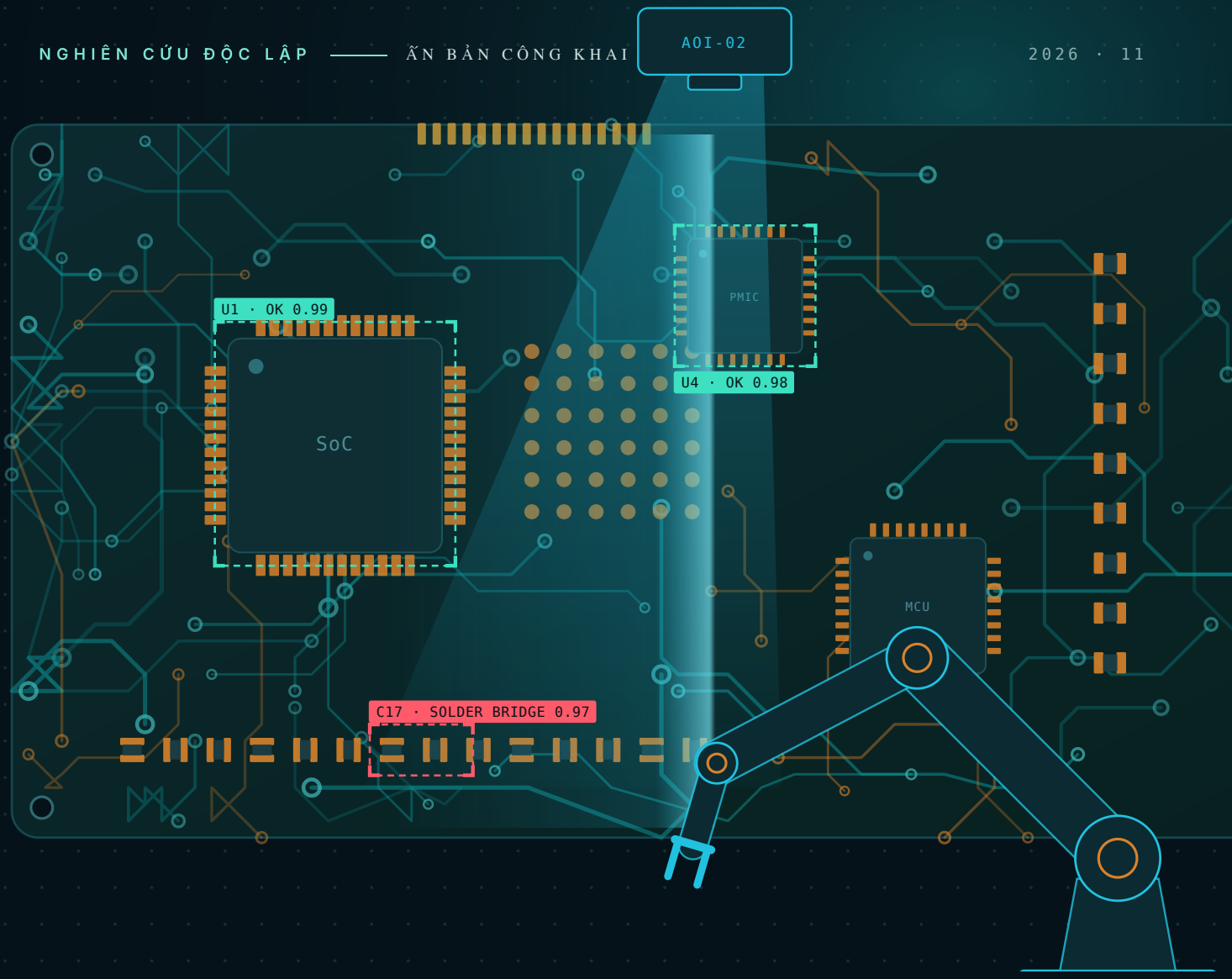


SẢN XUẤT ĐIỆN TỬ · KIỂM TRA AI × ROBOT

Kiểm tra thị giác bằng AI và cánh tay robot trong các nhà máy điện tử

Đánh giá 2026

Các cấu hình chính thống · Tính khả thi theo khu vực · Triển khai tại chỗ · Mô hình thầy–trò
Lợi tức đầu tư · Cạnh tranh · Điều chỉ có thí điểm mới trả lời được

PHÁT HIỆN CHÍNH

Rà soát AI đáng làm, trong phạm vi hẹp. Cánh tay robot hiện thân có thể chờ.

Tan Shuai
tanshuai.comi@tanshuai.com

Dữ liệu tính đến ngày 11 tháng 11 năm 2026
Trung Quốc đại lục · Đài Loan · Đông Nam Á · Hoa Kỳ

MỤC LỤC

Mục lục



◆ **TÓM TẮT ĐIỀU HÀNH** — Đáng làm, nhưng trong phạm vi hẹp, ngay bây giờ và cùng đối tác — 5

01 **Dây chuyền ngày nay** — Các nhà máy điện tử đang vận hành gì ngày nay, và điều gì khả thi về mặt kỹ thuật — 9

02 **Tại chỗ** — Chạy AI bên trong nhà máy: phòng sạch, cách ly mạng, phần cứng và lượng tử hóa — 17

03 **Thầy và trò** — Thầy và trò: mô hình frontier có thể gán nhãn trước dữ liệu kiểm tra; với cánh tay robot người thầy thực sự vẫn là con người — 26

04 **Chi phí và hoàn vốn** — Chi phí và hoàn vốn: rà soát AI hoàn vốn trong vài tháng, phục vụ máy trong vài năm, cánh tay gắp bỏ thì không bao giờ — 35

05 **Thị trường và cạnh tranh** — Thị trường, cạnh tranh và cách người mua thanh toán — 43

06 **AI hiện thân** — AI hiện thân: trước tiên nó thay đổi cách robot được dạy, sau đó là cách chúng di chuyển — 52

07 **Các câu hỏi bỏ ngỏ** — Những gì hồ sơ công khai không thể trả lời, và cách một thí điểm giải quyết nó — 58

◆ **Về tác giả** — Tan Shuai — 61

Các nhãn mức độ tin cậy trong văn bản

- **verified** hai nguồn độc lập trở lên, hoặc hai tài liệu gốc độc lập, đồng thuận
- **single source** chỉ một nguồn duy nhất, bao gồm cả tài liệu gốc
- **vendor claim** do người bán sản phẩm hoặc dịch vụ công bố
- **estimate** phép tính hoặc suy luận của tác giả từ dữ liệu được trích dẫn

Cách đọc báo cáo này

Các con số then chốt mang nhãn mức độ tin cậy. **verified** nghĩa là hai nguồn độc lập trở lên đồng thuận, hoặc hai tài liệu gốc độc lập đồng thuận; **single source** nghĩa là chỉ một nguồn (bao gồm cả tài liệu gốc); **vendor claim** nghĩa là do người bán sản phẩm hoặc dịch vụ công bố, kể cả khi được lặp lại rộng rãi; **estimate** nghĩa là phép tính hoặc suy luận của riêng tôi từ dữ liệu được trích dẫn, với các đầu vào đã dẫn nguồn. “Primary” mô tả loại nguồn (báo cáo thường niên, bản cáo bạch, quy định, tiêu chuẩn, thống kê chính thức); nó không thay đổi lớp mức độ tin cậy. Khi các nguồn mâu thuẫn, các con số được trình bày cạnh nhau và không lấy trung bình. Quy đổi tiền tệ dùng tỷ giá trung tâm của Ngân hàng Nhân dân Trung Quốc ngày 9 tháng 11 năm 2026, $1 = 6.7188 RMB (Xinhua), và cho Đài Loan là tỷ giá cố định của SMBC cùng ngày, $1 = 32.12 NT$ (SMBC). Dữ liệu lao động và trợ cấp ở đại lục dùng Đồng bằng Châu Giang, nơi sản xuất điện tử tập trung nhất (dữ liệu lương chính thức từ Thâm Quyến), với Đồng bằng Dương Tử để so sánh. Các tuyên bố từ phỏng vấn ngành được coi là tuyên bố cần kiểm chứng, không bao giờ là sự thật. Các câu hỏi mà hồ sơ công khai không thể trả lời được tập hợp ở Chương 7.

HÌNH MINH HỌA

Hình minh họa



01	AI phù hợp ở đâu trên sàn nhà máy hôm nay	7
02	Dây chuyền SMT sản lượng lớn: người ngồi sau AOI	10
03	Khoảng 97–99% cảnh báo AOI là báo sai	11
04	Camera đa năng không phải là AOI	13
05	Triển khai cách ly mạng	17
06	Khu sạch và ESD quyết định chỗ đặt máy tính	18
07	Giới hạn tốc độ theo mục đích	20
08	Tính toán on-prem: bộ nhớ, giá và thứ mua hợp pháp ở Trung Quốc đại lục	21
09	Một máy trạm lớn không thành trợ lý dùng chung	22
10	Lượng tử hóa: 8-bit an toàn, 4-bit tùy công thức, 3-bit trở xuống loại	23
11	Ứng viên thầy: điểm chung tăng gấp đôi trong một thế hệ	27
12	"Có lỗi không?" vẫn khó với mô hình chung	28
13	Hai pipeline thầy–trò	29
14	Bộ chuẩn vàng phải lớn cỡ nào	30
15	Độ trễ chính sách biên	32
16	Robot trò: gần cấp sản xuất chỉ khi trạm có cấu trúc	33
17	Chi phí năm của một công nhân kiểm hoặc vận chuyển	36
18	Bậc thang giá	37
19	Hoàn vốn theo vùng	39
20	Vì sao nhặt tấm lỗi không hoàn vốn	40
21	Thuế Mỹ với robot, AOI và máy tính biên	41
22	Quy mô thị trường chênh nhau hơn mười lần	44
23	Từ dưới lên: 40,000 dây chuyền → 16,000 phục vụ được → ≈RMB 320M một lần	45
24	Cửa sổ cải tạo khoảng năm năm	46
25	Nêm tái kiểm bị bao vây ba mặt	46
26	Biên lợi nhuận nằm ở phần cứng sản phẩm hóa	48
27	Người mua trả thế nào	49
28	Số ngày phải thu tồn đọng	50
29	Ngăn xếp AI embodied	52
30	Điều triển khai robot trong nhà máy thực sự công bố	53
31	Nhân lực thay thế ròng mỗi robot	54
32	Dòng thời gian tiêu chuẩn và quy định	55

TÓM TẮT ĐIỀU HÀNH

Đáng làm, nhưng trong phạm vi hẹp, ngay bây giờ và cùng đối tác

PHÁT HIỆN CHÍNH

Kết luận: đối với các nhà máy điện tử, điều đã chín muồi trong năm 2026 và điều các nhà máy sẽ trả tiền là rà soát AI và kiểm tra ngách trên ảnh AOI, cộng với phục vụ máy bằng robot được lập trình, chứ không phải "một mô hình lớn nhìn các tấm mạch cộng với robot hiện thân gắp chúng".

