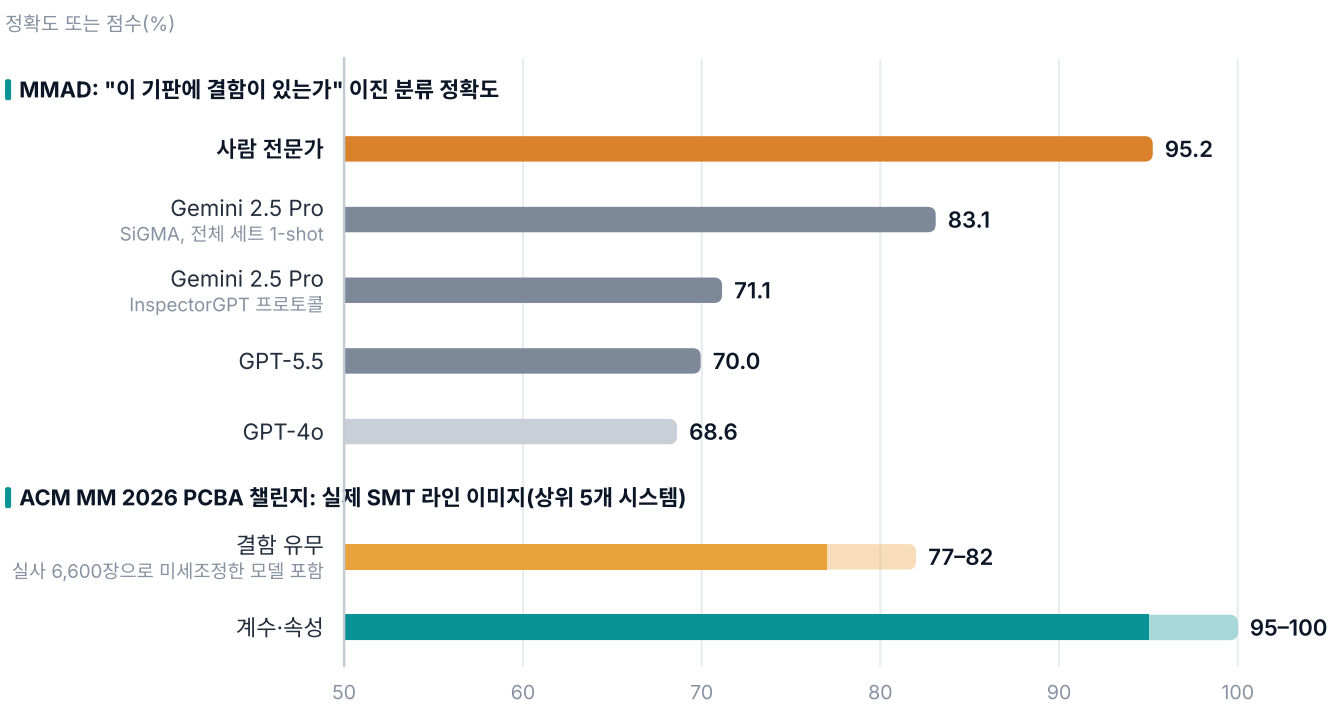
산업 결함 벤치마크(MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): 2026년 플래그십 공개 점수 0건
범용 검출 점수는 한 세대에 배가됐지만 결함 판정을 대체 못 함; 자사 골든 세트로 실측 필요

출처: Roboflow Vision Evals; 모델 릴리스 노트; MMAD, MVTec AD 2 등(2026년 11월 11일 기준).

교사 후보(2026년 11월)	백만 토큰당 가격(입/출)	Roboflow 검출 mAP@50 / 카운팅	산업 결함 벤치마크	중국 본토 사용 가능
Claude Fable 5.2	$8 / $40	81.6 / 86.5	없음	불가
Claude Opus 5.5	$4 / $20	74.5 / 80.6	없음	불가
GPT-6 Astra	$10 / $50	82.1 / 80.2	없음	불가
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	없음	불가
Gemini 4 Carbon	$2 / $10 (프로모션)	84.5 / 90.1	없음	불가
Qwen3.8-Max (비전은 API만)	$2 / $6	76.7 / 81.1	없음	가능
Qwen3.8-27B (오픈)	로컬	65.7 / 64.9	없음	가능
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench 트리아지 0.884	가능
Gemma 4 31B (오픈)	로컬	47.5 / 51.4	LDU-Bench 트리아지 0.504	가능(로컬 웨이트)
참조: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD 이진 69.95% / 68.63%	불가
사람 전문가	—	—	MMAD 이진 95.24%	—

최신 논문은 여전히 이전 세대를 테스트하지만 같은 방향을 가리킨다. 전체 MMAD 세트에서 사람 전문가는 결함 유무 95.2%, Gemini 2.5 Pro는 83.1%인데(SiGMA), 같은 모델이 다른 논문에서는 71.14%에 그쳤다(InspectorGPT). 테스트 프로토콜이 모델 세대만큼 점수를 움직인다. CVPR 2026 VAND 4.0 챌린지에서 기성 비전-언어 모델은 AP 61.88로 정점을 찍은 반면 지도학습 방식은 90.84였고, 주최측은 이들이 "현재 파인튜닝된 분류기를 대체할 수 없다"고 결론지었다(VAND 4.0) single source. ACM MM 2026 PCBA 챌린지에서는 상위 5개가 결함 유무 76.92–82.47인 반면 부품 유형과 핀 카운팅은 95–99.5에 도달했다(arXiv 2609.21276) single source. 인식은 판정보다 훨씬 쉽다.

FIG 12 **"결함이 있는가?"는 범용 모델에 여전히 어렵다** ***최고 범용 모델은 MMAD에서 사람 전문가보다 12 pts 이상 뒤진다. 실 라인 이미지에서는 튜닝 시스템이 낫다.***



출처: MMAD; SiGMA; InspectorGPT; ACM MM 2026 PCBA 챌린지.

세 가지 결론이 나온다. 보고 세는 능력은 한 세대에 두 배가 됐지만, 이상 판정은 느리게 움직이고 프로토콜에 크게 의존하며 체계적 편향을 동반한다. 신모델의 이득은 공장 자체 이미지로만 측정할 수 있다. 학습 분포 안에서는 파인튜닝된 7B 전문가가 여전히 범용 모델을 이긴다(Anomaly-LR, UniPCB). 그리고 실제 라인 데이터는 점수를 무너뜨린다. 조립된 전자 패널에서 이미지 수준 AUROC는 MVTec AD의 97–98에서 31.5–78.6으로 떨어진다(arXiv 2606.01992) single source. 중국 본토 기업이 쓸 수 있는 가장 강한 교사 Qwen3.8-Max는 글로벌 프런티어보다 약 8점 아래다.

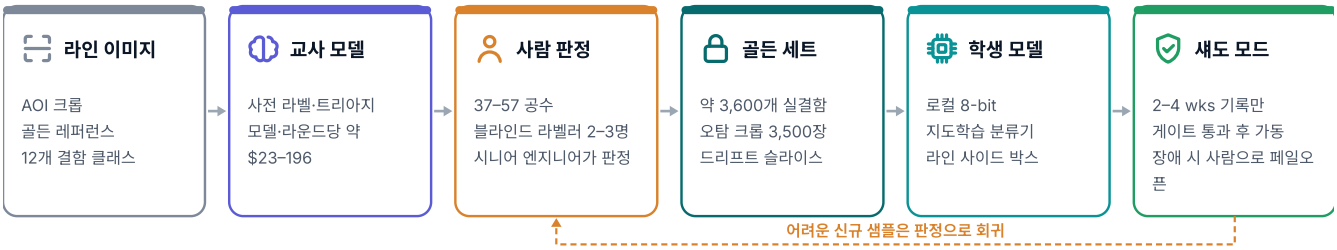
수십·수백 라운드 돌리기: 문제는 돈이 아니라 라벨 품질이다

돈은 문제가 아니다. 2026년 11월 리스트 가격 기준, 1메가픽셀 이미지 1,000장은 라운드당 약 $3(GPT-6.1 Terra)에서 $32(GPT-6 Astra)이고, 100,000장에 100라운드는 배치 할인 전 약 $32,000–135,000이다 estimate (Claude vision docs, OpenAI vision guide, Roboflow Vision Evals). 단순 라벨은 사람 비용이 Terra와 비슷하므로, 교사의 장점은 가격이 아니라 속도와 규모다. 고객의 NDA는 보통 제품 이미지를 공장 안에 묶어두므로, 골든 세트는 현장에서 라벨링된다.

진짜 장애물은 품질이며, 라운드를 늘려도 체계적 오류는 고쳐지지 않는다. Roboflow는 모든 과제를 세 번 돌리고 점수는 ±0.1–1.4점만 움직인다(Roboflow). 즉 모델은 대체로 같은 실수를 반복하며 다수결 투표로도 구제할 수 없다. 라운드는 각각 정보를 더할 때만 도움이 된다. 좋은 참조 이미지 1–2장은 약 3–4점을 더하고, 더 추가해도 소용없다(AD-Copilot). 좋은 참조에 대해 여러 턴에 걸쳐 모델이 반박하게 하면 GPT-5.5의 교차 도메인 AUROC가 6.23점, MMAD가 2.95점 오르며 약 3회 호출에서 정점을 찍는다(AnomalyClaw) single sources. "교사 라벨로 학습한 학생이 사람 라벨에 근접한다"는 것은 자연 이미지에서만 보였다. COCO와 Objects365의 Qwen3-VL 라벨로 학습한 ResNet-101은 사람 라벨 결과의 95.7–99.4%에 도달했고, 저자들은 도메인 데이터는 미검증이라고 밝힌다(arXiv 2602.20972) single source.

FIG 13 **두 개의 교사–학생 파이프라인** ***검사에서는 프런티어 모델이 사전 라벨하고 사람이 판정한다. 로봇 암의 진짜 교사는 여전히 사람 데모다.***

검사: 프런티어 모델이 사전 라벨, 사람이 판정, 골든 세트가 게이트



교사 라벨 자동 채택 기준

클래스당: 결함 ≥299개·미검출 0(95% 신뢰에서 재현율 ≥99%), 모델-정답 일치가 사람 간 이상. 아니면 사전 라벨만

암 제어: 교사는 여전히 사람 데모, 실기 RL, 시뮬레이션

프런티어 모델은 오프라인 프로그래밍·교체 시에만 도움



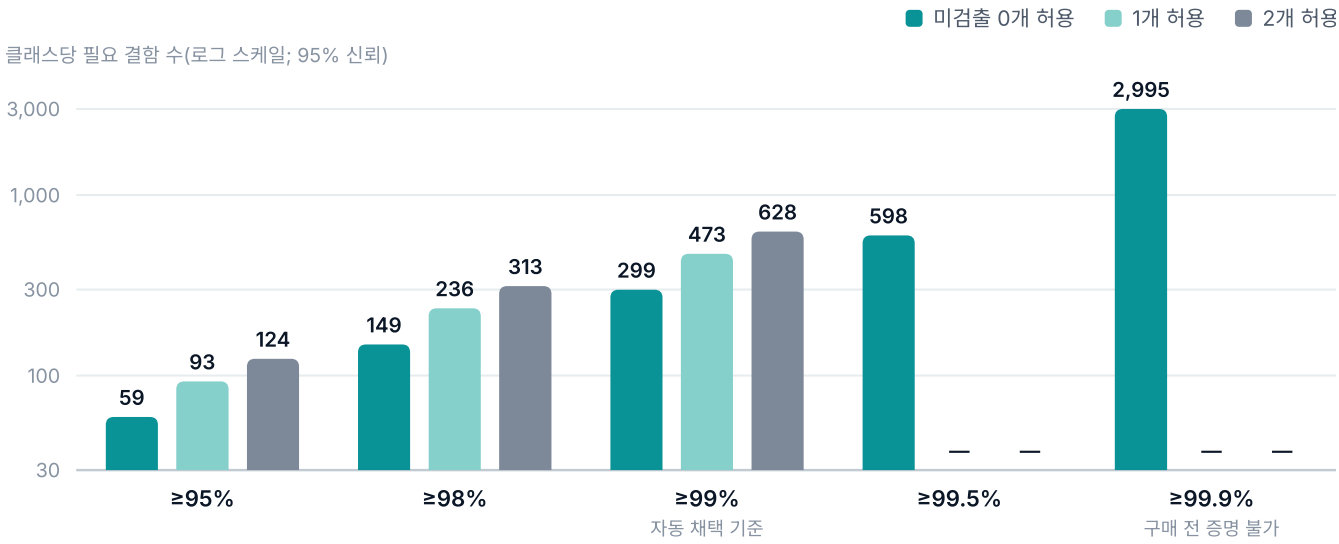
출처: 저자 종합(제3장); 비용·샘플 수는 추정.

실행 가능한 프로토콜 estimate :

1. **공장 자체 AOI 아카이브와 보관 중인 NG 기판에서 골든 세트를 수집하라.** 12개 실제 결함 클래스(누락, 오프셋, 묘비, 극성, 오부품, 브리지/쇼트, 솔더 부족, 솔더 과다, 솔더 볼, 리프트 리드/오픈, 부품 손상, 이물질)와 4개 오탐 원인(그림자, 플럭스, 실크스크린, 공차), 그리고 두 번째 라인 또는 다른 조명·기판 색·부품 번호에서 온 최소 300개 결함의 "드리프트 슬라이스"로 층화한다.
2. **IPC-A-610 교육을 받은 검사원 2–3명이 블라인드로 라벨링하라.** 시니어 엔지니어가 결정하고 X-ray, 전기 테스트, 단면으로 모호한 샘플을 확정한다. 검사원 간 일치도가 κ ≥0.75(치명 클래스는 ≥0.90)에 도달해야만 클래스 정의를 채택한다. LDU-Bench의 11개 전자 결함 형태에 대한 사람 간 일치도는 0.727에 불과했으므로 일부 클래스는 병합하거나 재정의해야 한다. AIAG 방식의 속성 일치(효과성 ≥90%, 미검출 ≤2%, 오탐 ≤5%, κ >0.75; 릴리스는 >0.90)가 합리적 최소선이다(Fabrico, iFactory).
3. **정확 이항(Clopper–Pearson) 단측 95% 경계로 크기를 정하라**(binomial confidence intervals, rule of three:

증명할 재현율	미검출 0 허용	1 허용	2 허용
≥95%	결함 59개	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 ***골든 세트는 얼마나 커야 하나 미검출 0으로 재현율 99%를 증명하려면 클래스당 결함 299개가 필요하다. 99.9%는 구매 전 증명 불가.***



출처: 정확 이항 신뢰구간(95%).

클래스당 결함 60개면 합격/불합격 스크린에 충분하고, 자동 채택을 결정하려면 300개가 필요하다. 전체 세트는 결함 약 3,600개 + 오탐 크롭 3,500개로, 라벨링·판정에 약 37–57 공수가 들고 교사 후보를 돌리는 비용은 $23–196이다. 진짜 비용은 가장 드문 클래스에 의해 정해지는 2–6주의 수집 기간이다 estimate. 기판 전체가 아니라 단일 부품 AOI 크롭을 보내되, 목표 크기는 최소 42–48 픽셀이다(arXiv 2610.09313).

1. **교사 라벨은 골든 결함 ≥299개에서 미검출 0인 클래스에만 자동 채택하라.** 모델-정답 일치도는 사람-사람 일치도보다 나쁘지 않아야 하고, 오탐의 최소 50%를 해소해야 한다. 쇼트, 부품 누락, 극성, 오부품 같은 치명 클래스는 99.5–99.9%(결함 598–2,995개)가 필요하다. 그렇지 않으면 AI가 높은 신뢰도로 해소하지 못한 모든 것은 사람에게 간다. 결함의 0.1% 미만 유출률은 구매 전에 증명할 수 없다(미검출 0으로 2,995개; 0.01%는 29,956개). 따라서 라인은 AI가 해소한 호출에 대한 인력 샘플링, 다운스트림 ICT/FCT 백스톱, 유출에 대한 SPC를 유지한다. 검증된 라벨은 그다음 기존 검출기나 분류기로 증류되고 신제품마다 재검증된다.

조건과 데이터 규칙도 교사를 결정한다. Anthropic은 자사 서비스로 경쟁 모델을 학습하는 것을 금지한다(Anthropic commercial terms). OpenAI는 출력으로 제3자에게 판매하지 않는 사내 분류기를 학습하는 것을 허용한다(OpenAI business terms). DeepSeek은 증류를 명시적으로 허용한다(DeepSeek terms) primary terms. 어느 쪽도 "경쟁 모델"을 정의하지 않으므로 상업적 사용에는 법률 검토가 필요하고, 고객의 데이터 규칙은 보통 이미지를 현장에 묶어둔다. 가장 안전한 기본값은 로컬로 돌리는 오픈 웨이트 교사이며, 클라우드 프런티어 모델은 고객이 서면으로 허용할 때만 쓴다.

검사 학생: 작고 빠르다. 병목은 광학, 데이터, 검증이다

학생이 병목은 아니다. YOLO26n은 $249–399 Orin Nano Super에서 INT8로 이미지당 3.80 ms이고(Ultralytics) vendor test, NVIDIA는 약 600장으로 파운데이션 모델을 적응시킨 뒤 세 PCB 결함 클래스에서 98.51%를 보고하며(NVIDIA) vendor claim, 합성 결함은 mAP를 0.9–2.5점만 더한다(arXiv 2605.04635). 제약은 광학, 라인 데이터, 검증, 라이선스다. Ultralytics YOLO는 AGPL-3.0 또는 엔터프라이즈 라이선스이므로(Ultralytics), 폐쇄형 제품은 라이선스가 필요하거나 RF-DETR 같은 Apache-2.0 검출기를 써야 한다.

로봇 학생: 구조화 스테이션은 생산급에 근접, 유연·동적 작업은 아니다

2026년 11월 로봇 교사는 세 종류가 있고, 로컬 학생이 될 수 있는 것은 하나뿐이다. 오픈·파인튜닝 가능한 VLA(π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA)다. 폐쇄형 로봇 모델(베타 API 경유 π0.6/π0.7, 신뢰받는 테스터용 Gemini Robotics 2)은 현장 배포가 불가능하다(The AI Insider, DeepMind). 프런티어 범용 모델은 느리고 정밀 조작에 약하다. Artificial Analysis 테스트에서 46–69점, 그랩 성공률 28–45%, 과제당 약 30분으로 "생산용으로는 너무 느리고 비싸다"고 했다(AA-Robotics). GPT-6 Astra는

정밀 슬롯을 20회 중 2회만 성공했고(RoboCurve), 관절 직접 구동은 필요한 약 83 Hz 대비 0.2–0.4 Hz에서 LIBERO 0–5.5%다(Anthropic) verified across three sources. Fable 5.2의 시뮬레이션 RoboDojo 벤치마크 64.9%는 내부 실행이다(Anthropic) vendor claim.

10월 이후의 제3자 연구도 일치한다. 세 선두 에이전트의 LIBERO-10 약 33–44%(arXiv 2610.30051), 하위 과제의 24–39%에서 성공을 허위 주장(arXiv 2611.08209), Opus 5.5가 준정적 실기 과제 108개 중 104개를 완수했지만 과제당 280초에 삽입·조립 없음(AMP1) single sources. 프런티어 모델은 오프라인에 속한다. 프로그램 작성, 과제 오케스트레이션, 성공 기준 초안 작성이다.

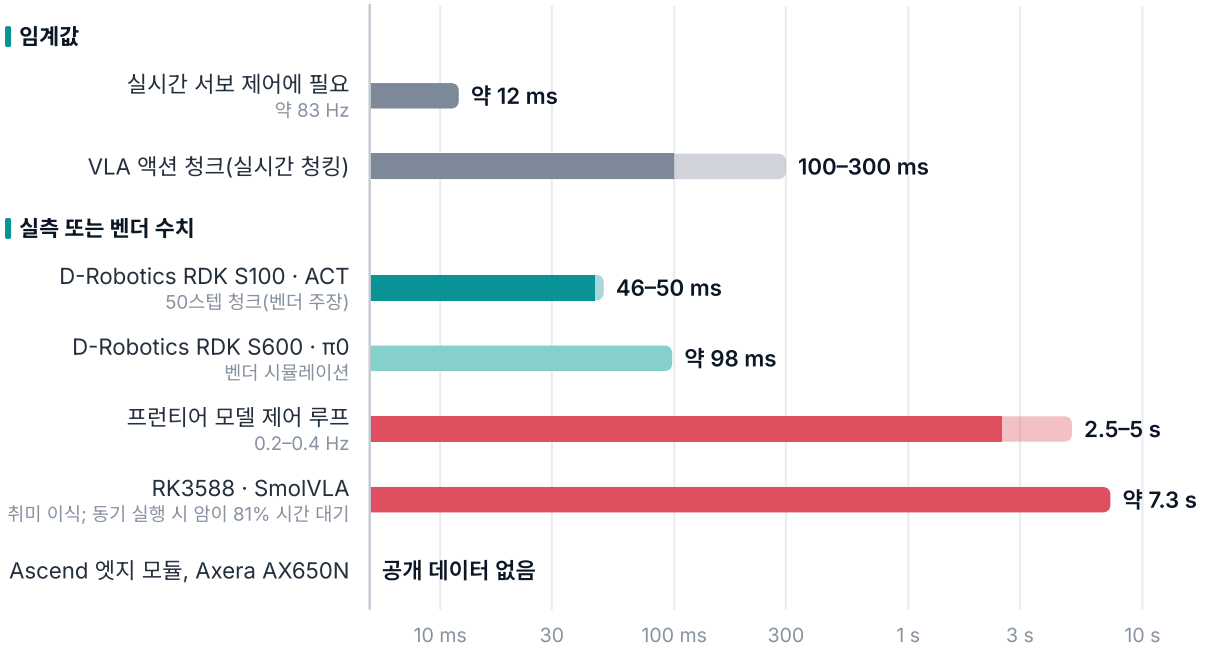
학생 품질에는 분명한 특징이 있다. 학습 분포 안에서는 좋고, 밖에서는 무너진다. π0.7은 본 과제에서 "종종 90% 초과", 미확인 과제에서 60–80%(π0.7 paper) vendor research. ContactDP는 학습된 커넥터에서 97.8%지만 미확인 커넥터에서 28.6–66.7%다(ContactDP). 약간의 카메라·포즈 변화만으로 LIBERO-Plus에서 95%가 30% 아래로 떨어진다(LIBERO-Plus, CVPR 2026). 이동 컨베이어에서 학습된 그랩은 58.89%로 정점을 찍는데(ForeTime-VLA) single sources, 이는 전통 비전 트래킹이 수년째 양산에서 해온 일이다.

생산급에 가까운 증거는 모두 구조화 스테이션, 사람 시연, 실기 보정에서 나온다. Neuromeka는 상태 머신과 사전 교시 모션을 유지하고 그랩, 케이블 삽입, 납땜만 각각 20분 미만의 데이터로 학습시켜 생산 324개 중 322개, 사람보다 12.8% 느렸다(Neuromeka) single source, vendor. Fraunhofer는 68분의 시연으로 4개 커넥터에서 92.8%에 도달했다(Fraunhofer). Berkeley의 HIL-SERL은 1–2.5시간의 실기 RL 후 RAM·SSD·USB 삽입에서 100회 중 100회 성공했다(HIL-SERL). AgiBot은 태블릿 테스트 라인에서 20초 택트 안에 단계당 18–20초로 움직였다(IThome) vendor claim.

증류는 속도를 사지 일반화를 사지 않는다. π0.5는 Jetson Orin 원시에서 0.70 Hz, 비동기 스케줄링 후 10.30 Hz로 돌고(EagleVLA), 레이어 증류는 약 3 Hz에서 약 10 Hz로 올린다(Shallow-π) verified. 학생은 교사가 커버한 것만 물려받으며, 신제품마다 여전히 과제 데이터가 필요하다.

FIG 15 ***엣지 정책 지연소형 ACT 정책은 엣지 보드에서 약 50 ms. 취미용 칩에 이식한 VLA는 7.3 s.***

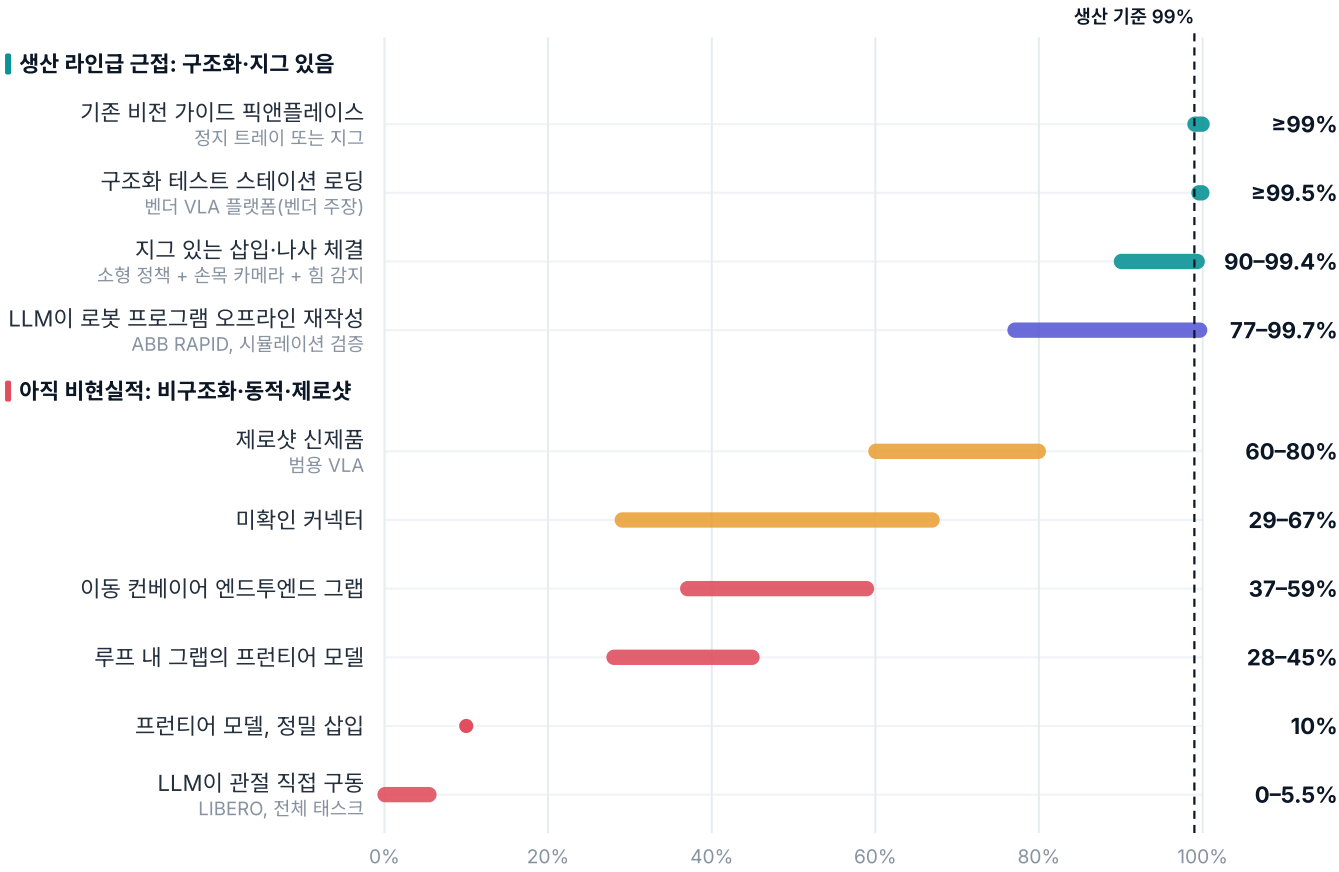
액션 청크·제어 주기당 추론 지연(로그 스케일)



출처: D-Robotics 문서·포럼(벤더 수치); GitHub 커뮤니티 이식; Anthropic 연구.

중국 엣지 칩에서 측정된 유일한 학습 정책은 D-Robotics RDK S100의 ACT 정책으로, 50스텝 청크당 46–50 ms이고 성공률은 공개되지 않았다(D-Robotics RDK S docs) vendor claim. RK3588에 취미로 이식한 SmolVLA는 청크당 약 7.3초가 걸린다(GitHub). 양자화가 속도를 공짜로 주지 않는다. 배치 크기 1에서 16-bit보다 빠른 4-bit π0.5 빌드는 없었다(LoopQ). 시연은 비싸다(Reuters는 초보 텔레오퍼레이터가 쓸 만한 모션 하나당 약 300회 시도가 필요하다고 보도했다(Notebookcheck)). 검증은 더 비싸다. 실패율 ≤0.1%를 95% 신뢰도로 증명하려면 연속 성공 약 3,000회, 17–22 로봇시가 필요하다 estimate. 학습된 조작 정책 중 안전 인증을 받은 것은 없다.

FIG 16 **로봇 학생: 스테이션이 구조화될 때만 생산급에 근접*지그 있는 스테이션은 90–99%+. 비구조화·동적·제로샷은 아직 비현실적.***



출처: 제3장 인용 논문·벤더 자료; 일부 수치는 벤더 주장.

전자 과제	실행 가능한 학생	기대 성공률	사람 대비 속도	전환당 비용	2026–2028 판정
하이믹스 픽앤플레이스, 키팅, 머신 텐딩(정지 트레이·지그)	학습된 인지 + 프로그래밍·계획된 모션	전통 비전 가이던스로 ≥99%	기존 스테이션과 동일	수 시간	지금 실현 가능
지그 있는 커넥터·FFC·케이블 삽입, 나사 체결, 납땜	상태 머신 + 소형 BC/확산/ACT 정책 + 손목 카메라 + 힘/토크(이상적으로 터치) + 사람 보정	90–99.4%	10–15% 느림	데모 10–300회(≈15–70분) + 검증 1,000–3,000회(로봇시 6–22시간)	실현 가능, R&D 집약; 첫 스테이션은 수개월
구조화 테스트 스테이션 텐딩과 OK/NG 분류	벤더 VLA 플랫폼 + 디지털 트윈 + 실기 RL	≥99.5% vendor claim, conflicting	단계당 18–20초, 20초 택트 내	재학습 ≤4시간 + 캘리브레이션 15분	벤더 플랫폼에서 실현 가능
전환 프로그래밍	LLM이 로봇 프로그램을 오프라인 작성·수정, 시뮬레이션 검증	ABB RAPID 재작성 77–99.7%	택트에 영향 없음	프로그램당 분~시간	엔지니어링 도구로 실현 가능
이동 컨베이어 엔드투엔드 학습 그랩	VLA 학생	37–59%; 고속 벨트 ≤37%	—	셋업당 데모 ≥458회	2028년 이전 불가; 전통 트래킹 + 학습 인지 사용
새 데이터 없는 신제품(제로샷)	범용 VLA	60–80%; 미확인 커넥터 29–67%	—	—	비현실적

전자 과제	실행 가능한 학생	기대 성공률	사람 대비 속도	전환당 비용	2026–2028 판정
암을 구동하는 루프 내 프런티어 LLM	GPT-6, Claude, Gemini 에이전트	그랩 28–45%; 정밀 삽입 10%	과제당 ≈30분 또는 280초 vs ≈20초	—	비현실적
자유 배선 하네스와 다단계 유연 조립	—	2026년 생산 증거 없음	—	—	2028년 이전 불가

출처: 위 연구 및 (University of Passau on RAPID rewriting, GraspVLA, Komax) **mostly single source and vendor data** . 2026–2028년에 대형 모델이 로봇에 주는 가치는 사전 학습된 사전 지식과 오프라인 프로그래머이지, 학생을 "가르치는" 무엇이 아니다. 내가 찾은 생산급 근접 사례(Neuromeka, Longcheer의 AgiBot, Fraunhofer, ContactDP)는 모두 사람 시연, 힘·터치 센싱, 시뮬레이션 또는 디지털 트윈, 지도 RL에 의존한다. 프런티어 모델의 라벨에 의존하는 사례는 없다.