0.46–0.65 năm

Thời gian hoàn vốn của rà soát AI tại Mỹ, Đài Loan và Thâm Quyến, ở mức 50 nghìn RMB mỗi dây chuyền

≥95% · <20 PPM

Độ bao phủ và tỷ lệ lọt lưới của hệ thống rà soát AI nội bộ của một EMS hàng đầu: chuẩn mực để bán cho EMS cấp 1

100–700 triệu RMB

Thị trường nâng cấp do bên thứ ba thực hiện có thể phục vụ tại Trung Quốc đại lục, một lần, trong khoảng thời gian chừng năm năm

0 điểm

Điểm benchmark khuyết tật công nghiệp đã công bố của các mẫu flagship tháng 11/2026: là thầy, không phải người kiểm tra

1.1–1.5 năm

Thời gian hoàn vốn của một cụm cobot phục vụ máy chạy hai ca tại Mỹ, Đài Loan và Thâm Quyến

0.003–0.14 FTE

Lao động mà một cánh tay robot gắp bỏ linh kiện lỗi tiết kiệm được mỗi ca trên một dây chuyền SMT: nó không bao giờ hoàn vốn

Kiểm tra
ROI

Học sâu có giám sát áp dụng lên những ảnh mà một AOI sẵn có đã tạo ra cắt giảm các cảnh báo giả hoặc số giờ rà soát từ 60–90% (vendor claims). DBG, một công ty EMS lớn, trong báo cáo thường niên cho biết hệ thống nội bộ của họ bao phủ hơn 95% các linh kiện bị gắn cờ với tỷ lệ lọt lưới dưới 20 PPM và giảm hơn 85% khối lượng rà soát thủ công; đó là chuẩn mực mà một sản phẩm bên thứ ba phải vượt qua để bán cho EMS cấp 1. Ở mức 50,000 RMB mỗi dây chuyền, rà soát AI hoàn vốn trong vòng một năm tại Trung Quốc đại lục, Đài Loan và Mỹ (Thâm Quyến 0.65 năm, Đài Loan 0.53, Mỹ 0.46), các cụm cobot phục vụ máy trong 1.1–1.5 năm và Đông Nam Á trong 2–6 năm, trong khi một cánh tay robot thêm vào dây chuyền SMT chỉ để gắp ra các tấm mạch lỗi thì không bao giờ hoàn vốn ở bất kỳ đâu, bởi vì trên 90% tỷ lệ đạt ngay lần đầu, việc gắp chỉ chiếm 0.003–0.14 người mỗi ca estimate.

Mô hình
ROBOT

Các mô hình lớn đa phương thức vẫn chưa thể làm người kiểm tra chính. Tính đến ngày 11 tháng 11 năm 2026, không có mẫu flagship nào (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 và các mẫu khác) có điểm đã công bố trên bất kỳ benchmark khuyết tật công nghiệp hoặc PCB nào, và trong thử thách PCBA của ACM MM 2026, năm hệ thống dẫn đầu, được huấn luyện trên 6,600 ảnh thực từ dây chuyền, chỉ đạt 77–82 điểm ở hạng mục "có khuyết tật hay không". Các mô hình frontier có thể gán nhãn trước và phân loại sơ bộ. Đối với cánh tay robot, người thầy thực sự vẫn là trình diễn của con người, học tăng cường trên robot thật và mô phỏng: học trò đạt 90–99.4% ở các trạm cố định, có cấu trúc, nhưng chỉ 30–80% trên băng chuyền đang chuyển động, linh kiện chưa từng thấy

và sản phẩm mới; một tác tử frontier hoàn thành 104 trên 108 nhiệm vụ robot bán tĩnh trên robot thật, ở mức 280 giây mỗi nhiệm vụ.

Tại chỗ
TRIỂN KHAI

Triển khai cách ly mạng trong các nhà máy sạch, có kiểm soát là hoàn toàn khả thi. Một xưởng SMT thông thường thường là khu vực bảo vệ ESD chứ không phải phòng sạch phân cấp: một hộp Jetson không quạt hoặc hộp NPU nội địa đặt tại dây chuyền và các máy chủ GPU đặt trong phòng máy chủ. Theo chính phân loại xuất khẩu của NVIDIA, Jetson Thor và DGX Spark có thể bán cho một nhà máy đại lục thông thường mà không cần giấy phép Mỹ; DGX Station thì không. Độ chính xác mặc định là 8-bit; 4-bit nhận biết lượng tử hóa của nhà cung cấp dùng được kèm điều kiện; 3-bit trở xuống là loại.

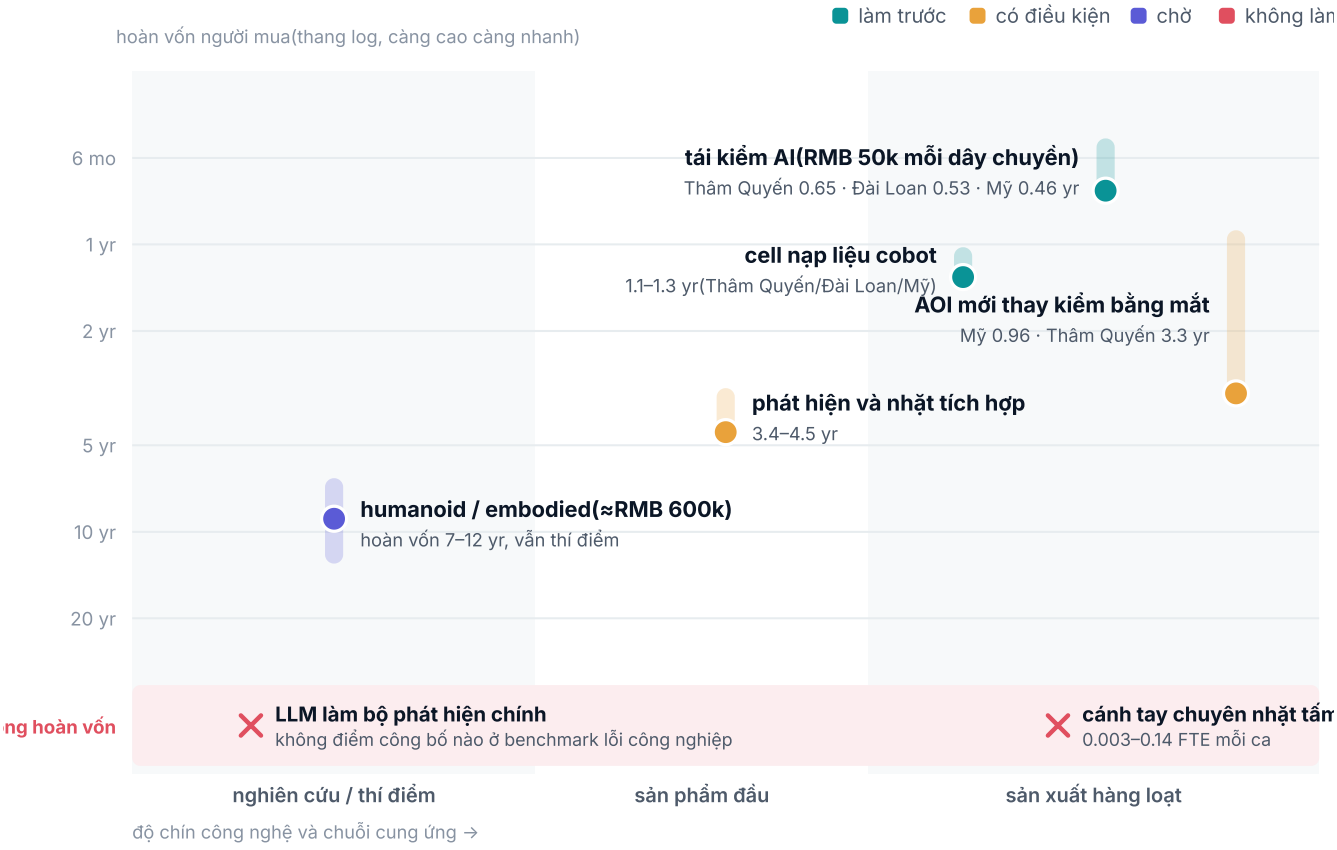
Thị trường
CẠNH TRANH

Thị trường là thật nhưng hữu hạn. Trung Quốc đại lục có khoảng 35,000–46,000 dây chuyền SMT và khoảng 26,000–66,000 AOI nội tuyến chưa chạy AI; thị trường có thể phục vụ cho nâng cấp của bên thứ ba là khoảng 100–700 triệu RMB một lần, trong một khoảng thời gian chừng năm năm, và các hãng sản xuất AOI đã bắt đầu tích hợp rà soát vào các máy mới estimate. Cạnh tranh vốn đã mạnh: mảnh đất rà soát bị bao vây bởi các hãng AOI, các đội EMS nội bộ và các nhà cung cấp nền tảng, còn lợi nhuận nằm ở phần cứng được sản phẩm hóa.

Câu hỏi	Đánh giá	Các con số then chốt (nguồn và mức độ tin cậy ở các chương)
① **Các nhà máy điện tử chính thống đang vận hành gì ngày nay?**	Một dây chuyền SMT khối lượng lớn là "3D SPI + AOI trước và sau reflow + X-ray lấy mẫu + ICT/FCT + sửa chữa và kiểm tra thị giác", với rà soát thủ công phía sau mỗi AOI. AI tham gia vào rà soát theo ba cách: do hãng AOI tích hợp sẵn, do EMS hàng đầu tự xây nội bộ, hoặc do bên thứ ba thêm vào. Lắp ráp và xử lý dùng cánh tay, SCARA và AMR với nhận thức được học và chuyển động được lập trình; AI hiện thân vẫn đang ở giai đoạn thí điểm	Takt ≈13–22 giây mỗi tấm mạch; một ODM đã dự chi 30.6 triệu RMB cho AOI cho 18 dây chuyền, ≈3–5 AOI mỗi dây chuyền; ≈97–99% cảnh báo AOI là cảnh báo giả
② **Công nghệ đi được bao xa, tiền có mua được "không người" không, và mỗi khu vực dùng được gì?**	Kiểm tra đạt cấp công nghiệp ở dạng "bộ phát hiện chuyên dụng + rà soát AI + vài người"; ngay cả tiền không giới hạn cũng không đạt tới không người. Đại lục chỉ dùng các mô hình nội địa hoặc trọng số mở nội địa; các nhà máy tư nhân ở Đài Loan dùng được gần như mọi thứ, nhưng các nhà máy wafer và hạ tầng trọng yếu loại trừ thương hiệu đại lục trên thực tế; Đông Nam Á ít hạn chế nhất; Mỹ có đầy đủ giải pháp cộng thêm thuế quan và chứng nhận	Rà soát AI cắt giảm cảnh báo giả hoặc giờ rà soát 60–90% vendor claims ; một camera 4K trên toàn tấm mạch ≈78 µm mỗi pixel so với 8–20 µm của một AOI; cánh tay sản xuất tại đại lục chịu ≈40% thuế Mỹ tổng hợp, các thiết bị dạng AOI ≈37.5%
③ **Có thể chạy tại chỗ trong phòng sạch và nhà máy cách ly mạng không, trên phần cứng nào, đủ nhanh và đủ chính xác không?**	Có; các ràng buộc là về quy trình, không phải tính toán. Một hộp không quạt ≤100 W tại dây chuyền, máy chủ GPU trong phòng máy chủ; Thor cho các trạm, Mac Studio hoặc thiết bị nội địa cho Hỏi&Đáp của nhà máy; không có DGX Station ở đại lục. Tốc độ ổn khi kích thước mô hình khớp với nhiệm vụ; giữ 8-bit, dùng 4-bit nhận biết lượng tử hóa của nhà cung cấp kèm điều kiện, không bao giờ 3-bit	≈0.5 giây mỗi ảnh cắt cho mô hình thị giác 8B trên Thor estimate ; bộ phát triển Thor 59,104 RMB tại Trung Quốc; DGX Station $123–140k, tổng 139 token/s cho 32 người dùng trên mô hình 743B
④ **Mẫu flagship mới nhất có thể làm thầy qua hàng chục vòng và được chưng cất thành học trò nội bộ không, kể cả cho sản xuất linh hoạt và cánh tay robot?**	Kiểm tra: gán nhãn trước và phân loại sơ bộ; chỉ tự động chấp nhận nhãn của thầy cho các lớp có "≥299 khuyết tật, không lần bỏ sót" trên bộ dữ liệu vàng. Robot: người thầy vẫn là trình diễn của con người, RL trên robot thật và mô phỏng, và học trò chỉ tiệm cận cấp sản xuất ở các trạm có cấu trúc	Không mẫu flagship tháng 11 nào có điểm khuyết tật công nghiệp; học trò 97.8% trên đầu nối đã huấn luyện, 29–67% trên đầu nối chưa thấy, 37–59% trên băng chuyền đang chuyển động; mô hình frontier điều khiển trực tiếp khớp 0–5.5%

Câu hỏi	Đánh giá	Các con số then chốt (nguồn và mức độ tin cậy ở các chương)
⑤ **Chi phí là bao nhiêu, và ROI theo khu vực ra sao?**	Rà soát AI hoàn vốn trong vòng một năm tại Trung Quốc đại lục, Đài Loan và Mỹ; cobot phục vụ máy 1.1–1.5 năm; Đông Nam Á 2–6 năm; cánh tay gắp bỏ trên dây chuyền SMT không bao giờ hoàn vốn	Hộp rà soát 50k RMB mỗi dây chuyền: Thâm Quyến 0.65 năm, Đài Loan 0.53, Mỹ 0.46, Việt Nam 2.2 **estimate**; giá chấp nhận được cho tiện ích bổ sung của bên thứ ba ≈10–30k RMB mỗi dây chuyền một lần hoặc 5–10k RMB mỗi năm
⑥ **Đối thủ đã đủ mạnh chưa, và nên bắt đầu ngay hay chờ?**	Nhu cầu chưa bão hòa, nhưng nguồn cung đông đúc; AI của OEM thu hẹp khoảng thời gian nâng cấp xuống còn khoảng năm năm. Hãy tham gia ngay với một sản phẩm hẹp qua các đối tác kênh phân phối, không đối đầu trực diện với các nhà cung cấp nền tảng; hãy chờ dữ liệu sản xuất 2027–2028 trước khi đặt cược vào AI hiện thân	≈26–66k AOI nội tuyến chưa có AI ở đại lục, thu hẹp 8–12% mỗi năm **estimate**; nhà cung cấp thị giác AI hàng đầu có doanh thu 2025 là 1.086 tỷ RMB và lỗ ròng 991 triệu RMB; doanh thu mảng 3C của công ty niêm yết tương đương gần nhất giảm 79.9% so với cùng kỳ trong tháng 1–tháng 5 năm 2026

FIG 01 **AI phù hợp ở đâu trên sàn nhà máy hôm nay** ***Tái kiểm AI các báo sai của AOI và nạp liệu bằng cobot đã chín muồi và hoàn vốn nhanh; cánh tay nhặt riêng và LLM làm bộ kiểm chính thì không đáng làm.***



Nguồn: Hoàn vốn từ mô hình của tác giả(Chương 4); trục độ chín là nhận định của tác giả

Nên xây dựng gì và không nên xây dựng gì

	Khuyến nghị
Nên xây dựng	Rà soát AI và phân loại khuyết tật trên ảnh AOI, cho các EMS quy mô trung bình với AOI cũ hoặc đa thương hiệu và không có nâng cấp AI từ hãng AOI; kiểm tra những gì AOI SMT bao phủ kém (xuyên lỗ và hàn tay, phủ conformal, cáp, đầu nối và FPC, ngoại quan cuối cùng và ốc vít); các cụm nạp và dỡ bằng cobot nơi mọi linh kiện đều được xử lý (đồ gá kiểm tra, khay, đóng gói), với kiểm tra như một tiện ích bổ sung; một bộ dữ liệu vàng và chế độ chạy song song trên dây chuyền thực làm nền tảng cho mọi tuyên bố
Không nên xây dựng	Một nền tảng tổng quát "AI cộng cánh tay robot"; một LLM đa phương thức làm người kiểm tra chính; một cánh tay robot trên dây chuyền SMT chỉ để gắp ra các tấm mạch NG (một bộ tách NG hoặc bộ đệm làm việc đó); robot hình người và robot-dạng-dịch vụ; mô hình frontier bên trong vòng điều khiển hoặc chức năng an toàn; những đống bộ nhớ GPU tại các trạm để chạy mô hình "kích thước đầy đủ"; bán trực tiếp cho nhà máy wafer, hạ tầng trọng yếu Đài Loan hoặc các nhà máy điện tử quốc phòng Mỹ
Làm việc cùng ai	Các hãng AOI và nhà phân phối (giao diện kênh và ảnh; nhà phân phối mang về 46.68% doanh thu năm 2025 của Sinic-Tek); các hãng cobot (cánh tay đã chứng nhận và chức năng an toàn); các công ty EMS có 10–40 dây chuyền (mở rộng cuốn chiếu và đặt cọc); các nhà tích hợp địa phương ở Thái Lan và nơi khác (để đáp ứng quy tắc liên kết nội địa BOI)

Dây chuyền ngày nay

01 Các nhà máy điện tử đang vận hành gì ngày nay, và điều gì khả thi về mặt kỹ thuật

Dây chuyền khối lượng tiêu chuẩn: vài AOI, X-ray lấy mẫu, rà soát tập trung, 13–22 giây mỗi tấm mạch

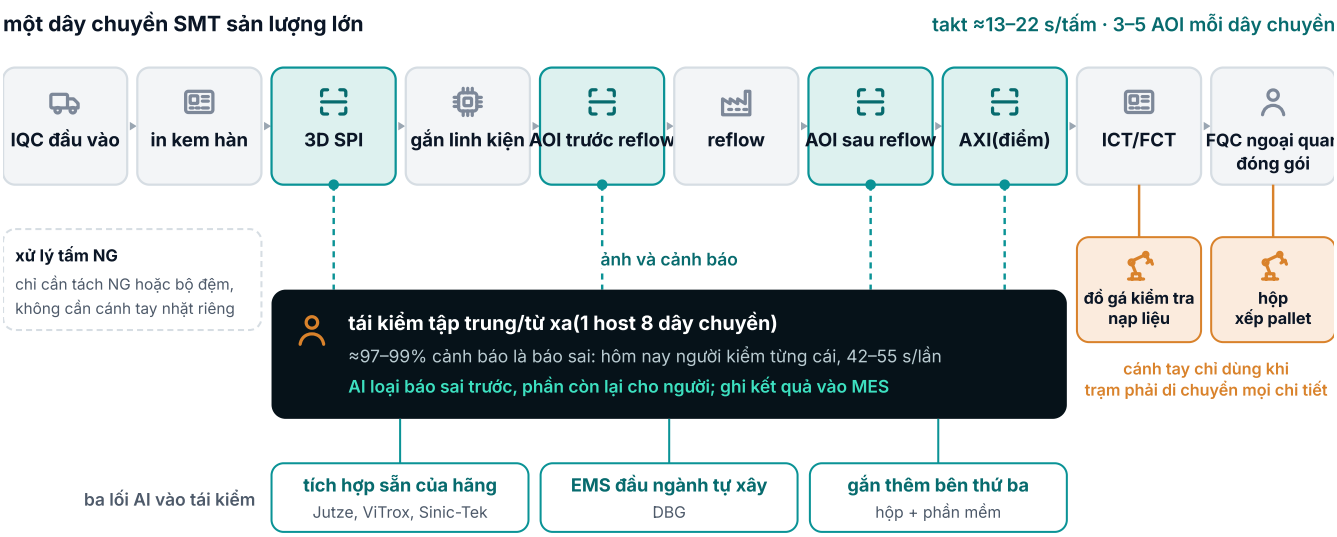
Các hồ sơ gốc từ các ODM niêm yết cho thấy một dây chuyền khối lượng giai đoạn 2025–2026 thực sự trông như thế nào. Bản cáo bạch Hồng Kông của Longcheer (dữ liệu đến ngày 30 tháng 9 năm 2025) liệt kê các khâu kiểm tra chất lượng SMT của họ là "kiểm tra X-ray, kiểm tra kem hàn (SPI), kiểm tra quang học tự động (AOI), kiểm tra chức năng PCBA và kiểm tra thị giác"; luồng quy trình của họ chạy AOI trước reflow, AOI sau reflow, X-ray lấy mẫu, các tấm mạch NG chuyển đến trạm sửa chữa SMT, rồi kiểm tra thị giác cuối cùng. Các nhà máy Huệ Châu, Nam Xương và Việt Nam của họ có 34, 24 và 8 dây chuyền SMT và 67, 60 và 13 dây chuyền lắp ráp cuối cùng (Longcheer H-share prospectus) single source, primary filing. Nhà máy laptop 2023 của Huaqin đã lên kế hoạch 18 dây chuyền SMT với ngân sách AOI 30.6 triệu RMB, khoảng 1.7 triệu RMB mỗi dây chuyền, tương đương ở giá AOI 3D nội địa là ba đến năm AOI mỗi dây chuyền cho kiểm tra trước và sau reflow của mạch hai mặt (Huaqin A-share prospectus) primary filing + estimate. Một dây chuyền khối lượng do đó thường có hai đến bốn điểm rà soát, không phải một.