비용과 회수

04 비용과 회수: AI 재판정은 수개월, 머신 텐딩은 수년, 선별 전용 암은 결코 회수 안 됨

인건비: 미국은 선전의 약 5.5배, 대만은 약 1.5배, 대부분의 동남아시아는 절반 미만

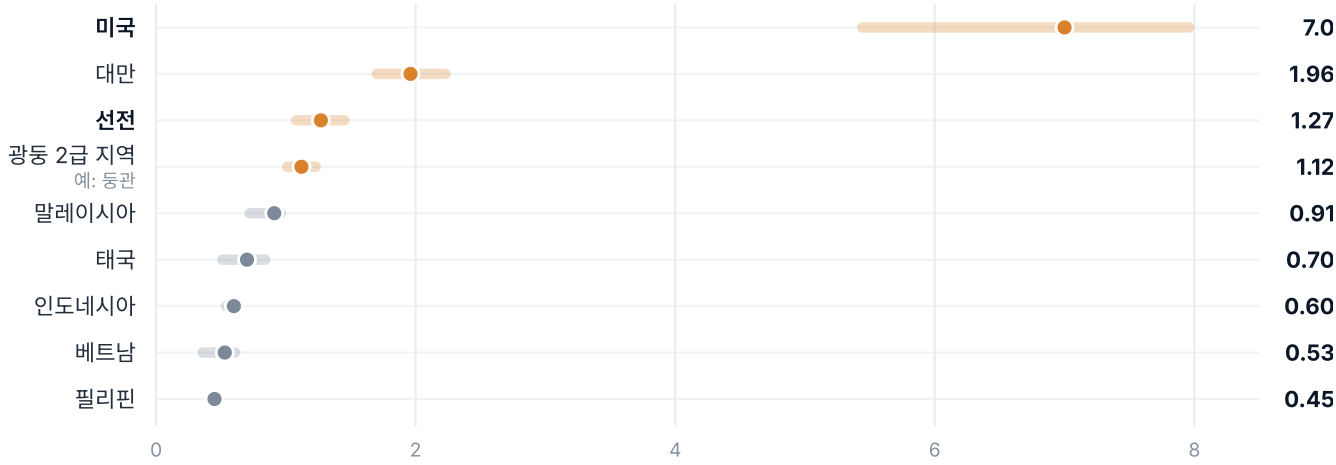
노동 대체의 수익은 노동 비용에서 시작한다. 선전의 공식 임금 지침은 2024년 제조 품질 검사원 중위값을 연 **RMB 74,074**로 제시한다. 임금은 전자 제조 총 노동 비용의 86.92%이므로, 전액 부담 비용은 약 RMB 85,000이다. JETRO의 2025년 조사는 중국 공장 노동자의 연 비용을 $12,667, 약 RMB 85,100으로 제시하며 0.2% 이내로 일치한다(선전 wage guidance, JETRO 2025) verified .

지역	검사원·핸들러 1인의 연 전액 부담 비용 (저 / 기준 / 고)	근거
중국 본토, 선전	RMB 70k / **85k** / 100k ($10.4k / **12.7k** / 14.9k)	공식 중위값 × 1.15; JETRO와 일치 verified
Guangdong 2급 임금 지역 (Dongguan 등)	RMB 65k / **75k** / 85k	Guangdong 2023 중위값 × 1.15 + 2년 인상분 estimate
Yangtze River Delta(Suzhou, Kunshan, Wuxi)	선전과 동일	더 저렴한 검사 노동의 근거 없음; Jiangsu 최저임금이 선전보다 약 5.6% 높음 estimate
대만	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	노동부 직종별 임금 조사 + ≈1.18 법정 부담금; 고점은 JETRO estimate
말레이시아	$6.8k / **9.1k** / 10.0k	JETRO와 최저임금 estimate
Thailand	$4.7k / **7.0k** / 8.8k	동일
Indonesia	$5.0k / **6.0k** / 6.3k	동일
베트남	$3.2k / **5.3k** / 6.5k	동일
필리핀	$4.0k / **4.5k** / 4.9k	동일
미국	$54k / **70k** / 80k	BLS 2025년 5월 검사원 중위값 $48,570 ÷ 임금이 보상의 65.5% estimate

출처: (JETRO 2025, Guangdong wage guidance, Wuxi HR bureau, 대만 occupational wage survey, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **검사·이송 작업자 1인의 연간 비용*총비용 기준 미국은 선전의 약 5.5배, 대만은 약 1.5배.***

QC·이송 작업자 1인 연간 총비용($10k; 점=기준, 막대=저–고)



출처: 선전 공식 임금 지침, 대만 노동부, JETRO, BLS; 저자 추정.

세 가지 보정이 근거를 만든다. 첫째, 선전은 2026년 7월 연금 납부 하한을 월 RMB 4,775에서 5,521로 올렸다(선전 Bendibao) single source. 따라서 2026년 선전 검사원은 RMB 80–90k가 더 현실적이다. 둘째, 자리는 사람이 아니다. 12시간 2교대에는 Pearl River Delta에서 약 2.0명(주 6일)~2.4명(연중), 동남아시아에서는 초과근무 상한 아래 약 2.7–3.1명, 미국과 대만의 24×7은 4.2–4.5명이 필요하다 estimate. AI 재판정은 보통 재판정자 1명이 2–4개 라인을 커버하게 하므로, 현실적 절감은 **교대당 라인당 1명이 아니라 라인당 0.5–1.5명(정규직)**이다. 셋째, 대만·베트남·말레이시아의 구매자는 비용만 쫓지 않는다. JETRO 조사에서 지난 2년간 채용이 더 어려워졌다고 답한 기업은 대만 61.1%, 베트남 48.2%, 말레이시아 44.9%인 반면 중국 본토는 18.6%다(JETRO 2025) single source.

장비 가격: 평균내면 안 되는 네 가지 기준

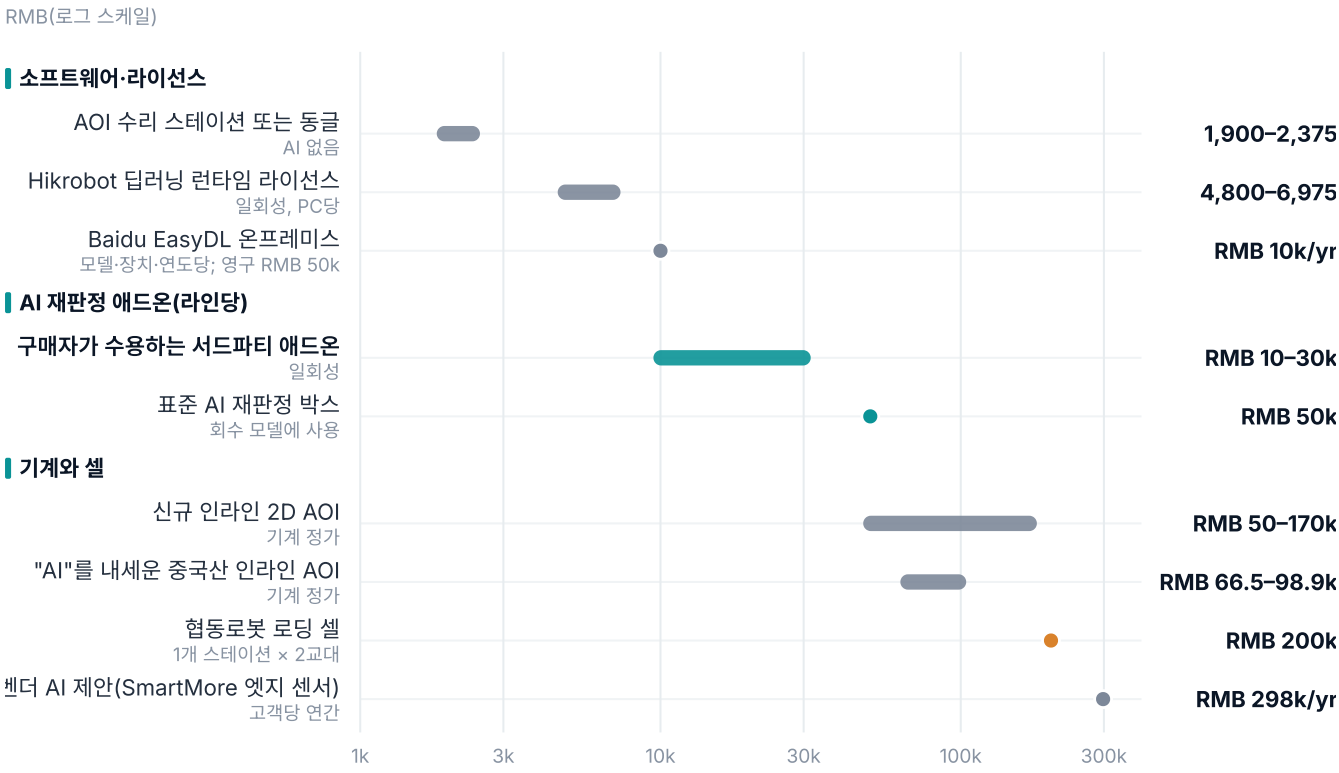
AOI 가격이 갈리는 주된 이유는 서로 다른 것을 측정하기 때문이다. 온라인 리스팅(2026년 11월, 신품·중고·렌탈 혼합)은 오프라인·데스크톱 2D 약 RMB 17–80k, 인라인 2D RMB 50–170k, 인라인 3D RMB 80–475k, 한국산 3D는 최대 RMB 700k다(1688 inline AOI, 1688 offline AOI, 1688 3D AOI) single source, list prices. 2025년 11월 입찰 하나는 3D AOI 1대 + 3D SPI 1대를 RMB 800k로 상한을 정했다(tender notice). Sinic-Tek의 증권신고서는 2019–2022년 3D AOI 대당 RMB 428–512k, 2025년 보고서 기준 약 RMB 340–470k를 시사한다(Sinic-Tek prospectus, Sinic-Tek 2025 annual report) primary + estimate. 해외 벤더는 데스크톱 2D $3,000–20,000, 인라인 2D $30–70k, 엔트리 3D $50–90k, 하이엔드 3D $100–200k+를 추정하며, 같은 인라인 장비가 중국에서는 $30–50k라고 한다(Averroes) vendor estimate from a seller of competing AI software.

AI 재판정 자체는 공개 RMB 가격이 없다. 벤더는 무료 현장 평가 후 건별로 견적을 낸다(Sohu on Hengtianyi). 이커머스 리스팅, 입찰, 벤더 기사, 대만 언론 검색에서도 공개 견적은 나오지 않았다(1688, 51SMT). 기준점은: Siemens 사례에서 시사되는 라인당 약 €27k estimate; 미국 유통사의 원격 재판정 네트워크 모듈 리스트 $21,924(Manncorp) vendor list price; SmartMore 엣지 AI 센서의 2025년 고객당 연 약 RMB 298k(36Kr) single source. 국산 인라인 2D AOI가 "AI"로 광고되며 RMB 66.5k에 리스팅되고(1688 inline AOI), 수동 재검사용 NG 버퍼는 약 RMB 50k다(1688 AOI review). 따라서 서드파티 재판정 애드온의 상한은 라인당 약 RMB 10–30k 또는 비슷한 연회비다. 그 이상이면 구매자는 AI가 내장된 국산 AOI로 장비를 교체하는 편을 택한다 estimate.

기준점	가격(RMB)	제공하는 것	출처
AOI 수리 스테이션 또는 소프트웨어 동글(구형 TRI, Zhenhuaxing)	≈1,900–2,375	라이선스 키, AI 없음	1688 list price
범용 딥러닝 비전 런타임(Hikrobot VisionMaster 동글)	PC당 일회성 ≈4,800–6,975	플랫폼; 고객이 모델 구축·유지	1688 list price

기준점	가격(RMB)	제공하는 것	출처
Baidu EasyDL 온프레미스 배포	모델·장치·연당 10,000, 또는 영구 50,000	패키징된 추론 서비스	Baidu AI docs **2022 price page**
"AI"로 광고되는 국산 인라인 AOI	대당 66,500–98,900	장비 교체	1688 **list price**
SmartMore 엣지 AI 센서	고객당 연 ≈298,000	여러 장비를 아우르는 벤더 솔루션	36Kr **single source**

FIG 18 ***가격 레더수천 RMB의 소프트웨어 라이선스에서 로봇 셀·벤더 AI 제안의 RMB 200–300k까지.***



출처: 1688 리스팅, Baidu AI 문서(2022 가격표), 36Kr, 업계 인터뷰(미검증).

Baidu 방식의 연회비(추론 서버 1대로 1–4개 라인, 모델 1–4개)는 라인당 연 RMB 2.5–40k, 현실적으로 약 RMB 5–10k의 구독을 시사한다. 영구 라이선스의 20%면 구독과 구매가 5년차에 손익분기인데, 따라 할 만한 구조다 estimate. 구매자는 거의 0에 가까운 소프트웨어 요금에 익숙하다. Sinic-Tek의 "기타" 매출(부품, 서비스, 렌탈, 액세서리) 전체가 2025년 매출의 7.65%, 설치 장비당 연 약 RMB 2,300–3,100이었다(Sinic-Tek 2025 annual report) primary + estimate. 라인당 연 RMB 5–10k의 AI 구독은 구매자가 현재 AOI 소프트웨어에 쓰는 금액의 2–5배이므로, 기존 소프트웨어 요금과 비교해서가 아니라 그것이 풀어주는 재판정 공수(정규직 1인당 연 약 RMB 75–90k)로 정당화해야 한다.

로봇 암의 경우, 중국 협동로봇 평균 가격은 2025년 약 RMB 58,600으로 내렸고(GGII) estimate, Dobot의 6축 평균은 저가 모델 점유율 상승으로 그해 19.1% 내렸다(Huxiu) verified. 같은 모델의 리스트 가격은 2026년 5–15% 올랐다(Sina Finance). 미국에서 FANUC CRX-10iA는 $51,434에 리스팅되고(Vention), 집계에서는 UR7e가 $30–48k다(cobotcost) estimate. 암은 미국 셀의 40–50%에 불과하며(Standard Bots), 중국에서는 카메라·그리퍼·안전·엔지니어링이 RMB 40k 암의 2–4배에 쉽게 이른다. 따라서 배수보다 절대 엔지니어링 비용으로 예산을 잡아라 estimate.

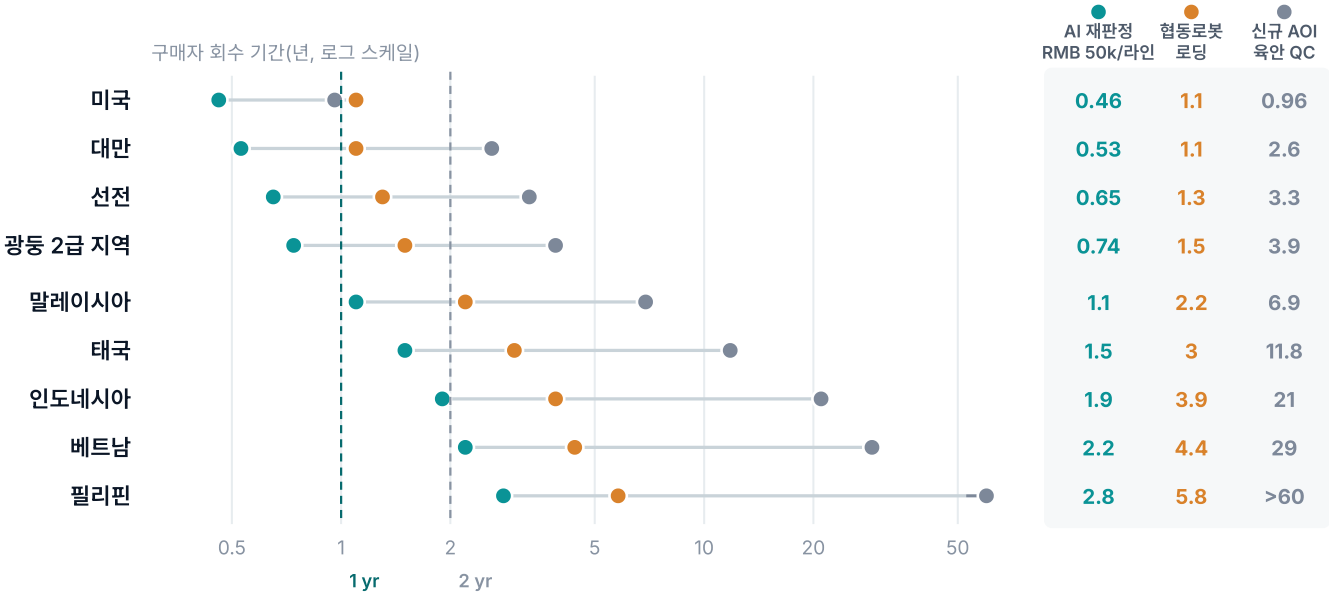
회수: 재판정은 수개월, 텐딩은 수년, 선별은 결코 안 됨

단순 회수 = 투자 ÷ (절감 인력 × 전액 부담 인건비 − 연 운영비). 보조금·관세 없는 2교대 라인 기준이다. 대만 장비는 중국의 1.3배, 베트남·말레이시아·Thailand는 1.15배, Indonesia·필리핀은 1.2배, 미국은 해외 가격으로 책정한다 estimate.