Các máy móc vốn đã tự động hóa cao: Huaqin báo cáo 96% tự động hóa trên các dây chuyền SMT của mình (Huaqin H-share prospectus) single source. Số người còn lại nằm ở rà soát, sửa chữa, kiểm tra thị giác, kiểm tra và lắp ráp cuối cùng, và không công ty nào công bố họ có bao nhiêu người rà soát mỗi dây chuyền (Chương 7).

Ba ước tính độc lập đồng thuận về takt **13–22 giây mỗi tấm mạch** (một hồ sơ môi trường 2026 cho năm dây chuyền Hà Bắc (Hebei SMT project EIA), các dây chuyền laptop của Huaqin và một ca nghiên cứu iFactory), tức 7–11 giây mỗi mặt cho mạch hai mặt estimate. Phần khó của một dây chuyền nhiều chủng loại không phải là khối lượng mà là giới thiệu sản phẩm mới và chuyển đổi sản xuất, và mỗi lần chuyển đổi lại là một lần khởi động nguội khác cho mô hình AI.

Về phía phần mềm, các trạm rà soát của chính các hãng AOI đã được tập trung hóa: phần mềm trạm sửa chữa của Sinic-Tek cho phép "một máy chủ điều khiển thiết bị của tám dây chuyền SMT" (Sinic-Tek prospectus) single source, primary, 2022, và sau khi Jabil triển khai điều khiển AOI từ xa ViTrox V-ONE vào năm 2026, một kỹ sư có thể quản lý 5–8 AOI từ xa so với 1–2 tại chỗ (AtlasPCB) vendor claim. Một sản phẩm rà soát AI của bên thứ ba cắm thẳng vào chính trạm trung tâm này, thông qua xuất ảnh và kết quả của AOI (IPC-CFX, IPC-HERMES-9852 hoặc định dạng riêng của nhà cung cấp). Bản thân các AOI cũng đang thêm AI rất nhanh: Yamaha cho biết phán đoán AI của họ mất khoảng một giây mỗi lần gọi, hỗ trợ người kiểm tra sau 30 ảnh NG và đạt chất lượng "rất cao" sau 1,000 ảnh (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Dây chuyền SMT sản lượng lớn: người ngồi sau AOI** ***Nhân lực tập trung ở các chỗ tái kiểm sau kiểm quang tự động. Đó là nơi AI đi vào.***



Nguồn: Hồ sơ công khai và tài liệu hãng(Huaqin, DBG, Sinic-Tek và các hãng khác); tác giả tổng hợp

Công đoạn	Sản phẩm và công nghệ chính thống	Nhà cung cấp tiêu biểu	AI đã tiến đến đâu
Sau khi in	3D SPI (ánh sáng cấu trúc, moiré, một số dùng tam giác hóa laser)	Koh Young, Sinic-Tek, TRI	Phán đoán kem hàn bằng AI của OEM trong sản xuất; Jutze nói một số mẫu là "hoàn toàn AI"
Sau khi đặt linh kiện và reflow	AOI 2D/3D với LED đa góc, đa màu và ánh sáng cấu trúc, 8–20 µm mỗi pixel, camera 8–12 MP	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek và các hãng khác	Rà soát AI của OEM (Jutze trong sản xuất, ViTrox V510Ai, Sinic-Tek với Tencent); EMS hàng đầu tự xây (DBG); tiện ích bổ sung của bên thứ ba
Mối hàn đáy và mối hàn bị che	X-ray lấy mẫu hoặc nội tuyến (AXI)	—	Hikvision tuyên bố mô hình Guanlan của họ phán đoán sai lỗ rỗng X-ray dưới 1% số lần vendor claim
Rà soát và sửa chữa	Trạm rà soát trung tâm (một máy chủ, tám dây chuyền), rà soát từ xa, trạm sửa chữa	Phần mềm của các hãng AOI, ViTrox V-ONE	AI khép lại các cảnh báo giả, người xử lý phần còn lại
Kiểm tra thị giác, IQC, FQC	Chủ yếu thủ công; một số máy ngoại quan AI có nạp bằng robot	SmartMore và các hãng khác	Máy ngoại quan trong sản xuất; không có dữ liệu nhân sự công khai cho IQC/FQC
Xử lý và phục vụ máy	Cánh tay sáu trục và cobot, SCARA, AMR, dẫn hướng thị giác 2D/3D	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Nhận thức được học cộng chuyển động được lập trình trong sản xuất; VLA và robot hình người vẫn đang thí điểm
Dữ liệu và hệ thống	MES, IPC-CFX, IPC-HERMES-9852	—	Kết quả được ghi ngược về MES; hỗ trợ ở thiết bị cũ không đồng đều

Nguồn cho bảng: các hồ sơ nêu trên và (Koh Young, Jutze 2025 annual report, Sinic-Tek 2025 annual report, ViTrox, Sohu on Hengtianyi, Xinhua on Hikvision).

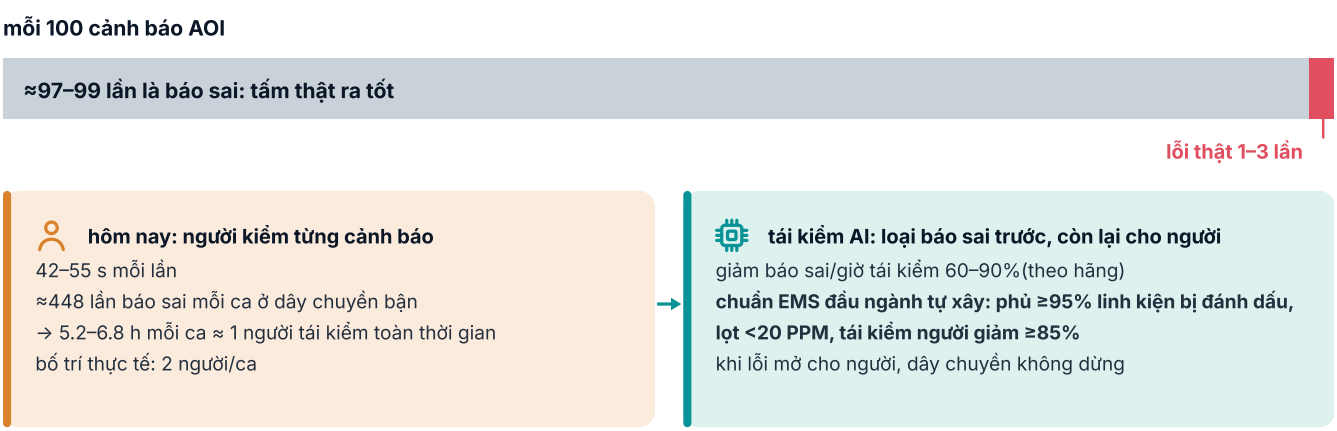
Con người ngồi phía sau AOI, vì hầu hết cảnh báo là giả

Chuỗi kiểm tra trên một dây chuyền SMT/PCBA chính thống chạy: 3D SPI sau khi in cho thiếu, không đủ hoặc thừa kem hàn, cầu hàn và lệch vị trí; AOI sau khi đặt linh kiện (tùy chọn) và sau reflow cho linh kiện thiếu, lệch, xoay,

ngược hoặc dựng đứng (tombstone) và lượng thiếc hàn; X-ray khi có BGA hoặc linh kiện hàn đáy khác; rồi kiểm tra in-circuit hoặc đầu dò bay, kiểm tra chức năng và kiểm tra thị giác thủ công. Mỗi AOI có một chỗ rà soát và một trạm sửa chữa phía sau (Sinic-Tek prospectus, JLCPCB) verified. AOI sau reflow là phổ biến nhất và đáng giá nhất, nhưng nó chỉ thấy hình dạng và diện tích; mối hàn, lỗ rỗng và độ thấm ướt dưới BGA cần X-ray hoặc kiểm tra điện (JLCPCB).

Lao động nằm ở rà soát vì hầu hết cảnh báo là giả. Một EMS Anh ghi nhận **54,019 cảnh báo giả so với 544 khuyết tật thật trên 917 mã linh kiện, khoảng 99% số cảnh báo** (arXiv 2505.03576) single source, preprint; tại nhà máy Rastatt của Siemens, **97.5% số tấm mạch bị AOI loại là tấm tốt** (Siemens case study) vendor claim. Một dây chuyền bận rộn gặp khoảng 448 cảnh báo giả mỗi ca ở mức 42–55 giây mỗi lần, tức 5.2–6.8 giờ mỗi ca, xấp xỉ một người rà soát toàn thời gian (iFactory) vendor claim; số giờ là ước tính.

FIG 03 **Khoảng 97–99% cảnh báo AOI là báo sai** ***Hôm nay người xác nhận từng cảnh báo; tái kiểm AI loại báo sai trước và chuyển phần còn lại cho người.***



Nguồn: Nghiên cứu Siemens, iFactory(theo hãng), báo cáo năm DBG 2025; giờ tái kiểm do tác giả ước tính

Từ "giờ rà soát mỗi ca = số tấm mạch × cảnh báo giả mỗi tấm × giây mỗi lần gọi", một người rà soát có thể phụ trách khoảng 0.7–1 dây chuyền khối lượng với AOI dựa luật. Một khi AI loại bỏ 60–90% cảnh báo giả, chỉ riêng thời gian phán đoán đã cho phép 2–14 dây chuyền mỗi người, nhưng việc đi lại, xử lý bộ đệm NG và bàn giao sửa chữa giới hạn phạm vi bao phủ tại chỗ ở hai đến bốn dây chuyền liền kề; chỉ rà soát trung tâm hoặc từ xa mới tiến gần trần lý thuyết estimate; giây mỗi lần gọi đến từ một nguồn. Vì các AOI chính thống quét 60–80 cm² mỗi giây (Yunheng 2026 AOI guide), một mặt 20×15 cm mất 4–5 giây; AI chỉ cần xử lý vài vùng bị gắn cờ trong một takt, không cần quét lại tấm mạch.

Điều đã hoạt động trong sản xuất: phân loại có giám sát trên chính ảnh của AOI

Mô thức đã được chứng minh là phân loại học sâu có giám sát trên ảnh từ AOI mà một nhà máy đã sở hữu. Tiện ích bổ sung của Fujitsu trên AOI của ADLINK tại Đài Loan cắt giảm cảnh báo giả **hơn 90%** (Fujitsu case); OKI đã dùng AI từ tháng 7 năm 2026 để cắt giảm xác nhận thủ công sau AOI khoảng **80%** trong mảng EMS của chính họ (Circuits Assembly); Siemens báo cáo giảm 60% phân tích thủ công, tiết kiệm €160,287 trong năm đầu và hoàn vốn tám tháng (Siemens case study) tất cả đều là vendor claims. Hai điểm chung của các ca này: hầu như không ai công bố tỷ lệ lọt lưới, và tất cả đều giữ người lại trên các cảnh báo còn lại. Ca duy nhất đưa ra con số lọt lưới (từ 0.8% xuống 0.3%, so với mức hợp đồng 0.5%) thì tự mâu thuẫn (iFactory). AI cắt giảm giờ rà soát, chứ không cắt bỏ chính công việc rà soát.

Rà soát AI đã chuyển từ tiện ích bổ sung của bên thứ ba sang tính năng OEM và hệ thống nội bộ, điều này thay đổi thước đo cho người mới. Báo cáo thường niên 2025 của Jutze liệt kê "nâng cấp AOI 3D" của họ là đang sản xuất khối lượng và nói nó "dùng thuật toán AI để bỏ rà soát thủ công ở đầu kiểm tra" (Jutze 2025 annual report); Sinic-Tek đã phát triển "hệ thống rà soát thủ công hỗ trợ AI đa phương thức" từ năm 2019 và đang đồng phát triển một

AOI 3D trên-dưới với Tencent, nhắm tới độ bao phủ khuyết tật 99.5%+ và độ chính xác nhận diện 98% (Sinic-Tek 2025 annual report) vendor claims in primary filings. Các công ty EMS lớn tự xây hệ thống. Báo cáo thường niên 2025 của DBG liệt kê "phán đoán AI cho kiểm tra thị giác" là một dự án đã hoàn thành, "bao phủ hơn 95% linh kiện bị AOI gắn cờ", với "tỷ lệ lọt lưới dưới 20 PPM" và "giảm hơn 85% rà soát thủ công", truy xuất đầy đủ (DBG 2025 annual report) single source, primary. Đó là con số nhà máy điện tử duy nhất tôi tìm thấy cho đồng thời độ bao phủ, tỷ lệ lọt lưới và mức giảm. **Một sản phẩm bên thứ ba bán cho EMS cấp 1 phải sánh được mức 95% bao phủ, 20 PPM lọt lưới và 85% giảm của DBG**; dưới cấp 1, chuẩn mực là mức giảm 50–85% mà các nhà cung cấp quảng cáo.

Phát hiện dị thường chỉ huấn luyện trên sản phẩm tốt thì nhanh (EfficientAD-S mất 2.2 ms mỗi ảnh (EfficientAD, WACV 2024)) nhưng yếu với dị thường "logic" như thiếu hoặc sai linh kiện, và các phương pháp tốt nhất trung bình dưới 60% AU-PRO trên MVTec AD 2 khó hơn (MVTec AD 2) single source. Thiếu và sai linh kiện thì tốt hơn nên đối chiếu với CAD và BOM.

Trần ở bất kỳ chi phí nào: nhiều lớp máy, rất ít người, nhưng không phải bằng không

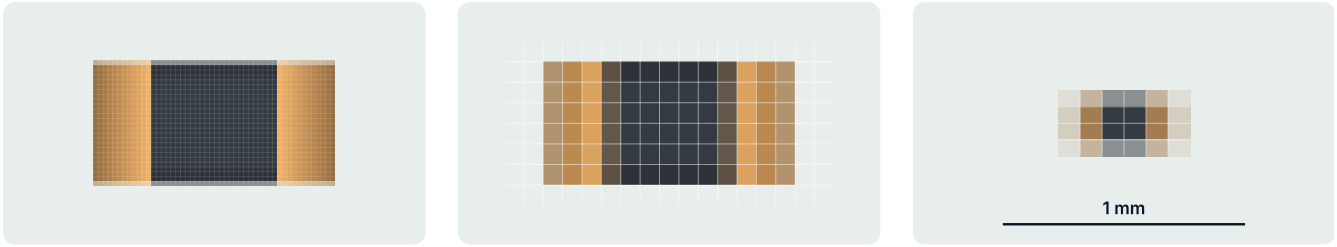
Tôi không tìm thấy nhà máy điện tử nào bỏ hoàn toàn rà soát thủ công và đạt không lọt lưới. Thực hành cho sản phẩm IPC-A-610 Class 3 (độ tin cậy cao) vẫn cần người kiểm tra đã đào tạo quyết định các ca ranh giới mà ngưỡng tự động không phân loại được, và một nhà lắp ráp PCB gọi "kiểm tra tự động có thể thay thế chữ ký của con người" là một "hiểu lầm phổ biến" (PCBCart) single source. Đường cơ sở con người cũng cần hiệu chỉnh: dẫn một nghiên cứu Sandia 2012, một người kiểm tra thị giác đơn lẻ giỏi nhất cũng chỉ bắt được khoảng **80%** khuyết tật, hai người nối tiếp khoảng 96% (Instrumental) single source, được một nhà cung cấp trích dẫn. Tuyên bố có thể bảo vệ là "AI cộng một vòng rà soát nhỏ của con người đánh bại kiểm tra hoàn toàn thủ công", chứ không phải "AI thay thế kiểm tra".

Không giới hạn ngân sách, trần là 3D SPI, AOI 3D, X-ray, rà soát AI và các trạm ngoại quan AI xếp chồng lên nhau, với người chỉ còn lại cho các ca ranh giới, sản phẩm mới và chuyển đổi, loại khuyết tật mới và chữ ký của khách hàng. Các nhà máy flagship báo cáo giảm 85–92% kiểm tra thủ công (I-Connect007, 2021, WiAdvance) vendor claims; ước tính của tôi cho trần không giới hạn ngân sách là lao động rà soát ở mức 8–15% so với hiện tại estimate. Lắp ráp cuối cùng vẫn tốn lao động: tự động hóa các nút của iPhone 16 bị bỏ vì tỷ lệ khuyết tật cao (9to5Mac) single source.

Một camera độ phân giải cao thông thường không phải là AOI

Các AOI sản xuất chụp ảnh theo từng ô ở 8–20 µm mỗi pixel, với camera 8–12 MP, chiếu sáng đa góc đa màu có thể lập trình và 3D ánh sáng cấu trúc (Koh Young, Saki) verified. Ngành đã nhận ra từ lâu rằng một linh kiện 01005 chỉ phủ khoảng 8×16 pixel ở 25 µm mỗi pixel, "có thể không đủ" cho phán đoán tin cậy (Circuits Assembly, 2007). Với cùng phép tính estimate, một camera 4K trên một tấm 300 mm đạt khoảng **78 µm mỗi pixel**: một linh kiện 0402 hệ inch là khoảng 13×6 pixel, đủ cho sự hiện diện và lệch thô nhưng không đủ cho hình dạng mối hàn, và một 01005 là khoảng 5×3 pixel, gần như vô hình. Đạt 15 µm mỗi pixel trên toàn tấm sẽ cần khoảng 330 megapixel, đó là lý do AOI di chuyển camera trên tấm mạch. Khi ai đó nói "0402", hãy kiểm tra họ muốn nói loại nào: 0402 hệ mét là 01005 hệ inch, vượt xa bất kỳ camera góc rộng nào.

FIG 04 **Camera đa năng không phải là AOI** ***Camera 4K trên tấm 300 mm cho ≈78 µm/px: linh kiện 0402 còn ≈13×6 px và 01005 biến mất.***



AOI: 8–20 µm/px
0402 ≈100×50 px

Camera 4K nhìn tấm 300 mm: ≈78 µm/px
0402 chỉ còn ≈13×6 px: đủ để biết có/không và lệch rõ

cùng camera nhìn 01005
≈5×3 px: thực tế không thấy

Nguồn: Circuits Assembly; tính toán của tác giả cho tấm 300 mm và cảm biến 4K. Lưới pixel là minh họa

Trên các dây chuyền đã có AOI, chỗ tốt nhất cho AI là ở trên chính ảnh của AOI. Chỉ các dây chuyền không có AOI, hoặc các trạm cuối dây chuyền, mới biện minh cho một trạm camera kín có chiếu sáng kiểm soát; một camera độ phân giải cao mở trên dây chuyền là nơi dễ bị ảnh hưởng nhất bởi ánh sáng môi trường và bóng đổ (CSDN).

Lắp ráp và xử lý: nhận thức được học cộng chuyển động được lập trình đang trong sản xuất

Báo cáo 2026 của IFR nói "thị giác máy cho kiểm tra và dẫn hướng robot là chín muồi và đạt cấp công nghiệp", trong khi "AI vật lý đa dụng và học từ mô phỏng và trải nghiệm thì chưa đạt cấp công nghiệp" (IFR World Robotics 2026 executive summary) single source, primary. Điện tử là ngành mua robot lớn nhất (142,235 chiếc lắp đặt toàn cầu năm 2025, 59% tổng số lắp đặt tại Trung Quốc (IFR, IFR China release)), nhưng Trung Quốc chỉ có 166 robot trên 10,000 công nhân sản xuất so với 307 ở Mỹ (IFR, April 2026). Mô thức trong sản xuất là nhận thức được học cộng chuyển động xác định; chuyển động học đầu-cuối chỉ tiệm cận cấp sản xuất ở các trạm có cấu trúc (Chương 3 và 6).

Mỗi khu vực dùng được gì

Trung Quốc đại lục dựa vào các mô hình nội địa và trọng số nội địa; ở Đài Loan danh tính khách hàng quyết định; ở Mỹ, thuế quan và chứng nhận quyết định.