시나리오	투자(중국 / 미국)	절감 인력	연 운영비
(a) AI 재판정 박스 또는 AI 카메라 박스	라인당 RMB 50k / $30k	라인당 1.0	15%
(b) 수동 육안 검사를 대체하는 신규 인라인 2D AOI(교대당 재판정자 0.5명 잔류)	RMB 200k / $60k	1.0	12%
(c) 협동로봇 로딩 셀(스테이션 1개 × 2교대)	RMB 200k / $140k	2.0	8%
(d) 검출·선별 일체형 셀(검사 1.0 + 선별 0.07)	RMB 300k / $200k	1.07	8%
(e) NG 기판 선별 전용으로 추가된 로봇 암	RMB 150k / $100k	0.07	8%

지역	(a) AI 재판정	(b) 신규 AOI	(c) 협동로봇 셀	(d) 검출·선별	(e) 선별 전용 암
선전	0.65 yr (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	회수 안 됨
Guangdong 2급	0.74	3.9	1.5	5.3	회수 안 됨
대만	0.53	2.6	1.1	3.6	회수 안 됨
말레이시아	1.1	6.9	2.2	9.2	회수 안 됨
Thailand	1.5	11.8	3.0	15	회수 안 됨
Indonesia	1.9	21	3.9	25	회수 안 됨
베트남	2.2 (1.6–4.5)	29	4.4	34	회수 안 됨
필리핀	2.8	>60 (운영비 ≈ 절감)	5.8	≈100	회수 안 됨
미국	0.46	**0.96**	**1.1**	3.4	회수 안 됨

FIG 19 ***지역별 회수 기간AI 재판정은 미국·대만·선전에서 수개월, 동남아에서 1–3년 내 회수. 협동로봇 로딩은 수년.***



불량판 선별 전용 암: 모든 지역에서 회수 안 됨
직통률 90% 초과 시 선별은 교대당 0.003–0.14 FTE만

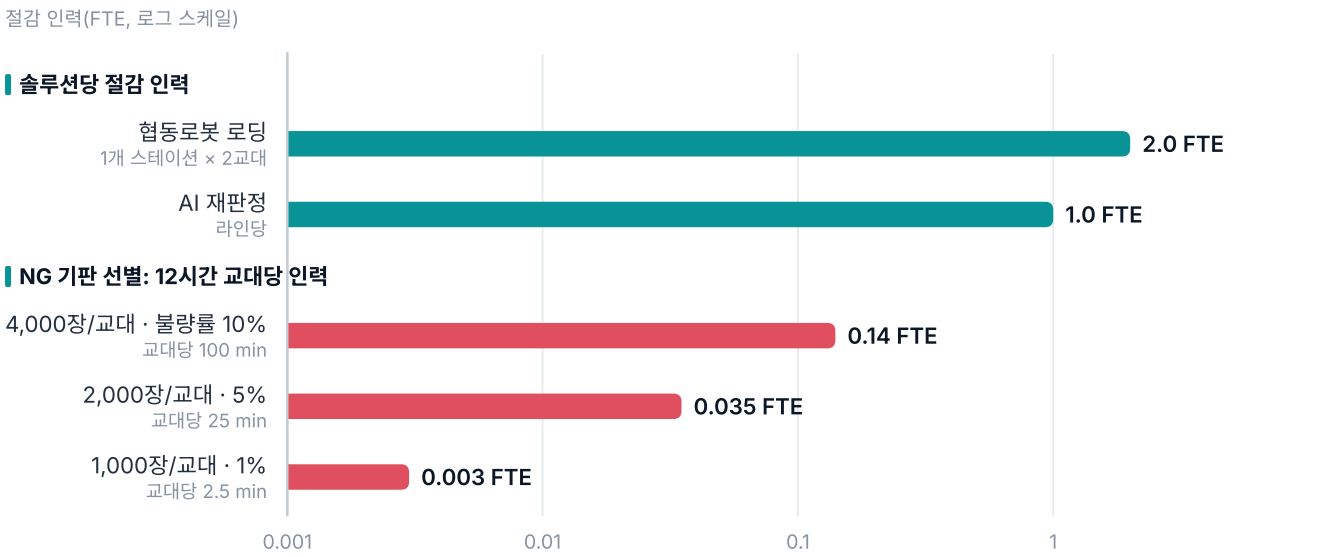
출처: 저자 모델: AI 재판정 박스 RMB 50k/라인; 협동로봇 로딩 셀 RMB 200k 또는 $140k, 2교대.

모든 값은 위 입력의 추정이다. 민감도: 절감 인력을 절반으로 하면 모든 회수 기간이 대략 두 배가 된다. 1교대면 (b)–(d)가 두 배, 미국의 100% 1년차 감가상각은 회수를 약 20% 단축하고, 중국·일본·한국 장비 관세는 약 10–40% 연장하며, 중국의 15% 지방 보조금은 약 15% 단축한다. 구매자 회수 임계값에 대한 독립 조사는 찾지 못했다. 벤더는 AOI가 "보통 9–18개월"에 회수된다고 하고(Averroes), Siemens 사례는 8개월을 주장한다 vendor claims. 저자는 1년과 2년을 선별 기준으로 쓴다. 이 모델은 품질 이득(보증, 반품)과 이직·교육 비용을 빼므로 보수적이다.

"검출과 선별"의 반직관적인 점은 노동이 재판정에 있지, 라인에서 실제 NG 기판 몇 장을 빼내는 데 있지 않다는 것이다. NG 기판당 처리 15초 기준 estimate:

교대당 기판	결함률	12시간 교대당 선별 시간	인력
1,000	1%	2.5분	0.003
2,000	5%	25분	0.035
4,000	10%	100분	0.14

FIG 20 **불량판 선별이 회수되지 않는 이유** ***직통률 90% 초과 시 선별은 교대당 0.003–0.14인분뿐.***



출처: 저자 추정.

기본 케이스(교대당 기판 2,000장, 결함 5%, 2교대)는 0.07명, 선전에서 연 약 RMB 6,000의 가치로, 암의 연 운영비 약 RMB 12,000보다 적다. 따라서 어디서도 결코 회수되지 않는다. 최악의 케이스(교대당 4,000장, 결함 10%)도 선전에서 약 13년, 미국에서 9년이 걸린다 estimate. 걱정이 선별 때문에 라인을 멈추는 것이라면, 표준 NG 분류기나 버퍼 컨베이어가 암 없이 해결한다. 업계 사람들은 "직통률 90% 이상이면 라인을 멈추고 손으로 꺼내라"고 말하는데, **선별에는 맞지만 검사에는 틀린 말이다**. 따라서 SMT 라인의 제품 설계는 "검사 + 마킹 또는 분류"이며, 로봇 암은 모든 부품을 이송하는 곳(테스트 지그 로딩, 트레이·매거진 핸들링, 포장)에 두는데, 거기서는 검사가 거의 공짜 애드온이 되고 시나리오 (c)가 적용된다.

보조금, 세금, 관세가 회수를 15–25% 움직이지만 순위는 바꾸지 않는다

중국 본토: 지원은 주로 자금 조달 비용을 낮추며, 단일 장비는 직접 해당되는 경우가 드물다. 국가 대출 보조금은 14개 우선 산업 사슬에서 최대 2년간 이자 1.5%p를 지원한다(Yicai) verified. RMB 300k 장비에서 약 RMB 9,000이다. Guangdong의 최대 20% 업그레이드 보상은 신규 장비 RMB 7 million 이상의 등록 프로젝트가 필요하며 선전은 제외된다(Guangdong rules). Dongguan은 2026년 7월 실제로 4.42%를 지급했다(Dongguan allocation) single source, primary. AI 검사 장치는 구매자의 등록 프로젝트 안에서 "지원 소프트웨어 포함" 장비로만 승인되므로, 기본 케이스는 보조금 없음으로 가정한다.

대만: 인센티브는 구매자에게 간다. 2025년 5월 AI를 포함하도록 개정된 산업혁신조례 제10-1조에 따라, 구매자는 2025–2029년 신규 스마트 기계 또는 AI(CNN/DNN 기반 검사 포함) 지출의 5%를 해당 연도 소득세에서 공제하거나, 3년간 연 3%를 공제할 수

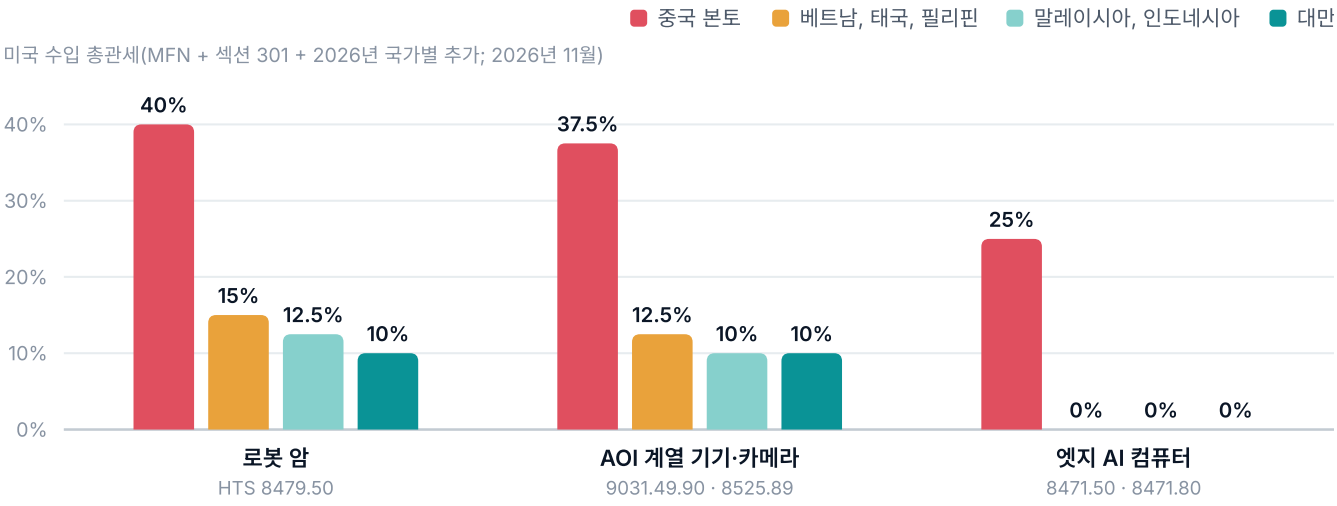
있다(Statute, credit rules) verified. 공급자 측 보조금은 중국 본토 자본에 닫혀 있다(MOEA) single source.

동남아시아: 말레이시아와 Thailand만 자동화 구매에 보상한다. 말레이시아의 첫 RM10 million에 대한 200% 자동화 자본공제는 흑자 구매자에게 capex의 약 24% 가치다(MIDA, PwC 말레이시아) verified. Thailand의 BOI는 업그레이드 투자의 50–100%까지 3년 면세를 주며, Thai 자동화 공급자가 관여하면 상한이 높아진다(BOI) single source, primary. 따라서 Thailand 진입은 현지 인테그레이터와 함께하는 게 낫다. 베트남·Indonesia·필리핀은 자동화 구매보다 공장 건설에 보상한다(베트남 Law Magazine, PwC 필리핀).

미국: 감가상각이 회수를 단축하고, 관세가 연장한다. 2025년 1월 19일 이후 구매 자산은 영구 100% 1년차 감가상각을 받으며, 흑자 구매자에게 약 21%의 연방 세금 방패다(IRS Notice 2026-11) single source, primary. 로봇·산업 기계에 대한 Section 232 조사는 2026년 10월 28일 신규 관세가 아닌 협상으로 끝났고 180일 내 재평가된다(White House, Mayer Brown) verified. 동남아시아 최종 조립은 암·AOI 계열 기기의 부가세를 약 37.5–40%에서 10–15%로 낮춘다(1장).

지역	주요 수단	구매자 회수에 미치는 영향 estimate
중국 본토	최대 2년간 1.5%p 대출 보조금; 성·시 업그레이드 보상(Guangdong은 ≥RMB 7M 등록 프로젝트 필요; Dongguan 실제 4.42% 지급); 선전 2026–2029 조치(요율 비공개)	단일 장비 약 0–5%; 대형 프로젝트 안에서 4–15%
대만	조례 §10-1 공제, 5% 또는 3% × 3년	구매자가 세금을 충분히 내면 약 5% 단축
말레이시아	200% 자동화 자본공제(첫 RM10M)	흑자 구매자 약 15–25% 단축
Thailand	BOI 3년 면세, 투자의 50–100%까지	세금 면제 기간 밖 흑자 구매자 약 15–25% 단축
베트남, Indonesia, 필리핀	공장 건설 소득세 인센티브; 필리핀 장비 추가 20% 감가상각	단일 자동화 구매에는 거의 없음
미국	100% 1년차 감가상각; 중국 본토산 장비 ≈37.5–40% 종합 관세	감가상각 ≈20% 단축; 관세 ≈10–40% 연장

FIG 21 ***로봇·AOI·엣지 컴퓨터의 미국 관세 중국 본토산 암은 약 40%. 동남아·대만 출하는 10–15%.***



출처: USITC 관세율표 제99장; Federal Register; USTR(2026년 11월).

지역별로: 중국 본토에서 AI 재판정 박스는 수개월, 협동로봇 셀은 1.3–1.5년에 회수되는 반면, 노동 절감만을 위해 구매한 신규 AOI는 3–4년, 검출·선별 일체형은 4.5–5.3년으로 모두 2년 선별 기준을 넘는다. 대만은 선전보다 약간 낫다(재판정 0.53년, 협동로봇 셀 1.1년). 채용도 더 어렵다. 말레이시아와 Thailand에서는 AI 박스가 1–1.5년, 협동로봇 셀이 2–3년 걸리며, 흑자 구매자는 현지 세제 혜택으로 약 15–25% 빨라진다. Indonesia·베트남·필리핀에서는 아주 저렴한 AI 박스만 2–3년 내 회수되고, 수요는 고객의 품질 요

구와 노동 부족에서 온다. 미국에서는 순수 선별 전용 암을 제외한 모든 선택지가 0.2–3.4년에 회수되며, 관세·Section 232 재평가·NRTL 인증이 주요 변수다.

05 시장과 경쟁
시장, 경쟁, 그리고 구매자의 지불 방식

시장 규모: 공개 수치는 배수로 갈린다. 유용한 것은 보험업 집계뿐이다

모든 시장 수치는 범위와 함께 읽어야 한다. 중국 머신 비전 시장은 2024년 RMB 18.15 billion이며(Zhiyan) verified, 최대 구매처는 3C 전자다. 머신 비전의 글로벌 추정은 약 3배, AOI는 약 6.5배 차이 난다(MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Koh Young, TRI, Jutze, Sinic-Tek, ViTrox의 매출로 보험업하면 글로벌 SMT 검사 장비는 $1–2 billion 규모다 estimate.

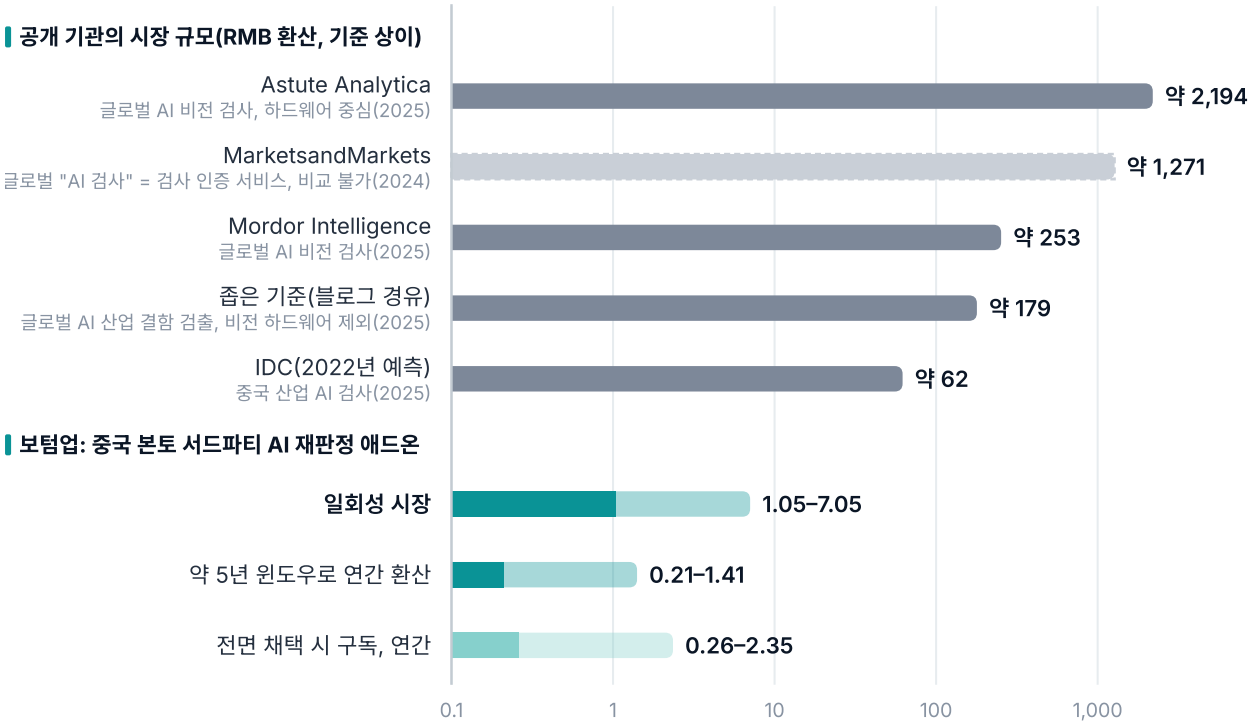
"AI 비전 검사 소프트웨어"라는 독립 시장은 없다. 공개 수치는 10배 이상 차이 나며 모두 하드웨어나 서비스를 포함한다.