Yếu tố	Trung Quốc đại lục	Đài Loan (nhà máy tư nhân)	Đông Nam Á	Hoa Kỳ
Claude (bất kỳ kênh nào)	Không khả dụng	Có điều kiện: nhà cung cấp không được sở hữu đa số hoặc bị kiểm soát bởi cá nhân hay thực thể ở Trung Quốc đại lục, Hồng Kông hoặc Ma Cao	Tương tự	Tương tự
Đám mây GPT và Gemini	Không khả dụng	Khả dụng	Khả dụng (danh sách chưa kiểm từng nước)	Khả dụng
Các mô hình đám mây Trung Quốc (Qwen, DeepSeek, Doubao, GLM, Kimi)	Khả dụng (mô hình đã đăng ký; cung cấp dịch vụ cần đăng ký)	Hợp pháp, nhưng hosting tại đại lục hiếm khi qua đánh giá an ninh của khách hàng; cấm trong khu vực công, doanh nghiệp nhà nước và hạ tầng trọng yếu	Hợp pháp (nút quốc tế)	Hợp pháp, đánh giá nghiêm ngặt

Yếu tố	Trung Quốc đại lục	Đài Loan (nhà máy tư nhân)	Đông Nam Á	Hoa Kỳ
Trọng số mở nội địa gốc Trung Quốc (Qwen, GLM, DeepSeek)	● Lựa chọn hàng đầu; triển khai nội bộ tư nhân không cần đăng ký mô hình hay thuật toán (cách đọc của luật sư)	● Khả dụng (khu vực công bị hạn chế)	● Khả dụng	● Có điều kiện: đánh giá nguồn gốc
Trọng số mở nội địa gốc không phải Trung Quốc (ví dụ Gemma 4)	● Khả dụng	● Khả dụng	● Khả dụng	● Khả dụng
Các mô hình thị giác không sinh truyền thống	● Khả dụng (ngoài các quy tắc AI sinh)	● Khả dụng	● Khả dụng	● Khả dụng
Jetson Thor/T5000/T4000, DGX Spark	● Khả dụng: NVIDIA phân loại chúng 5A992.c, không cần giấy phép Mỹ cho nhà máy dân sự thông thường; Spark có SKU Trung Quốc; giá kênh cao hơn giá niêm yết Mỹ khoảng 40–60%	● Khả dụng	● Khả dụng	● Khả dụng
RTX 5090, RTX PRO 6000, DGX Station	● Không khả dụng: hai cái đầu là 4A090.a, DGX Station 5A992.z; thay thế hợp pháp là 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	● Khả dụng	● Khả dụng, nhưng người mua có trụ sở tại hoặc sở hữu cuối từ Trung Quốc tính là Trung Quốc; Malaysia yêu cầu giấy phép xuất khẩu và chuyển tải	● Khả dụng
Mac Studio M5 Ultra 512 GB	● Khả dụng (Apple Store Trung Quốc)	● Khả dụng	● Được cho là khả dụng	● Khả dụng
Huawei Ascend và các bộ tăng tốc Trung Quốc khác	● Khả dụng	● Rủi ro cao	● Rủi ro cao	● Không dùng được trên thực tế
Cánh tay cố định, AOI, camera công nghiệp sản xuất tại đại lục	● Khả dụng	● Nhập được; từ ngày 1 tháng 12 năm 2025, nhà chức trách có thể can thiệp trực tiếp với thương hiệu đại lục, và các cơ quan, doanh nghiệp nhà nước, một số quỹ và nhà cung cấp hạ tầng trọng yếu có thể bị hạn chế hoặc cấm dùng	● Khả dụng	● Hợp pháp, nhưng thuế tổng hợp khoảng 40% với cánh tay, 37.5% với thiết bị dạng AOI và camera, 25% với máy tính biên
AMR và bộ thao tác di động ngoại	● Khả dụng	● Khả dụng	● Khả dụng	● Mẫu mới cần phê duyệt có điều kiện từ Bộ Chiến tranh Mỹ
Camera Hikvision và Dahua	● Khả dụng	● Dùng tư nhân được phép, khu vực công bị cấm	● Khả dụng	● Ủy quyền FCC mới bị chặn; cấm cho nhà thầu liên bang
An toàn và chứng nhận robot	● GB 11291.1-2011 và GB 11291.2-2013 bắt buộc, đang sửa đổi từ tháng 4 năm 2026; chứng nhận CR tự nguyện; GB/T 39404-2026 bao phủ an ninh mạng bộ điều khiển	● Mặc định có rào; hợp tác không rào phải đáp ứng CNS 14490 hoặc ISO 10218 với đánh giá bằng văn bản	● Theo khách hàng và điểm đến xuất khẩu, xác nhận bằng văn bản trước khi giao hàng	● Thiết bị điện phải được "phê duyệt" (trên thực tế là dấu NRTL hoặc đánh giá hiện trường); UL 1740 yêu cầu cụm robot đáp ứng ISO 10218-2 từ ngày 11 tháng 8 năm 2025

Yếu tố	Trung Quốc đại lục	Đài Loan (nhà máy tư nhân)	Đông Nam Á	Hoa Kỳ
Nhà máy wafer và nhà máy sản phẩm có kiểm soát	—	Hợp đồng mua của TSMC yêu cầu SEMI E187; thiết bị đại lục đang bị loại khỏi các fab 2 nm	—	Dữ liệu kỹ thuật ITAR bao gồm "ảnh chụp"; chính sách từ chối với Trung Quốc; nhà cung cấp do đại lục kiểm soát trên thực tế không thể phục vụ nhà máy điện tử quốc phòng

Nguồn cho bảng: (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, Taiwan BOFT import list, Taiwan Cyber Security Management Act history, Taiwan 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).

Trung Quốc đại lục: một bộ giải pháp đầy đủ, nhưng cả mô hình frontier đóng lẫn máy đơn mạnh nhất đều không vào được. Các mô hình đám mây nội địa đã đăng ký, trọng số mở nội địa (Qwen theo Apache-2.0, GLM và DeepSeek theo MIT), các mô hình thị giác truyền thống và AOI, camera và cobot nội địa đều dùng được. Phân loại xuất khẩu của chính NVIDIA liệt kê bộ Jetson AGX Thor, các module T5000/T4000 và DGX Spark là 5A992.c, nên một nhà máy EMS hoặc PCBA thông thường không cần giấy phép Mỹ, trong khi RTX 5090 và RTX PRO 6000 (4A090.a) và DGX Station (5A992.z, TPP 58,484) không thể mua ở đại lục (NVIDIA export classifications, eCFR) primary; cách đọc pháp lý là của tôi và không phải tư vấn pháp lý. RS China niêm yết bộ Thor ở mức 59,104 RMB đã gồm thuế, khoảng 1.4 lần giá niêm yết Mỹ (RS China).

Bốn công ty luật đồng thuận rằng triển khai nội bộ tư nhân cho mục đích riêng của nhà máy không cần đăng ký mô hình hay thuật toán, trong khi dịch vụ đa người thuê có hosting nằm trong vùng xám (Covington, FPF, JunHe) verified: các cách đọc nhất quán, không có phán quyết chính thức. GB 11291.1-2011 và GB 11291.2-2013 vẫn là chuẩn mực an toàn robot bắt buộc và đang được sửa đổi từ ngày 28 tháng 4 năm 2026; GB/T 39404-2026 về an ninh mạng bộ điều khiển có hiệu lực từ ngày 1 tháng 11 năm 2026 (revision plan 20262542, GB/T 39404-2026) single source, primary. Cho tới khi GB 11291 mới ra đời (2028–2029 estimate), hãy thiết kế theo ISO 10218:2025 và giữ hồ sơ tuân thủ GB 11291.

Đài Loan: nhà máy tư nhân dùng được gần như mọi thứ; danh tính khách hàng quyết định. Các API frontier, trọng số mở và thiết bị đại lục đều hợp pháp (data.gov.tw). Nhưng kể từ khi Đạo luật Quản lý An ninh mạng sửa đổi có hiệu lực ngày **1 tháng 12 năm 2025**, nhà chức trách có thể can thiệp trực tiếp với sản phẩm là "thương hiệu đại lục" và hạn chế chúng tại cơ quan chính phủ, doanh nghiệp nhà nước và nhà cung cấp hạ tầng trọng yếu (Laws & Regulations Database, review rules) verified; danh sách nhà cung cấp hạ tầng trọng yếu là bảo mật; và TSMC đã đưa SEMI E187 vào hợp đồng mua của mình (CNA) và được cho là đang loại thiết bị đại lục khỏi các fab 2 nm (TrendForce) single source. EMS tư nhân, lắp ráp và các nhà máy PCBA nhỏ thì mở; fab, tiện ích khu công nghiệp và khách hàng liên quan chính phủ thì đóng trên thực tế. Một nhà cung cấp do đại lục kiểm soát nên thiết kế theo E187, cung cấp danh mục vật liệu phần mềm (SBOM), cập nhật ngoại tuyến và không bao giờ kết nối từ xa từ đại lục.

Đông Nam Á: ít hạn chế nhất; vấn đề là hoàn vốn chậm và công ty mẹ Trung Quốc. Mọi thứ đều dùng được, nhưng hướng dẫn của Mỹ ngày 31 tháng 5 năm 2026 yêu cầu giấy phép để xuất khẩu các hạng mục điện toán tiên tiến cho các thực thể có trụ sở tại, hoặc sở hữu cuối từ, Trung Quốc ngay cả khi bên nhận ở nước thứ ba (Greenberg Traurig) single source, và Malaysia có giấy phép riêng cho chip tiên tiến (Fulcrum). Nhu cầu đang dịch chuyển về đây: TrendForce dự báo Thái Lan, Việt Nam và Malaysia sẽ tiến tới 50% công suất SMT của server-ODM trong năm 2026 (TrendForce) single source, dự báo 2023.

Hoa Kỳ: bộ giải pháp đầy đủ, với cái giá là thuế quan, chứng nhận và dữ liệu có kiểm soát. Theo chương 99 của biểu thuế Mỹ, một cánh tay robot sản xuất tại đại lục chịu khoảng 40% thuế tổng hợp, thiết bị dạng AOI và camera khoảng 37.5% và máy tính biên 25%; sản xuất tại Malaysia hoặc Indonesia khoảng 10–12.5%, tại Việt Nam, Thái Lan hoặc Philippines 12.5–15%, tại Đài Loan 10%, với máy tính ở 0% ngoài đại lục (USITC HTS chapter 99, Great Lakes Customs Law) primary schedule + estimate. Thông báo ngày 10 tháng 11 của USTR gia hạn 178 miễn trừ Section 301 cũ

đến ngày 9 tháng 1 năm 2027, không miễn trừ nào bao phủ robot, thiết bị kiểm tra, camera hay máy tính (Federal Register 2026-22925) verified.

OSHA yêu cầu thiết bị điện nơi làm việc phải được "phê duyệt" (trên thực tế là dấu NRTL, hoặc đánh giá hiện trường khoảng $3,000 đến $20,000+ mỗi cụm), và từ ngày 11 tháng 8 năm 2025 UL 1740 đưa toàn bộ cụm robot vào ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). Điện tử quốc phòng đóng với nhà cung cấp do đại lục kiểm soát trên thực tế: ITAR coi ảnh sản xuất là dữ liệu kỹ thuật và áp dụng chính sách từ chối với Trung Quốc (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Một cụm xuất sang Mỹ nên dùng cánh tay cố định, camera không hạn chế và mô hình nội địa hoặc đám mây Mỹ, với lắp ráp cuối tại Đông Nam Á nơi điều đó thực sự thay đổi xuất xứ.