범위	발행처	수치
글로벌 "AI 산업 결함 검출"(좁은 기준, 비전 하드웨어 제외)	미공개 업체, iFactory 블로그 경유	2025년 $2.66 billion, 2035년까지 CAGR 8.6%, 전자 34%
글로벌 "AI 비전 검사"	Mordor Intelligence	2025년 $3.76 billion, 2031년 $12.8 billion, CAGR 22.67%
글로벌 "AI 비전 검사"(대부분 하드웨어)	Astute Analytica	2025년 ≈$32.66 billion
글로벌 "AI 검사"(시험·인증 서비스; 비교 불가)	MarketsandMarkets	2024년 $18.92 billion
중국 "산업 AI 품질 검사"	IDC, 2022년 예측	2025년 $958 million(≈RMB 6.2 billion), 통신·전자 및 두 산업이 91.5%

출처: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) each single source, mostly paid-report summaries.

FIG 22 **시장 규모는 10배 이상 차이*의사결정에 유용한 것은 좁은 기준과 보험업 집계뿐.***

RMB 100M(로그 스케일; USD 6.7188 환산)



출처: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC(2022), iFactory; 저자 보험업 추정.

보험업으로 보면 SMT 라인의 공개 집계는 하나뿐이다. New Venture Research는 2020년 전 세계 약 73,231개 라인, 그중 47%가 중국(2000년 19%), 약 70% 하이엔드·30% 로우엔드로 추정했다(EMSNow, NVR SMT2020 summary) single source, 2020. NVR은 이제 IPC 소유이며 최신 데이터는 유료다(NVR). 연 3–5% 성장과 40–47% 중국 점유율로 계산하면 2026년 전 세계 약 87,000–98,000개, 중국 본토 약 35,000–46,000개(기준 40,000)다 estimate.

입력	저	기준	고	근거
중국 본토 SMT 라인(2026)	35,000	40,000	46,000	NVR 2020 기준 × 성장 × 점유율
AI 재판정 미적용 비중	75%	80%	85%	아래 개조 윈도우
서드파티가 도달 가능한 비중(사내·OEM AI에 묶이지 않고 지불 가능)	40%	50%	60%	가정: ≈30% 로우엔드 라인 + 일부 복수 브랜드 하이엔드 라인, DBG식 사내 구축 제외
대상 라인	**10,500**	**16,000**	**23,500**	곱
라인당 일회성 가격	RMB 10k	RMB 20k	RMB 30k	가격 레더와 상한(4장)
일회성 대상 시장(설치 기반)	**RMB 105M**	**RMB 320M**	**RMB 705M**	곱
≈5년 윈도우의 연간	RMB 21M	RMB 64M	RMB 141M	윈도우 연 8–12% 축소
구독(라인·연 RMB 2.5k / 5k / 10k, 전면 채택)	연 RMB 26M	연 RMB 80M	연 RMB 235M	Baidu식 연회비 기준점

FIG 23 **보험업: 40,000라인 → 대상 16,000 → 일회성 약 RMB 320M*중국 본토의 서드파티 AI 재판정 틈새는 실재하지만 작다.***



신규 진입자가 3년 내 대상 라인의 2–5% 확보: 약 320–800라인, 연 RMB 2–5M
구독(라인·연 RMB 5k)이면 RMB 1.6–4M 경상 수익, IDC 중국 산업 AI 검사 예측의 0.3–2.3%에 불과

출처: NVR(2020 기준); 저자 추정.

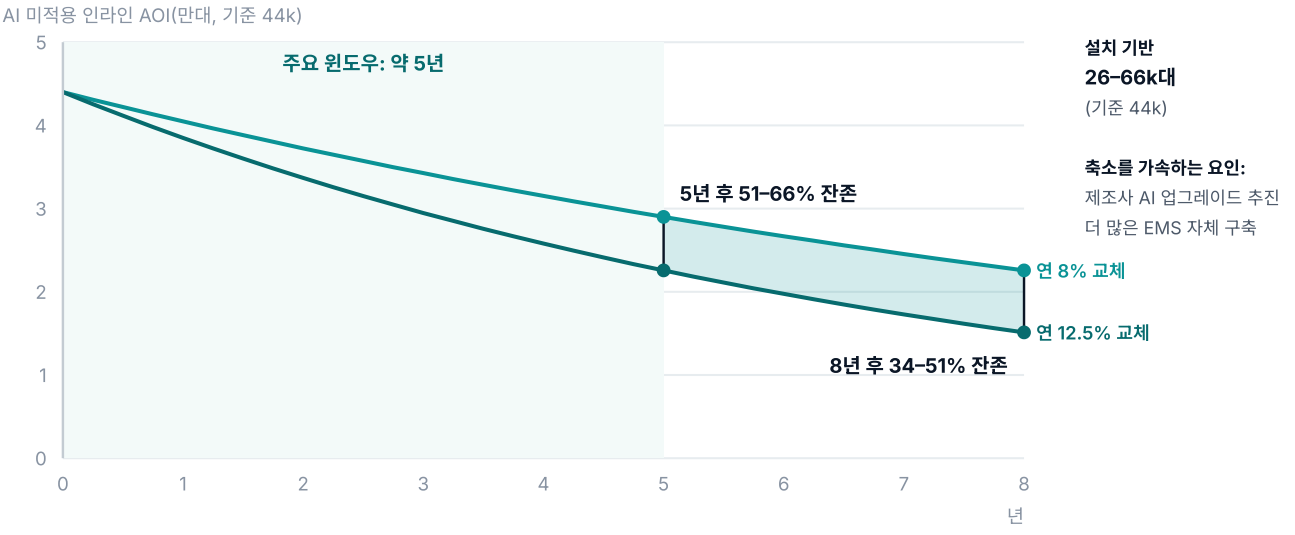
이 모든 것은 추정이다. 비중도 가격도 공개적으로 측정된 적이 없다. 이 틈새는 IDC가 예측한 2025년 중국 산업 AI 품질 검사 전체(약 RMB 6.2 billion)의 연 0.3–2.3%에 불과하다. 신규 진입자가 3년 안에 기준 대상 라인의 2–5%를 확보하면 약 320–800개 라인, 라인당 RMB 20,000로 총 RMB 6.4–16 million, 연 약 RMB 2–5 million, 또는 라인당 RMB 5,000의 경상 수익으로 연 RMB 1.6–4 million이다 estimate. 이는 회사 내부 사업부에는 충분하지만 벤처 규모의 독립 회사에는 부족하다. 대만·동남아시아·미국의 설치 라인 1,000개당 중국 본토 가격 기준 일회성 약 RMB 3–15 million이다. 대만과 미국은 가격이 더 높지만, 그곳 구매자는 AI 옵션이 있는 OEM 3D AOI를 더 자주 쓴다 estimate.

개조 윈도우는 약 5년이다

AOI 제조사 중 장비 수명을 공개하는 곳은 없다. EMS는 장비를 3–10년에 감가상각하고(DBG 2025 annual report), 구형 AOI는 수명 종료 후 소프트웨어 업데이트와 부품을 잃는다(I-Connect007) single source. 저자는 8–12년(기준 10)을 쓴다.

라인당 인라인 AOI 1.0–1.65대로 계산하면 중국 본토에는 약 35,000–78,000대의 인라인 AOI가 가동 중이고(기준 약 55,000), 이는 알려진 출하량과 일치한다(Circuits Assembly, Sinic-Tek 2025 annual report). AI 재판정을 돌리는 대수를 측정한 출처는 없다. 국산 장비가 약 2023년부터 제공해 왔고 최근 인도의 40–60%가 사용한다고 가정하면, **중국 본토에서 AI를 돌리지 않는 인라인 AOI는 약 26,000–66,000대(기준 약 44,000)**다 estimate. 모든 신규 장비가 이제 AI를 탑재해 출하된다면 이 풀은 연 8–12.5% 축소되어 5년 후 51–66%, 8년 후 34–51%가 남는다. AOI 제조사가 신형 장비에 AI 업그레이드를 추진하면 더 빠르다. **기회의 대부분은 향후 약 5년에 몰려 있으며**, 핵심은 구형 국산 2D AOI, 복수 브랜드, AI 업그레이드 경로가 없는 중견 EMS다.

FIG 24 **개조 윈도우는 약 5년*AI 미적용 인라인 AOI는 교체·업그레이드로 연 8–12.5% 축소.***



출처: 저자 추정; 설치 기반·점유율은 독립 검증 없음.

경쟁자는 이미 강하다: 재판정 쐐기는 포위당했다

AI 재판정 쐐기는 세 방향에서 좁아지고 있다. AOI·비전 선두 업체가 신규 플랫폼에 AI를 내장하고 있다. Jutze는 "AI 알고리즘을 전면 채택"했고(Jutze investor relations, June 2026), OPT의 산업 AI 매출은 2025년 96.77% 성장했으며(OPT 2025 annual report) single source, primary, ViTrox는 자율 재판정 정확도 최대 95%를 주장한다(ViTrox) vendor claim. 대형 제조사는 자체 구축한다(iThome). 벤처 자본을 받은 AI 기업은 규모로 판다. SmartMore의 엣지 센서 고객은 2년 만에 65곳에서 251곳으로 늘었고(36Kr) single source, Hikvision은 검사용 Guanlan 모델을 판매하며(Xinhua), SmartMore·Hikrobot·Mech-Mind는 이미 "눈·뇌·손" 통합을 제공한다(36Kr, Securities Star).

FIG 25 ***재판정 쐐기는 삼면에서 포위서드파티의 여지는 중견 EMS와 AOI가 약한 검사 틈새에 있다.***



출처: Jutze, Sinic-Tek, DBG 연차보고; Hikvision·SmartMore 공개 자료; 저자 편집.

신규 진입자에게 남은 공간은 좁다. 구형 2D/3D AOI를 갖고 제조사 AI 업그레이드가 없는 중견 EMS; 여러 AOI 브랜드를 섞어 쓰는 라인(한 개조 벤더는 이미 Omron, VI, Koh Young, Mirtec을 지원한다(Sohu on Hengtianyi)); 스루홀·수납땜 접합, 컨포멀 코팅, 케이블·커넥터·FPC, 최종 외관·나사처럼 SMT AOI가 잘 못 보는 틈새; 그리고 RMB 230–280k AOI를 감당 못 하는 소형 공장이다.

가장 가까운 상장 비교 기업이 경고를 준다. AI 검사 장비와 로봇을 파는 Weiyi Zhizao는 2026년 1–5월 3C 매출이 전년 동기 대비 79.9% 감소했고, "경쟁 심화와 가격 압박"을 인정했으며, 2026년 10월 25일 홍콩 상장을 연기했고, 완료 전 종료된 프로젝트 15건을 공시했는데 그중 12건이 임바디드 로봇 프로젝트다(Weiyi Zhizao prospectus, postponement notice) single source, primary. AInnovation은 2025년을 매출채권의 29.8%를 손상 처리하며 마감했다(AInnovation 2025 annual report). 3C 검사 쐐기는 붐비고, 라인당 가격 협상력은 약하다.

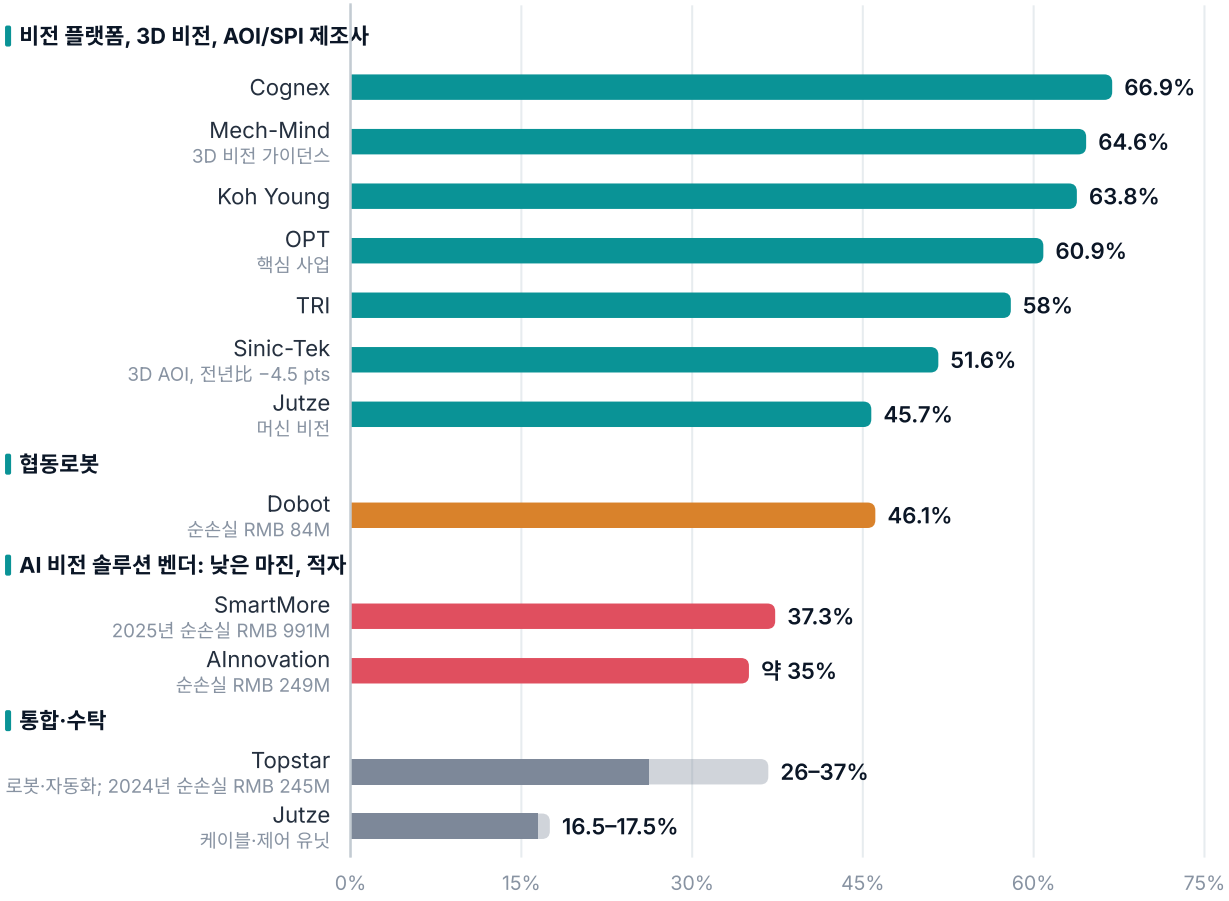
마진은 제품화된 하드웨어에 있다

사슬 내 위치	대표 기업과 2025년 매출총이익률	이익과 회수
비전 플랫폼	Cognex 66.9%	흑자; 저마진 매출 ≈$22M 매각 중
AOI/SPI 장비	Koh Young 63.8%; TRI 58%; OPT 핵심 사업 60.85%; Sinic-Tek 3D AOI 51.62%(전년比 −4.53점); Jutze 머신 비전 45.74%	흑자; 매출채권 회전일수 139(Jutze)~247–274(OPT)
3D 비전 가이던스	Mech-Mind 64.6%	—
AI 비전 솔루션 벤더	SmartMore 37.3%; AInnovation ≈35%	적자: SmartMore 2025년 순손실 RMB 991M, AInnovation RMB 249M

사슬 내 위치	대표 기업과 2025년 매출총이익률	이익과 회수
협동로봇	Dobot 46.1%	Dobot 순손실 RMB 84M; UR 매출 $326M(2022) → ≈$232M(2025, 예측)
시스템 통합	Topstar 로보틱스·자동화 26.2–36.7%	순손실 RMB 245M(2024); 매출채권 회전일수 180–240
수탁 하드웨어	Jutze 케이블·제어 유닛 16.5–17.5%	—

출처: (Cognex, StockAnalysis on Koh Young, Fugle on TRI, OPT 2025 annual report, Sinic-Tek 2025 annual report, Jutze 2025 annual report, Jiemian on Mech-Mind, 36Kr on SmartMore, Yicai on SmartMore, TMTPost, DoNews on AInnovation, The Paper on Dobot, The Robot Report on UR, Jiemian on Topstar).

FIG 26 **마진은 제품화된 하드웨어에 있다*비전 플랫폼·장비 제조사는 매출총이익률 46–67%. AI 솔루션 벤더는 35–37%에 적자.***



출처: 2025년 연차보고 또는 증권신고서.

신규 진입자는 검사를 제품·장비·소프트웨어로 만들어야만 마진을 지킨다. 프로젝트별 튜닝은 마진을 약 35%로 끌어내리고 지속적 적자를 만든다. 구조도 소기업에 불리하다. 매출은 현장 검수 후에만 인식되고 선수금은 작으며(Jutze, OPT), 파일럿은 보통 무료이고(A3/Mindtrace), 브랜드 소유주가 계약 제조사의 라인에 검사 장비를 지정하는 경우가 많다.

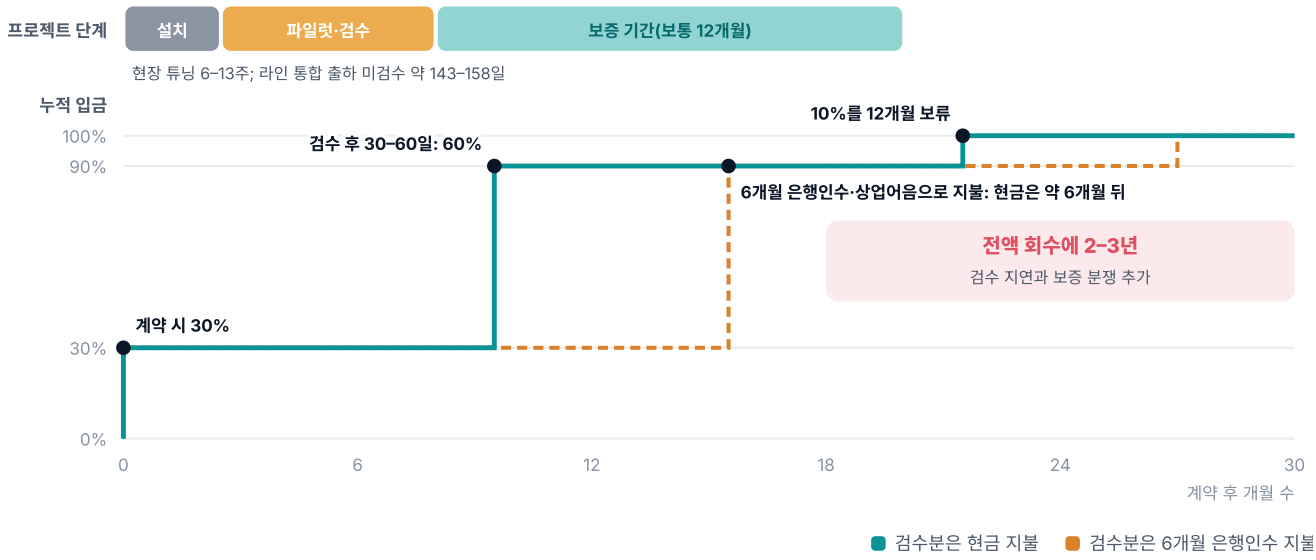
구매자의 지불 방식: 계약 시 30%, 검수 후 60%, 10%는 1년 보류

중국 본토 전자 공장이 검사 장비에 쓰는 지불 조건은 Sinic-Tek의 상장 문의 답변에 공시된 13개 실명 구매자 계약에 명시되어 있다(Sinic-Tek inquiry reply 8-1-1) single source: one primary filing, several independent buyers, 2020–2022 .

구매자(연도)	계약	인도	검수	보류 또는 잔금
Hikvision 계열(2021)	15일 내 30%	—	최종 검수 후 30일 내 65%	보증 보류 5%를 검수 후 12개월 지급; 일부는 6개월 이상 은행 인수어음
Nanchang Huaqin(2020–2021)	10영업일 내 30%	—	검수 후 60일에 60%	1년 보증 후 10%
Guangdong Yinuo(2021, 3D AOI)	10영업일 내 30%	—	30일 내 60%	무고장 3개월 후 10%
Xiamen Hongxin, Jingmen Hongyi(2022)	10영업일 내 30%	—	최종 검수 후 30영업일 내 60%, 6개월 어음	1년 보증 후 10%
AKM Panyu, Suzhou(2021–2022)	7–10영업일 내 30%	—	검수·송장 후 1개월 60%, 180일 어음	360일 후 10%
BYD Precision, 선전, Huizhou(2020)	—	도착 시 30%	검수·송장 후 70%	전액 6개월 상업어음
선전 H&T(2020)	—	—	검수·송장 후 90%	무고장 6개월 후 10%

다른 1차 공시도 같은 구조를 보여준다. Mech-Mind의 대형 프로젝트는 계약 시 30%, 인도 전 30%, 검수 후 30%, 보증 보류 10%를 12–15개월에 걸쳐 받는다(Mech-Mind prospectus). GKG는 장비 업계 보류가 보통 계약의 5–10%이며 검수 후 1년에 지급된다고 한다(GKG inquiry reply) single sources, primary. 따라서 표준은 계약 시 약 30%(20–30%, BYD 같은 일부 공장은 0), 검수 후 60–65%(30–60일 후 지급, 종종 6개월 은행·상업어음으로 지급되어 현금화가 반년 더 늦어짐), 보류 5–10%를 3–12개월(12개월이 가장 흔함)이다 verified across several contracts.

FIG 27 **구매자의 지불 방식*계약 시 30%, 검수 후 60%, 10%는 1년 보류: 전액 회수에 2–3년.***



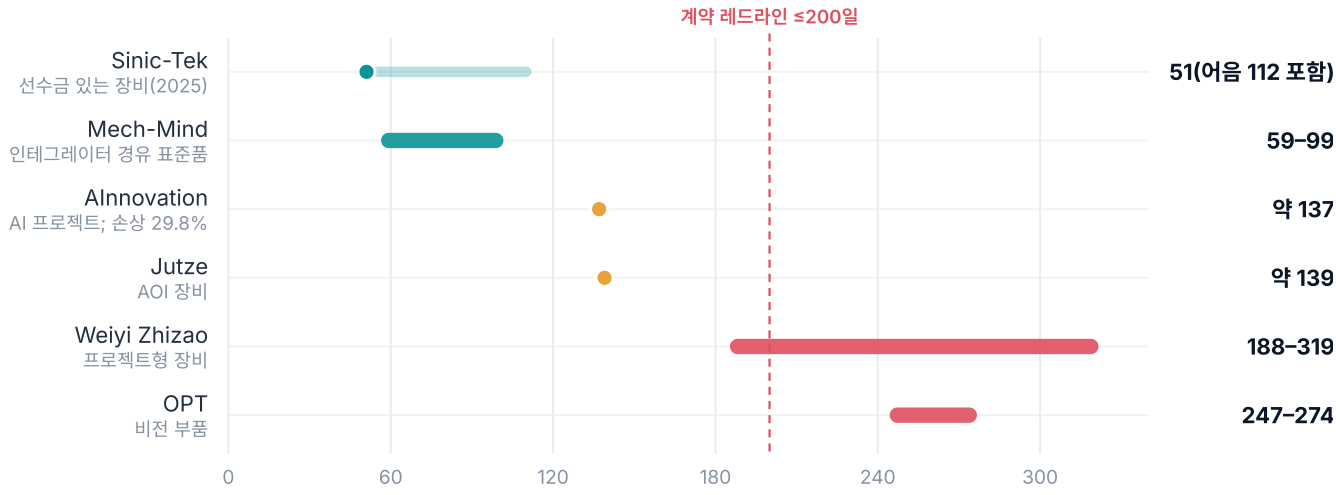
출처: Sinic-Tek 문의 답변, Mech-Mind, GKG, JT 등; 일반 조건 개요, 시기는 추정.

검수와 시운전은 느리고 판매자가 부담한다. Sinic-Tek의 출하 후 미검수 재고는 약 143–158일이 걸렸는데, 대형 EMS와 브랜드 고객이 전체 라인 통합 테스트와 연장 시운전을 요구하기 때문이다(Sinic-Tek inquiry reply 8-1-2). Weiyi의 AI 프로젝트는 평가에서 배포까지 4–8개월 걸린다(Weiyi Zhizao prospectus) single sources, primary. AI 재판정 파일럿의 전환 데이터는 없다. Sinic-Tek은 데모 장비의 약 18–19%만 직접 판매로 전환했다. 저자는 무료 파일럿은 3–5개 중 1개, 쇼트리스트 후 유료 파일럿은 약 80% 전환을 계획한다 estimate.

회사	비즈니스 모델	매출채권 회전일수
Sinic-Tek(2025)	예치금 있는 장비	≈51(어음 포함 ≈112)
Mech-Mind	인테그레이터 경유 표준품	59–99
Jutze	AOI 장비	≈139
AInnovation(2025)	AI 프로젝트	연말 ≈137, 손상 29.8%
Weiyi Zhizao	프로젝트형 장비	188–319
OPT	비전 부품	247–274

출처: 위 연차보고서와 증권신고서; 매출채권(계약자산 포함) ÷ 매출 × 365 primary + estimate.

FIG 28 **매출채권 회전일수*표준품 51–99일; 프로젝트형 장비 188–319일.***



출처: 기업 자료; 매출채권(계약자산 포함) ÷ 매출 × 365.

AOI 제조사나 인테그레이터를 통해 판매되는 AI 재판정 라이선스는 Mech-Mind식 60–100일에 회수되고, 직접 프로젝트 판매는 Weiyi식 190–320일이다. 현금 타임라인으로 합치면: 무료 파일럿 1–3개월, 대형 ODM 자격 심사 최대 약 12개월, 구축·인도·검수 3–7개월. 30%는 주문 시, 60%는 약 4–9개월 후, 마지막 10%는 약 16–21개월 후 도착한다. **대형 EMS의 첫 파일럿부터 전액 회수까지는 2–3년이 걸릴 수 있다.** 정식 입찰이 없는 중견 EMS는 자격 심사를 몇 주로 단축할 수 있다 estimate. 공시된 계약의 검수 기준은 전체 라인 통합, 처리량(대만 프로젝트 하나는 시간당 100대 요구), 무고장 운전 개월만 다루며, 유출률이나 오탐률을 명시한 것은 없다(Anda inquiry reply) single source, primary. AI 재판정의 서면 검수 기준은 파일럿 전에 합의해야 한다.

실패의 네 가지 패턴

기록된 실패는 네 가지 패턴으로 나뉜다. 제품이 신뢰성을 갖추기 전에 출하(Rethink(The Robot Report)), 선두 업체 하드웨어 위의 소프트웨어 계층이 나중에 그 생태계에 대체됨(READY Robotics(The Robot Report)), 사이클 바닥에서 자금 고갈(Dextrous Robotics(Robots & Startups)), 모델 회사가 인수합병(acqui-hire)으로 끝남(Covariant(GeekWire)). 교훈은: 프로젝트별 튜닝 대신 좁게 제품화하고, 검수 기반 매출과 긴 회수에 맞는 운전자본을 유지하며, 24시간 가동 공장에 현지 서비스를 제공하고, 제품이 검증되기 전에는 확장하지 말라는 것이다.

타이밍: 검사는 기다려도 얻을 게 없고, 암 하드웨어는 저렴해진다

기다림의 논거는 약하다. 2026년 하드웨어는 저렴해지지 않았다. 로봇 제조사는 가격을 5–15%, 일부 제품은 25–50% 올렸고(Sina Finance) single source, NVIDIA 키트는 세 번째로 올랐다(36Kr). 생산급 VLA는 윈도우 안에 도착하지 않는다(IFR,

Instrumental). 그 사이 AOI 교체마다 개조 대상이 사라지고, OEM AI 배포마다 라인의 데이터와 프로세스가 고정된다. 모델 능력은 상품화되고 있으며, 희소한 것은 공장 접근성과 라벨링된 공장 데이터다. 현지 지원에도 윈도우가 있다. 선전의 2026년 "AI 선도 도시" 규칙은 AI 응용 시범 프로젝트에 최대 RMB 2 million, 구매의 30%를 커버하는 모델 바우처를 제공하며 **2027년 6월 30일**까지다(Yicai, May 2026) terms single source.

여전히 빠르게 저렴해지는 것은 휴머노이드와 손재주 손이다(손은 2년 만에 약 RMB 50,000에서 10,000 미만으로 내렸다(Sina Finance)). 제품 믹스가 아래로 이동하면서 엔트리 협동로봇도 내려간다. 이는 임바디드 부분을 미루는 논거이지 검사 부분을 미루는 논거가 아니다. **종합: 검사는 지금, 좁은 제품으로. 로봇 셀은 지금, 파트너와 통합을 통해. 임바디드 AI는 나중에, 오늘은 R&D 스테이션 하나로.**

임바디드 AI

06 임바디드 AI: 먼저 로봇을 가르치는 방식이 바뀌고, 나중에 움직임이 바뀐다

바뀌는 것은 인지와 프로그래밍이다. 모션은 결정론적으로 남는다

모든 로봇 암 선두 업체가 이제 "피지컬 AI"를 마케팅하지만, 실제 출하된 것은 세 계층으로 나뉜다. 교시 없는 학습된 인지(FANUC의 AI Masterless Vision, 2027년 1월 출하(FANUC); KUKA의 신경망 그랩 선택(KUKA)); 프로그램을 작성하고 과제를 계획하는 대형 모델(KUKA의 iiQWorks.Copilot은 AI 생성 코드를 디지털 트윈에서 검증해야 한다고 경고한다(KUKA); Yaskawa의 Gemini Robotics-ER은 "생성형 AI가 무엇을 할지 결정"하고 로봇이 "어떻게 움직일지" 결정한다(Yaskawa)); 그리고 컴퓨트·데이터 플랫폼(Jetson Orin을 단 UR의 AI Accelerator(Universal Robots)) vendor claims. AI 기능의 생산 성공률을 공개하는 선두 업체는 없고, 일본 선두 업체들은 프런티어 VLA를 자체 구축하기보다 Google, SoftBank, Microsoft 등과 협력한다(FANUC news index). 학습된 정책 계층은 모델 회사에서, 인증된 암과 컨트롤러는 암 제조사에서 나온다.

FIG 29 **임바디드 AI 스택** *대형 모델은 오프라인에서 계획, 소형 정책은 셀에서 실행, 모션·안전은 결정론적 유지.*



중소기업의 기회는 통합·센싱 하드웨어: 학습형 인지와 소형 정책을 인증 로봇 셀에 넣어 스테이션별로 도입

출처: 저자 종합(제3·6장); ISO 10218:2025; EU 기계 규정.