Ở bất kỳ ngân sách nào, trần khác nhau theo khu vực nhưng không nơi nào đạt không người estimate. Đại lục có thể xếp chồng 3D SPI, AOI 3D, X-ray, rà soát AI nội địa và một đội AI nội bộ để hạ lao động rà soát xuống 8–15%, nhưng thiếu mô hình frontier đóng và DGX Station, nên Hỏi&Đáp LLM toàn nhà máy phải chạy trên trọng số mở nội địa và tính toán nội địa (một mô hình kích thước đầy đủ cho nhiều người dùng cần ít nhất hai máy chủ Ascend tám card). Đài Loan và Đông Nam Á về kỹ thuật dùng được thứ tốt nhất của mọi thứ, kể cả API frontier và DGX Station, bị giới hạn bởi danh tính khách hàng, thương hiệu đại lục và cấp phép công ty mẹ Trung Quốc. Mỹ không có giới hạn kỹ thuật, với cái giá là thuế quan, chứng nhận NRTL và tách biệt nhân sự cho dữ liệu có kiểm soát. Ở mọi nơi, con người vẫn còn cho các ca ranh giới, sản phẩm mới và chuyển đổi, loại khuyết tật mới và chữ ký khách hàng.

Một sản phẩm bán xuyên khu vực do đó cần hai bộ mô hình: một bộ cho đại lục trên các mô hình nội địa đã đăng ký hoặc trọng số mở nội địa, và một bộ cho nước ngoài trên API frontier cộng một mô hình nội địa gốc không phải Trung Quốc. Chính sách của Anthropic có hiệu lực từ ngày 12 tháng 11 năm 2026 phán đoán theo sở hữu và kiểm soát, nên một nhà cung cấp do đại lục kiểm soát không thể xây trên Claude ngay cả qua công ty con ở nước ngoài, nền tảng đám mây hay đại lý (Anthropic) single source, primary.

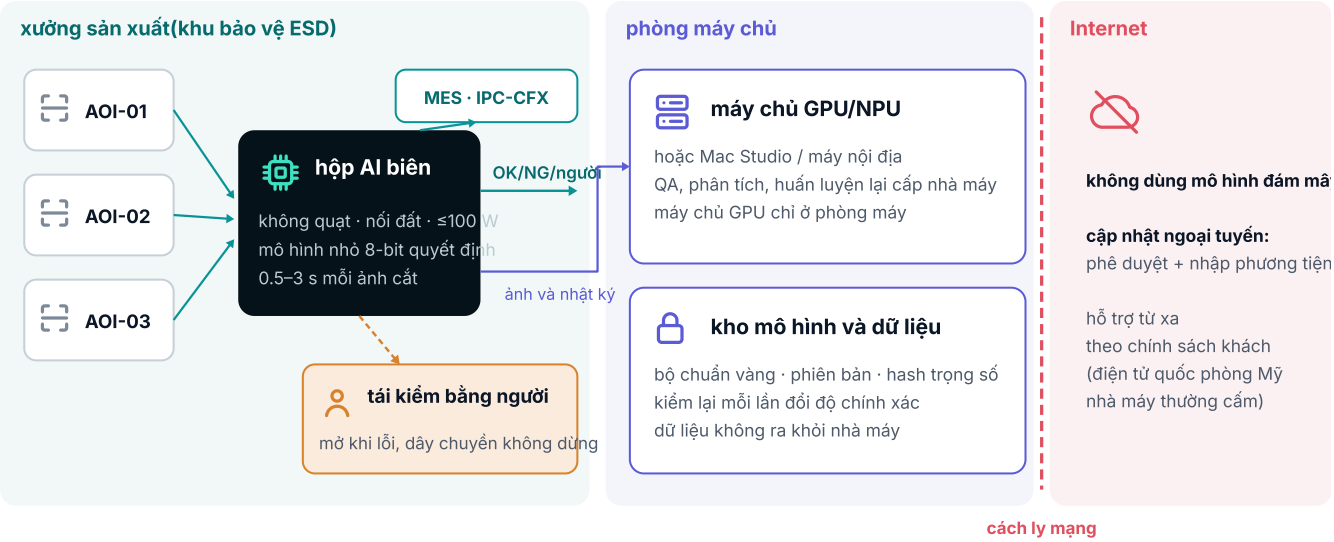
Tại chỗ

02 Chạy AI bên trong nhà máy: phòng sạch, cách ly mạng, phần cứng và lượng tử hóa

Trong một nhà máy cách ly mạng, ràng buộc là quy trình, không phải tính toán

Kiểm tra thị giác hoàn toàn ngoại tuyến là thông lệ; các ràng buộc là về quy trình. Hướng dẫn MIIT 2024 của Trung Quốc về an ninh hệ thống điều khiển công nghiệp yêu cầu phân vùng, tường lửa hoặc diode dữ liệu, bảo trì từ xa có giới hạn thời gian được kiểm toán, phương tiện được quét và bản vá đã kiểm thử (guide text) single source ; khách hàng bán dẫn thêm SEMI E187 và E188, thứ TSMC đã đưa vào hợp đồng mua năm 2023 (SEMI, TXOne) verified . Kết quả được ghi về MES qua IPC-CFX và danh tính tấm mạch truyền qua IPC-HERMES-9852 (Hermes standard). Bẫy phổ biến là cấp phép giả định có mạng: chế độ ngoại tuyến của Roboflow lưu trọng số trên hợp đồng thuê 30 ngày (Roboflow). Một sản phẩm ngoại tuyến cần kích hoạt ngoại tuyến, cập nhật có chữ ký trên phương tiện đã quét, danh mục vật liệu phần mềm, kiểm soát USB, nhật ký kiểm toán nội bộ và một bộ điều hợp CFX.

FIG 05 **Triển khai cách ly mạng** ***Hộp nhỏ tại dây chuyền quyết định; phòng máy chủ trả lời câu hỏi; khi lỗi chuyển về người và dữ liệu không rời nhà máy.***



Nguồn: Tổng hợp của tác giả(Chương 2)

Các mô hình frontier đóng trên thực tế không vào phòng máy chủ của nhà máy. Gemini chỉ được cung cấp cách ly mạng qua Google Distributed Cloud cho chính phủ và doanh nghiệp lớn (Google Cloud); Foundry Local của Microsoft chỉ mở các mô hình lớn cho "khách hàng đủ điều kiện" và không nêu tên mô hình GPT nào (Microsoft); Claude chỉ được cung cấp ở các khu vực đám mây chính phủ, không có triển khai tại chỗ (Claude for government); cả ba đều không được cung cấp ở Trung Quốc đại lục. Thiết bị nhà máy phải được thiết kế quanh trọng số mở; các mô hình frontier nhiều nhất chỉ có thể gán nhãn hoặc giải thích ảnh tại trụ sở, ngoại tuyến, khi khách hàng cho phép ảnh ẩn danh ra ngoài.

Phần cứng phòng máy chủ nên được coi là cấp văn phòng: Mac Studio được định mức 10–35 °C và DGX Spark 5–30 °C, với các báo cáo về tắt máy do nhiệt dưới tải suy luận liên tục (Apple, NVIDIA forum) verified . Dây chuyền dùng hộp công nghiệp không quạt, như một PC công nghiệp Jetson T4000 định mức −25 đến 60 °C (AAEON) vendor

claim . Vì một giờ ngừng máy khiến một nhà máy lớn mất $36,000 đến $2.3 triệu (Siemens True Cost of Downtime 2024) vendor survey , mọi bước AI trên dây chuyền phải **mở khi lỗi (fail open)**: khi có bất kỳ sự cố nào, nó rơi về quy trình AOI-cộng-người hiện có.

Các khu vực sạch và ESD: một xưởng SMT thông thường là khu vực bảo vệ ESD, không phải phòng sạch phân cấp

Điều giới hạn nơi đặt tính toán là tĩnh điện và luồng khí, không phải cấp sạch. Hầu hết các khu vực lắp ráp SMT "không được kiểm soát hạt chính thức", và các nhà máy đẳng cấp thế giới "có thể đạt hoặc tiệm cận ISO 8" (Circuitnet expert panel) single source . PCBA y tế và module camera thường ở ISO 7, die trần và căn chỉnh quang ở ISO 5–6 cục bộ (PCBSync) single source ; các nhà xây phòng sạch báo giá ISO 7–8 cho SMT nói chung (Angstrom Technology) vendor claim .

Yêu cầu phổ quát là kiểm soát ESD: ANSI/ESD S20.20-2021 yêu cầu vật dẫn cách ly dưới 35 V và giữ vật cách điện trường mạnh cách thiết bị nhạy ít nhất 30 cm (ANSI, ESDA, Desco) verified . Các fab đặt tính toán ở sub-fab hoặc trung tâm dữ liệu, trong các tủ làm mát bằng chất lỏng đáp ứng SEMI S2/S8/F47 (Prodrive) vendor claim .

Việc thiết bị có được vào khu vực sạch hay không được kiểm tra phát thải hạt theo ISO 14644-14; robot có các chứng chỉ như vậy (Fraunhofer IPA), nhưng tôi không tìm thấy PC công nghiệp, hộp edge AI hay máy chủ GPU nào có. CE, FCC và IEC 60068 cho thấy một hộp chịu được môi trường của nó, không phải rằng nó được vào khu vực bảo vệ ESD; điều phối viên ESD của khách hàng kiểm tra điều đó. Máy chủ GPU không bao giờ đặt trên dây chuyền: một DGX Station rút 1,292 W ở GPU trong thử nghiệm (StorageReview).