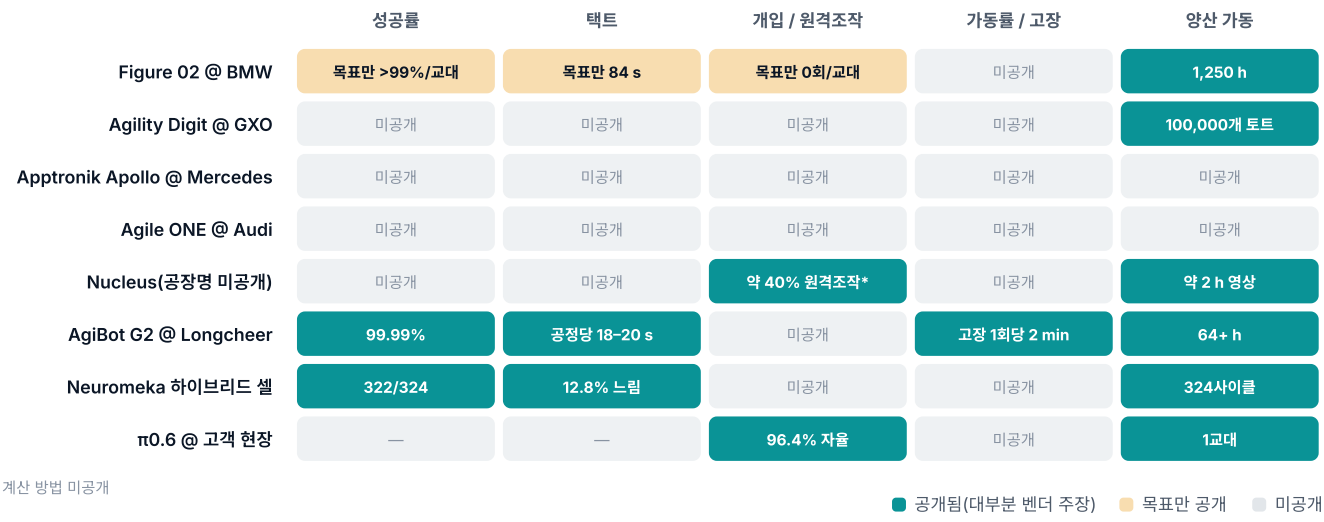
비구조화·유연·소량 작업과의 거리: 구조화 스테이션은 1–3년, 유연 조립은 5년+

실제 공장 수치가 나오기 시작했지만, 벤더 보고이고 좁으며 구조화 스테이션에서 나온 것이다. AgiBot의 휴머노이드 8대가 Longcheer 태블릿 최종 테스트 라인에서 64시간 이상 64,828개 과제를 99.99%로 기록했다(AgiBot, Yicai Global) vendor and customer. 그러나 이는 비전을 갖춘 기존 암이 할 수 있는 텐딩 작업이다. UBTech는 Walker S2가 사람 효율의 30–50%로 일한다고 한다(Irish Times/FT) single source. BMW Spartanburg에서 Figure 02의 11개월은 물량 수치(1,250시간+, 부품 90,000+)를 만들었지만, 사이클·성공·개입 수치는 결과가 아니라 목표다(Figure, BMW Group) verified, numerators only. Nucleus는 공장 영상의 약 40%가 텔레오퍼레이션이라고 한다(Interesting Engineering) vendor claim. Reuters 조사는 일부 중국 휴머노이드가 작업자 산출의 약 20%라고 인용했다(Notebookcheck) single source. 성공률, 사이클 타임, 개입률을 함께 공개하는 배포는 없고 감사된 것도 없지만, 그 세 가지가 바로 회수 계산에 필요한 것이다.

배포	성공률	택트	개입 또는 텔레오퍼레이션	가동률 또는 고장 간격	물량 가동
Figure 02 @ BMW Spartanburg	목표만(교대당 >99%)	목표만(84초)	목표만(교대당 0)	미공개(전완 상단이 고장점)	1,250시간
Agility Digit @ GXO	미공개	미공개	미공개	미공개	100,000 토트
Apptronik Apollo @ Mercedes-Benz	미공개	미공개	미공개	미공개	미공개
Agile ONE @ Audi	미공개	미공개	미공개	미공개	미공개
Nucleus(공장명 미공개)	미공개	미공개	≈40% 텔레오퍼레이션(방식 미공개)	미공개	≈2시간 영상
AgiBot G2 @ Longcheer	64,828개 과제의 99.99%	단계당 18–20초	미공개	통신 고장 1회, 복구 ≈2분	64시간+
Neuromeka 하이브리드 셀	322/324	사람보다 12.8% 느림	미공개	미공개	324사이클
Physical Intelligence π0.6 @ 고객 현장	—	—	교대 중 96.4% 자율	미공개	1교대

출처: 이 장과 3장 mostly vendor claims .

FIG 30 **공장 내 로봇 배포가 실제로 공개하는 것*대부분 수량만 공개. 성공률·택트·개입률을 함께 공개하는 사례는 없음.***

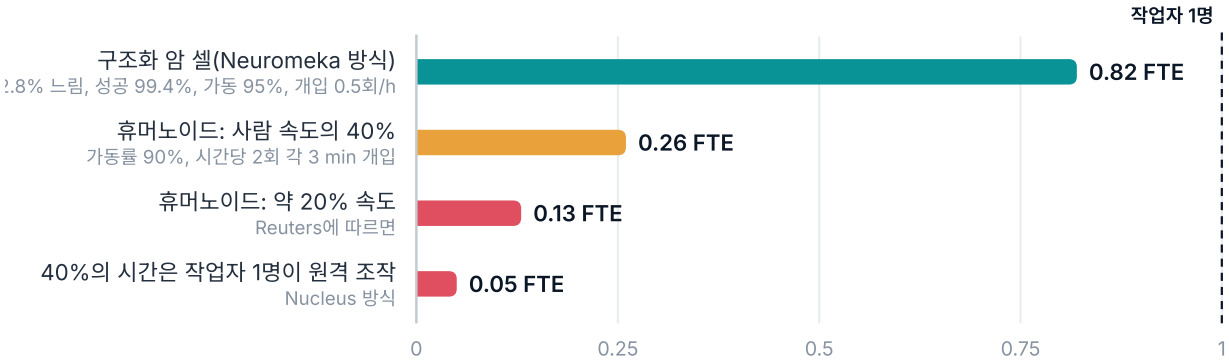


출처: BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence 등의 공개 자료.

텔레오퍼레이션은 성공률보다 경제성에 더 중요하다. 로봇당 교대당 순 대체 인력 ≈ (사람 사이클 ÷ 로봇 사이클) × 가동률 × 성공률 − 개입 시간 − 텔레오퍼레이션 estimate, my formula . Neuromeka식 구조화 스테이션(12.8% 느림, 성공 99.4%, 가동률 95%, 시간당 2분 개입 0.5회)은 약 0.82명을 대체한다. 사람 속도의 40%, 가동률 90%, 시간당 3분 개입 2회인 휴머노이드는 약 0.26명. Reuters의 20%에서는 약 0.13명. Nucleus처럼 작업자 1명이 40%의 시간을 텔레오퍼레이션하면 약 0.05명으로 거의 0이다. 중국 본토에서 휴머노이드 대당 약 RMB 600k, 인당 연 RMB 85k 기준으로 마지막 셋은 수십 년에 걸쳐 회수된다. 구조화 암 스테이션만 현실적 회수를 갖는다(4장) estimate .

FIG 31 ***로봇당 순 대체 인력 구조화 암 셀은 약 0.82인분 대체. 40% 원격조작 휴머노이드는 약 0.05인분.***

로봇 1대·교대당 순 대체 인력



출처: 저자의 수식 기반 추정.

구조화 스테이션을 벗어나면 범용 모델은 신규 과제에서 여전히 40–90%다. Gemini Robotics 2는 랩 픽앤플레이스 74.2%, 툴 키팅 78.9%, 봉지 묶기 44%에 도달한다(eesel/DeepMind). 30개 실기 로봇 과제인 RoboChallenge Table30에서 π0.5는 43.7%, 최고 오픈 모델은 50.3%다(RoboChallenge, QbitAI).

외부 전망은 갈리지만, 2년 내 광범위한 사용을 지지하는 것은 없다. IFR은 범용 피지컬 AI가 "아직 산업급이 아니며" 대부분의 휴머노이드가 "여전히 프로토타입 또는 파일럿"이라고 한다(IFR). Interact Analysis는 2025년 생산된 휴머노이드의 약 10%만 실제 사용에 들어갔고 상업적 전환점은 "2032년 이후"로 추정한다(automation.com). UBS는 2030년 이전 휴머노이드의 "EV 순간"은 없을 것으로 본다(Yicai Global). 중국 내부자는 공장 규모 사용까지 2–3년 더 걸린다고 추정한다(Gasgoo) single sources. 저자의 판단 estimate: 자원이 풍부한 전자·자동차 공장에 벤더가 상주하는 경우, **구조화 스테이션 과제**(텐딩, OK/NG 분류, 지그 있는 삽입)는 약 1–3년 안에 사람 속도에 가깝게 ≥99%에 도달한다. 범용 VLA를 쓴 **혼합 부품 키팅**은 약 3–5년, **자유 배선 하네스와 다단계 유연 조립**은 5년 이상이다. 휴머노이드 회수는 오늘 작동하지 않는다. Walker S2는 약 RMB 600k(10jqka) single source, 2교대에 사람 효율의 30–50%로 약 0.6–1.0명에 해당하며, 중국 본토 인건비 기준 통합·유지보수·다운타임 전에 단순 회수 7–12년이다 estimate.

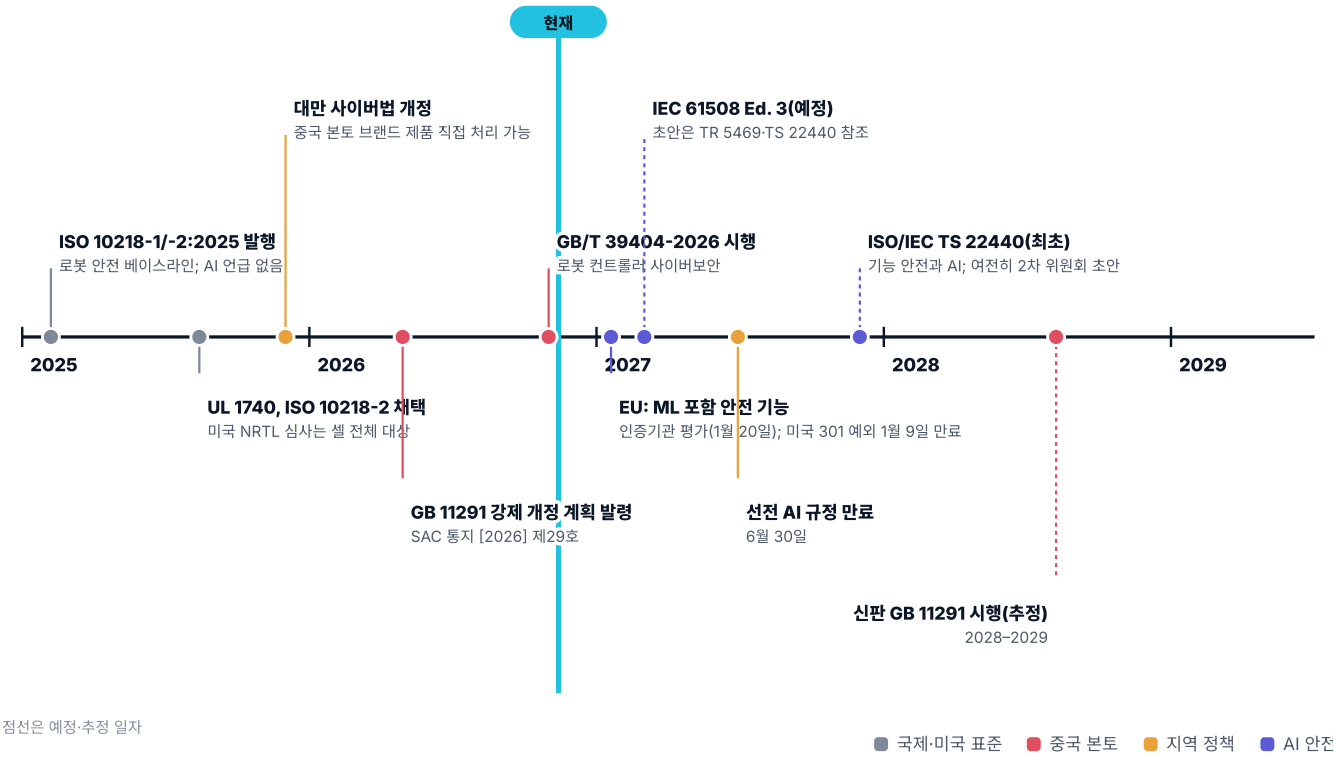
막힌 곳

고리	2026년 증거
데이터	업계는 고품질 실제 데이터 최소 1,000만 시간이 필요하고 2026년 초까지 약 50만 시간만 수집됐다고 추정한다(Gasgoo). Reuters는 약 100 million 시간 필요 대비 약 500,000시간을 보도한다(Notebookcheck). 초보 텔레오퍼레이터는 쓸 만한 모션당 약 300회 시도가 필요하다. 공장은 공정 데이터를 공유하지 않는다. 시뮬레이션으로 학습한 정책은 실기에서 급락한다
신뢰성	구조화 스테이션 98–99.99%(벤더 주장, 실패 정의 없음); 신규 과제에서 범용 모델 40–90%; 미확인 커넥터 29–67%
속도	휴머노이드는 사람 속도의 30–50%(UBTech), 일부 약 20%(Reuters); 프런티어 에이전트는 준정적 실기 과제당 약 280초; LLM 제어 루프는 필요한 약 83 Hz 대비 0.2–0.4 Hz(Anthropic)
비용	Walker S2 평균 약 RMB 600k; 회수 7–12년 estimate
안전 인증	ISO 10218-1/-2:2025는 기능 안전을 더 명시적으로 만들고 협동 요구를 흡수하지만 AI에 대해서는 아무 말도 없다(The Robot Report). ISO/IEC TS 22440(기능 안전과 AI)은 2026년 11월에도 여전히 2차 위원회 초안이다(ISS, ISO Open Data). IEC 61508-3 지침은 SIL 1 이상에서 AI를 권하지 않는다(Analog Devices). EU Machinery Regulation에 따르면 머신러닝이 있는 안전 기능은 2027년 1월 20일부터 인증기관 평가가 필요하다(EUR-Lex)
손과 도구	손재주 손은 손가락당 약 30 N(≈3 kg)으로 사람 손의 약 10 kg에 못 미친다(Gasgoo). 도구 교체는 여전히 수동이다
통합	센서 오판, 프로토콜 불일치, 현장 MES 연결; 개입률은 거의 공개되지 않음

안전 함의는 직접적이다. 표준은 학습된 정책을 금지하지도 인정하지도 않는다. ISO 10218-1/-2:2025는 AI에 대해 아무 말도 없다. ISO/IEC TS 22440(기능 안전과 AI)은 여전히 위원회 초안이며 2027년 말이나 2028년 전에는 나오기 어렵다 estimate. ISO/IEC TR 5469:2024는 지침일 뿐이며, 비AI 폴백, 감독 제약, 중복 투표를 권장한다. IEC 61508 3판은 2027년 초 예상된다(ISO Open Data, BSI, 61508 Association) verified. 머신러닝 인지를 안전 기능으로 인증하는 수용된 경로를 제공하는 공개 표준은 아직 없

다. 지역 규칙도 움직인다. 중국은 GB 11291을 개정 중이다(약 2028–2029년 시행)(revision plan 20262543). UL 1740은 2025년 8월부터 ISO 10218-2를 미국 NRTL 심사에 도입했다. EU는 2027년 1월 20일부터 머신러닝이 있는 안전 기능의 인증기관 평가를 요구한다. ISO 25785-1이 초안인 동안 펜스리스 공장 사용 인증을 받은 휴머노이드는 없다(i-SCOOP). 방어 가능한 아키텍처: 머신러닝은 공정 결정("어떤 부품", "이 기판은 양품인가")만 내리고, 인증된 결정론적 안전 계층(보통 PL d / Category 3)과 안전 PLC가 속도·힘·정지를 결정한다. PL d 또는 SIL 2 인증서에 머신러닝 부품이 명시되기 전까지 "펜스리스 안전용 카메라 AI" 주장은 미검증으로 취급하라.

FIG 32 **표준·규제 타임라인** ***안전 기능 내 AI의 인증 경로는 2027–2029년 전에는 확정되지 않는다.***



출처: ISO, IEC, UL, EUR-Lex, SAC 및 국가 규제 데이터베이스.

기회가 있는 계층: 모델은 거인, 엣지 박스는 칩 제조사, 소기업은 통합과 센싱 하드웨어

모델 계층은 자본과 데이터가 필요하며 거인에게 속한다(AgiBot은 GO-2를 "수만 시간"으로 학습했다(iFeng Tech). 가장 강한 모델은 폐쇄형이거나 클라우드 전용이다). 엣지 박스 계층은 칩 제조사에 속한다. Jetson Thor가 사실상 표준이고, 중국 본토에서는 D-Robotics의 RDK 보드(D-Robotics)와 Rockchip 보조 프로세서(CSDN)가 경쟁하지만 대부분 VLA 결과를 공개하지 않는다.

소기업에는 두 개의 진입점이 있다. 하나는 신뢰성을 직접 사 주는 센싱 하드웨어다. 터치는 ContactDP의 삽입 성공률을 58.7%에서 97.8%로 올리고, 힘/토크 센서와 손목 카메라도 비슷한 일을 한다. 데이터 수집 키트도 팔린다(10jqka). 다만 손, 센서, 암 모두 가격이 내리고 있다. 다른 하나는 암 제조사와 모델 회사 사이의 통합 계층으로, 각 스테이션의 과제 데이터, 안전 범위, MES 연결을 소유하는 것이다. 암 제조사는 오늘 이를 제품으로 팔지 않는다. 경쟁자는 네 그룹으로 나뉜다. 암 선두(FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot 등), 비전+암 기업(Mech-Mind, Hikrobot, SmartMore), 임바디드 AI 기업(AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA), 자동차 칩 진입자다.

구매자 수용: 분모 없는 숫자는 없다

벤더가 개입률, 고장 간격, 장기 성공 횟수를 공개하지 않으므로, 이것들은 계약의 검수 조건에 들어가야 한다. 통계적으로 연속 생산 사이클 3,000회 무고장은 95% 신뢰도에서 실패율 최대 0.1%를 증명한다. 3,000회 중 10회 실패는 ≤0.57%만 증명한다. 무고장으로 ≤1%, ≤0.5%, ≤0.1%, ≤0.01%를 증명하려면 299, 598, 2,995, 29,956 사이클이 필요하다 estimate, exact binomial. 그 기준에서 Neuromeka의 322/324는 양측 95% 구간이 97.8–99.9%로 ≥99%를 증명하지 못한다. AgiBot의 64,828개 과제 중 약 6회 실패는 99.98–100%를 시사하지만, "실패"가 구매자가 정의하는 방식으로 정의될 때만이다 estimate. 다음은 내가 어떤 로봇 암·임

바디드 셀에도 요구할 검수 조건이며, 임계값은 BMW에서의 Figure 자체 목표와 Neuromeka·AgiBot 사례를 참고한 판단이지 표준이 아니다 estimate .

항목	요구	권장 합격 임계값
정의	사이클, 성공, 실패(부분 성공, 재그랩, 낙하, 품질 유출), 개입(터치, 원격 텔레오퍼레이션, 리셋, 비상정지)의 서명된 정의와 자율성 계산 방식	정의가 모든 지표보다 먼저; 공식 없는 백분율은 없다
개입·텔레오퍼레이션	로봇시당, 유형·지속시간별; 개입 간 평균 시간; 가동 시간 대비 텔레오퍼레이션 분; 로봇당 작업자 수	구조화 스테이션은 로봇시당 개입 ≤0.5회 (≥2시간 간격); 검수 중 텔레오퍼레이션 0
장기 성공	≥3교대, ≥2개 부품 번호, 정상 조명·공급 변동에서 연속 생산 사이클 ≥3,000회; 실패 분류; 정확 하한	라인 정지 인라인 스테이션 하한 ≥99.5%, 오프라인 ≥99.0%; 품질 유출은 별도 보고
사이클 타임	복구 포함 부품당 P50/P90/P99, 같은 스테이션의 30회 이상 실측 사람 사이클 대비	사람 P50의 ≤1.15배, 또는 라인 택트 내
가동률·고장	가동 시간, 다운타임 로그, 부품별 고장, 부품 교체 시간	≥2주간 가동률 ≥95%; 부품 현장 보유 시 MTTR ≤20분
복구	스크립트 고장 주입: 낙하, 가림, 조명 변화, 컨베이어 정지, MES·네트워크 손실, 정전, 비상정지	자율 복구 ≥90%, 2분 내 생산 복귀
전환	신규 부품 번호·지그당 시간, 시연 횟수, 작업자 숙련도, 데이터 수집 비용	구조화 스테이션: 신규 부품 번호당 ≤1교대, 데모 ≤100회
안전	암·셀은 ISO 10218-1/-2:2025, 셀 위험 평가, 현지 규칙(중국 본토 GB 11291; 대만 CNS 14490 + 서면 평가; 미국 ANSI/A3 R15.06-2025 및 UL 1740)	머신러닝으로 구현된 안전 기능 없음; 인증 안전 계층(보통 PL d / Cat 3)이 속도·힘 제한; ISO 25785-1 공표·평가 전까지 펜스리스 보행 로봇 없음
컴퓨트·오프라인 동작	추론 하드웨어, P99 지연, 네트워크 없는 동작, 모델 버전 관리, 업데이트마다 재검증	택트 중요 추론은 클라우드에 의존하지 않음; 모델·양자화·엔진 변경 시 전면 재검수
증거	검수 기간 전체의 무편집 영상과 원본 로그; 고객 품질 팀 또는 제3자 입회	하이라이트 릴이나 벤더 요약 금지
상업 조건	검수 지표에 연동된 지불 마일스톤; 로봇시 과금 시 가용성 위약; 데이터 소유권과 원격 접속 규칙	3,000사이클 검수와 ≥4주 안정 가동 후에만 최종 지불·보류 해제
경제성	순 대체 인력 = 속도비 × 가동률 × 성공 − 개입 시간 − 텔레오퍼레이션 비중	주장하는 회수를 뒷받침하지 못하는 순 대체는 거부

출처: (Figure, Neuromeka, Yicai Global, i-SCOOP); 임계값은 저자 제안.

07 미해결 질문
공개 기록으로 답할 수 없는 것, 그리고 파일럿이 해결하는 법

공개 출처로 답할 수 없는 질문은 다섯 그룹으로 나뉜다. AI 재판정이 실제로 얼마에 팔리고 구매자가 얼마를 낼 의향이 있는지; AOI 제조사가 AI를 어떻게 번들하고 구형 장비를 어떻게 업그레이드하는지; 설치된 AOI가 얼마나 오래됐고 얼마나 이미 AI를 돌리는지; 라인당 재판정 인력과 공수가 실제로 얼마인지; 파일럿이 어떻게 전환되고 구매자가 어떻게 지불하는지. 그 위에 최신 모델, 양자화 빌드, 로봇 암이 특정 공장 조건에서 실제로 어떻게 동작하는지가 얹힌다. 각각은 명확한 임계값을 가진 테스트로 바꿀 수 있다. 아래 임계값은 이 보고서의 증거에서 내가 도출한 결정 규칙이다 estimate. 첫 공장 데이터가 들어오면 다시 계산하라.

미해결 질문	해결 방법	좋은 결과의 모습
AI 재판정의 가격과 지불 의향	여러 서드파티 벤더에 표준 견적 요청(RFQ)을 보낸다(예: 구형 2D와 신형 3D AOI가 섞인 4개 라인, 2교대, 교대당 기판 약 2,000장, 결함 클래스 12개, 온프레미스 전용, MES 또는 IPC-CFX 통합); EMS 공장 20–30곳에 4문항 가격 민감도 조사	4개 라인 기준 일회성 RMB 40–120k 또는 연 20–40k 범위의 견적이 대부분; 공장의 절반 이상이 12개월 내 회수로 라인당 RMB 10–30k 수용
AOI 제조사의 AI 번들과 구형 장비 업그레이드	AOI 제조사·유통사 약 12곳에 짧은 구조화 인터뷰	2025–26년 모델에서 AI 재판정이 표준·옵션·부재인지; 2020년 이전 장비 업그레이드 가능 여부와 가격
설치 AOI의 연식과 AI 가동 비중	규모·지역별로 층화한 EMS 공장 약 30곳 현장 조사, NDA 하에 장비 목록 검토	AOI 중위 연식 ≥5년이고 ≥60%가 AI 없음이면 개조 풀이 실재
재판정 인력과 공수	2곳 이상 공장에서 인터뷰와 2–4주간 AOI 알람·재판정·수리·교대 로그, 작업 샘플링	라인당 교대당 전담 재판정자 ≥1명, 라인당 교대당 재판정 공수 ≥3시간, 라인당 최소 0.5 FTE를 공장이 재배치할 의향
AI 재판정 자체의 성능	실제 라인의 골든 세트 + 섀도 모드(3장): 0주차 데이터 합의, 1–2주차 기준선·라벨링, 2–4주차 섀도 가동, 4–8주차 선택적 보조 가동	오탐 또는 재판정 공수 ≥50–60% 절감; 클래스별 유출 경계가 사람 기준선보다 나쁘지 않음; 치명 클래스 미검출 0; p95 지연 ≤½ 택트; 사람으로의 자동 페일오픈
프런티어 모델의 교사 라벨 품질	같은 골든 세트에 후보 2–3개, 각 3개 프롬프트 문구, 좋은 참조 이미지 1장, 단일 부품 크롭 ≥42–48 px	결함 ≥299개에서 미검출 0이고 사람 수준 일치인 클래스만 자동 채택; 아니면 사전 라벨·트리아지만
양자화 빌드	정밀도만 변경; 가중치 해시, 템플릿, 전처리, 디코딩, 엔진, 배치 크기 고정; 프리픽스 캐시 비활성화·리셋; 참조 빌드가 벤더 결과를 재현하는지 확인; 골든 세트와 드리프트 슬라이스에서 쌍 비교	치명 클래스 신규 미검출 없음; 결함 1,000개 중 신규 미검출 ≤1; 해소율 하락 ≤2점; 3회 실행 간 ≥99.5% 동일 결정; 아니면 FP8/INT8로 폴백
라인 사이드 하드웨어, 청정도, ESD	각 후보 박스 ≥2대를 기대 알람률의 2배로 스트레스 테스트; 실제 캐비닛 공기 흐름에서 72시간 소크; 고객 코디네이터의 ESD 확인; ISO 7 이상 고객용 ISO 14644-14 테스트	p95 ≤½ 택트; 72시간 무크래시·무쓰로틀링; ESD 서명; 적정 등급의 입자 보고서
지불 조건, 파일럿 규칙, 전환	조사 공장의 최근 검사 장비 계약 지불 조항; 파일럿 규칙·검수 기준·자격 심사 기간 인터뷰	약 30/60/10, 검수 후 60일 내 현금; 파일럿 ≤8–12주에 서명 책임자 지정; 3건 중 ≥1건 전환; 1년차에 ≥5–10개 라인 프레임워크
데이터, IT, 책임	법률 고문과 인터뷰·계약 검토	로컬 박스가 연결·데이터 접근 가능; 책임은 소프트웨어 결함으로 제한
해외 진출 전 지역 규정 준수	대만: 고객의 Cyber Security Management Act상 서면 지위, E187 자가평가; 미국: 셀 분류에 대한 관세 판정, NRTL 견적, ITAR 확인; 동남아시아: BOI 로컬 연계와 말레이시아 공제 적격성, 고객의 모기업; 중국 본토: 번들 PC·전원의 CCC 범위	고객이 통제 기관이 아니고, 관세·인증 포함 회수가 ≤2년 유지

미해결 질문	해결 방법	좋은 결과의 모습
로봇·임바디드 셀의 검수	6장의 12항목 검수표: 서명된 정의, ≥3교대·≥2개 부품 번호의 연속 생산 사이클 ≥3,000회, 스크립트 고장 주입, 실측 사람 사이클 ≥30회, 입회	인라인 하한 ≥99.5%(오프라인 ≥99.0%); 사이클 ≤ 사람의 1.15배; 시간당 개입 ≤0.5회·텔레오퍼레이션 0; 가동률 ≥95%; MTTR ≤20분; 복구 ≥90%
중국 엣지 칩의 ACT·VLA 정책	4주: 포팅과 마이크로 벤치마크(추론 ≥1,000회, 지연 분포, 전력, 60분 소크); 측정 지연을 주입한 폐루프 시뮬레이션; GPU 대비 실기 A/B; 8–24시간 내구(≥3,000사이클)	ACT 청크 p95 ≤100 ms, VLA ≤300 ms; 실기 성공률 GPU 대비 3점 이내, 사이클 ≤사람의 1.2배; 아니면 Jetson·GPU에 정책 유지
협동로봇·AMR·통합 가격	여러 협동로봇 유통사의 모델별 견적, 테스트 지그 로딩(스테이션 1개, 2교대)에 대한 인테그레이터 2–3곳의 풀셀 견적	풀셀 ≤RMB 200k, 선전 2교대 회수 ≤1.5년
대만·동남아시아·미국의 SMT 라인 수	IPC(구 New Venture Research) 데이터 또는 현지 장비 유통사 설치 기반	현지 응용 엔지니어링을 정당화할 만한 지역별 대상 라인

이 중 네 가지가 이 보고서에서 가장 민감한 수치를 결정한다. 구매자가 낼 가격, AI 재판정이 실제로 푸는 인력, AI 재판정이 정말 사람보다 더 놓치지 않는지, 현금이 얼마나 빨리 돌아오는지다. 구매를 고려하는 공장이든 제품을 고려하는 공급자든, 큰 지출 전에 이 네 가지를 해결할 가치가 있다.

양쪽 모두에 작동하는 파일럿 순서

공장에게도, 신뢰할 만한 레퍼런스를 원하는 벤더에게도 순서는 같다.

1. **0–2주차: 데이터와 정의 합의.** 데이터 계약에 서명하고, 결함 클래스와 무엇이 유출인지 정의하고, 기준선을 기록한다. 기판당 오탐, 호출당 재판정 초, 다운스트림에서 발견된 유출.
2. **1–4주차: 골든 세트 구축.** AOI 아카이브와 보관 NG 기판에서, 블라인드 라벨링과 판정으로(3장).
3. **2–4주차(또는 그 이상): 섀도 모드.** 사람이 이전처럼 결정하는 동안 모델이 결정을 기록; 클래스별 비교.
4. **4–8주차: 무작위 교대에서 보조 가동.** 게이트를 통과한 클래스만 모델이 해소하고 나머지는 모두 사람에게.
5. **확장 전 서면 검수:** 클래스별 결과, 지연, 페일 오픈 동작, 모델·라벨 소유권. 라인별로 확장하고, 신제품·모델 버전·양자화·엔진 변경 시마다 재검증.

이 보고서의 한계

이것은 데스크 리서치다. 이번 판에서는 공장 대상 1차 인터뷰를 하지 않았고, 많은 라인 데이터가 벤더 출처이며 독립 감사를 받지 않았다. 여러 정부 웹사이트, 입찰 데이터베이스, 채용 플랫폼에 접근할 수 없어서 선전의 2026년 업그레이드 보조금 요율, Yangtze River Delta 임금 지침, AI 재판정 거래 가격, 라인당 재판정자 수, 파일럿 전환율은 추정 또는 유추이며, 각각 위에 미해결 질문으로 나타난다. 환율은 2026년 11월 9일 People's Bank of China 기준환율($1 = RMB 6.7188)과 같은 날 SMBC 고시환율($1 = NT$32.12)이며, 베트남 동은 2026년 5월 데이터를 쓴다. 모든 임계값은 인용된 증거에서 도출한 결정 규칙이지 표준이 아니다.

저자 소개

Tan Shuai

Tan Shuai는 기술자이자 창업자로, 반도체·하드웨어·시스템 소프트웨어에서 20년 이상의 경력과 등록 특허 20건을 보유했다. Silicon Valley, New York, Chicago, 대만, 광둥-홍콩-마카오 대만구(선전·홍콩), 베트남, 말레이시아에서 일했으며, 첨단 제조와 공급망에 오래 집중해 왔다. 현재는 AI 연구자이자 개발자로 일하며, AI가 공장 현장에서 실제로 어떻게 쓰이는지에 집중하고 있다.

제조 분야 AI 종사자와의 대화를 환영한다. 팀 구축, 파트너 발굴, 의견 교환 모두: i@tanshuai.com · tanshuai.com

주석과 면책 조항

이 보고서는 2026년 11월 11일까지 공개된 문서에 기반한 독립 데스크 리서치다. 정보 제공 목적으로만 제공되며 투자·법률·세무·엔지니어링 자문이 아니다. 제품명, 가격, 모델 출시, 수출 분류, 관세, 규정은 빠르게 변하므로, 의존하는 모든 것은 1차 출처와 대조해 확인하라. vendor claim 또는 estimate 로 표시된 수치는 그렇게 읽어야 한다. 회사·제품명은 소유자에게 있으며 식별 목적으로만 사용된다. 어떤 회사도 이 보고서를 검토하거나 후원하지 않았다.

© 2026 Tan Shuai. 이 보고서는 출처 표시와 함께 변경 없이 공유할 수 있다.

전자 제조 · AI 검사 × 로보틱스

전자 공장의 AI 비전 검사와 로봇 암: 2026 평가

모든 핵심 수치는 출처와 신뢰도 태그를 달고 있다: verified, single source, vendor claim, estimate. 공개 기록으로 답할 수 없는 것은 7장에 모았고, 각 질문을 파일럿이 어떻게 해결하는지도 함께 담았다.

Tan Shuai
tanshuai.comi@tanshuai.com

공개판 · 데이터 기준 2026년 11월 11일