FIG 06 **Khu sạch và ESD quyết định chỗ đặt máy tính** ***Xưởng SMT thường là khu bảo vệ ESD, không phải phòng sạch. Máy chủ GPU và NPU thuộc về phòng máy chủ.***



Nguồn: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, tiêu chuẩn SEMI và tài liệu hãng; tác giả tổng hợp

Khu vực	Môi trường điển hình	Tính toán tại dây chuyền	Máy chủ GPU/NPU
Xưởng SMT/PCBA thông thường (tiêu dùng, công nghiệp)	Khu vực bảo vệ ESD theo S20.20 hoặc IEC 61340-5-1, thường không phải phòng sạch phân cấp (≈tương đương ISO 8); 40–60% RH	Hộp không quạt trong vỏ kim loại tiếp đất (Jetson T4000/Thor, hộp NPU Trung Quốc) cạnh AOI hoặc chỗ rà soát; màn hình và vỏ nhựa ra khỏi khu vực hoặc cách tấm mạch ≥30 cm	Phòng máy chủ hoặc tủ IT có điều hòa
Y tế, module camera, điện tử EV (phòng sạch ISO 7–8)	Phòng sạch cộng bảo vệ ESD	Không quạt, kín IP65/66, nhẵn và dễ lau; yêu cầu dữ liệu ISO 14644-14 hoặc chứng chỉ IPA	Ngoài phòng sạch, kết nối mạng vào

Khu vực	Môi trường điển hình	Tính toán tại dây chuyền	Máy chủ GPU/NPU
Die trần, hàn dây, trạm quang (ISO 5–6 cục bộ)	Môi trường mini	Chỉ camera và đèn bên trong; hộp tính toán bên ngoài	Phòng máy chủ
Phòng sạch fab (ISO 5 và sạch hơn)	Kế hoạch ESD và hút tĩnh điện riêng của fab, thiết bị theo SEMI E78; an ninh mạng theo SEMI E187/E188	Chỉ bộ điều khiển tích hợp công cụ, đã qua phê duyệt của fab; không PC bổ sung có quạt	Tủ làm mát lỏng sub-fab hoặc trung tâm dữ liệu fab

Bảng này là tổng hợp của tôi từ các nguồn trên estimate. Với một sản phẩm, điều đó nghĩa là một hộp không quạt cạnh dây chuyền dưới khoảng 100 W, vỏ kim loại nối đất bảo vệ, đầu vào DC 12–24 V, hoạt động dải nhiệt rộng ở luồng khí tủ thực tế, và một báo cáo phát thải hạt ISO 14644-14 sẵn sàng cho khách hàng ISO 7 trở lên. Bất kỳ thiết kế nào mà "một máy chủ GPU phục vụ bốn dây chuyền" đều đặt máy chủ trong phòng máy chủ và chỉ để một hộp nhỏ hoặc một camera tại dây chuyền.

Giới hạn tốc độ theo mục đích: độ chính xác khó hơn tốc độ

Các nhà máy và nhà cung cấp hầu như không bao giờ công bố mục tiêu độ trễ hoặc tính sẵn sàng của AI. Các ngưỡng dưới đây được suy ra từ takt, thời gian rà soát thủ công, tốc độ điều khiển robot và nghiên cứu tương tác estimate:

Mục đích	Giới hạn tốc độ	Giới hạn độ chính xác	Tính sẵn sàng
Kiểm tra nội tuyến từng linh kiện	Trong ½–1 takt: 240–1,000 ảnh cắt trong takt 10–30 s, 8–100 ảnh cắt mỗi giây	Lọt lưới không tệ hơn "AOI cộng người" hiện nay (ví dụ hợp đồng <0.5%)	Mở khi lỗi, không bao giờ dừng dây chuyền; mục tiêu ≥99.9% (không có SLA công khai)
Ý kiến thứ hai nội tuyến chỉ trên ảnh cắt bị AOI gắn cờ	Các cờ (thường ≤5–20 mỗi tấm) trong ½ takt: 0.5–3 s mỗi ảnh cắt	Tương tự; huấn luyện trên chính ảnh của dây chuyền	Rơi về rà soát thủ công
Rà soát AI các cảnh báo giả	2–5 s mỗi lần gọi, nhanh hơn mười lần so với 42–55 s của người	Lọt lưới không tệ hơn người rà soát	Suy giảm thì thêm người, không bao giờ dừng dây chuyền
Điều khiển robot (VLA)	100–300 ms mỗi đoạn hành động với phân đoạn thời gian thực, đầu ra 10–50 Hz	≥99% thành công nhiệm vụ để thắng chuyển động được lập trình	Hệ thống an toàn độc lập với mô hình
Lập kế hoạch và chọn kỹ năng robot	1–3 s khi cần mỗi chu kỳ; ≤1 phút nếu chỉ khi ngoại lệ hoặc chuyển đổi	Đầu ra phải là mã kiểm chứng được hoặc một ID kỹ năng	Rơi về chương trình cố định
Hỏi&Đáp (SOP, truy vấn MES)	Token đầu lý tưởng ≤1–2 s, tối đa 10 s; ≥15–20 token/s mỗi người dùng; 5–20 người dùng đồng thời	Trích dẫn chính xác; trung thành trên ngữ cảnh dài	Cấp IT văn phòng (≈99–99.5%)
Báo cáo phân tích và nguyên nhân gốc	Vài phút là được (batch)	Cao nhất; trung thành trên ngữ cảnh dài	Cấp IT văn phòng

Physical Intelligence đo được 97 ms suy luận cho một VLA cỡ 3B (khoảng 108–139 ms gồm mạng và tiền xử lý), thấy rằng phân đoạn thời gian thực giữ thông lượng không đổi tới 200 ms độ trễ thêm, và hoàn thành các nhiệm vụ tinh như đánh diêm hoặc cắm cáp mạng ở hơn 300 ms (Physical Intelligence) vendor research; nghiên cứu tương tác đặt giới hạn để giữ mạch suy nghĩ ở một giây và cho chú ý ở mười giây (NN/g). Độ chính xác mới là phần khó: Qwen3-VL-8B mở chỉ đạt trung bình 56.8 zero-shot trên năm bộ dữ liệu dị thường công nghiệp (IndusAgent) single source.

Không phần cứng nào cứu được một mô hình ngôn ngữ-thị giác chưa được tinh chỉnh trên chính ảnh của dây chuyền.

FIG 07 **Giới hạn tốc độ theo mục đích*Kiểm inline tính bằng mili giây, QA bằng giây, phân tích bằng phút. Người mất 42–55 s mỗi lần.***



Nguồn: Tổng hợp của tác giả(Chương 1–2)

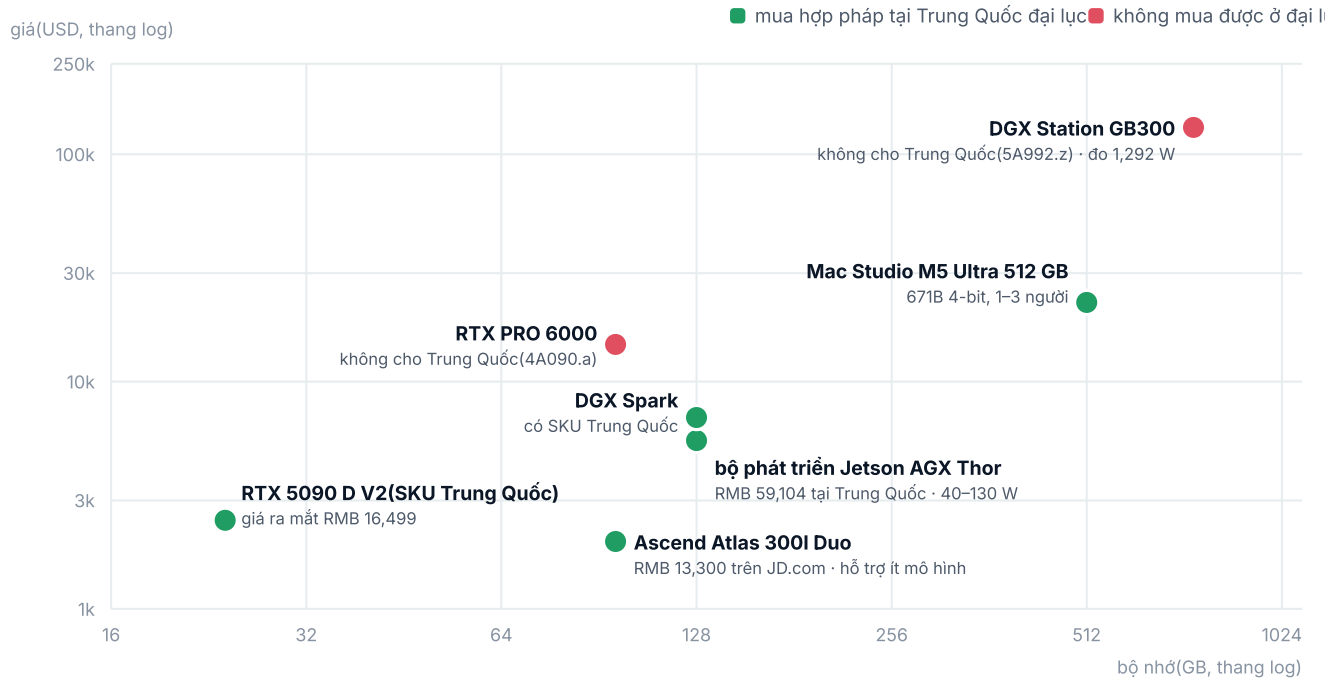
Phần cứng: Thor cho trạm, Mac Studio hoặc thiết bị nội địa cho Hỏi&Đáp nhà máy, và không DGX Station ở đại lục

Thiết bị	Bộ nhớ và băng thông	Giá (tháng 11 năm 2026)	Phù hợp cho việc gì	Trung Quốc đại lục
Bộ phát triển Jetson AGX Thor / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 khi ra mắt → $5,499 (tháng 7 năm 2026); T5000 $4,999, T4000 $2,999; bộ phát triển RS China 59,104 RMB đã gồm thuế	Mô hình thị giác 2–8B kiểm lại ảnh cắt; VLA cỡ 3B thời gian thực	Có (5A992.c, không cần giấy phép cho nhà máy dân sự)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; mẫu Trung Quốc ASUS GX10 cùng chip 32,999 RMB trên JD (tháng 10 năm 2025)	MoE 8–35B đa người dùng	Có (5A992.c, mẫu Trung Quốc, kênh ủy quyền)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (tháng 6 năm 2026) → ≈$16,005 giá niêm yết eBay trung bình tháng 11	FP8 27–35B đa người dùng; MoE cỡ 120B ở 4-bit	Không (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB hợp nhất, 1.2 TB/s; 480 W duy trì tối đa	Từ $22,299 tại Mỹ; ≈173–186k RMB ước tính tại Trung Quốc	DeepSeek 671B ở 4-bit cho 1–3 người dùng	Có
DGX Station GB300	748 GB (252 GB HBM3e ở 7.1 TB/s + 496 GB LPDDR5X ở 396 GB/s); 1,292 W đo ở GPU	$123–140k giá thị trường tháng 11 (≈830–940k RMB trước thuế); $95–117k tháng 8	Đồng thời cao trên MoE 120–235B vừa HBM; chỉ vài người dùng trên mô hình 433–743B	Không (5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	13,300 RMB từ một nhà bán JD	Rà soát batch khối lượng thấp; ít mô hình được hỗ trợ	Có (nội địa)

Thiết bị	Bộ nhớ và băng thông	Giá (tháng 11 năm 2026)	Phù hợp cho việc gì	Trung Quốc đại lục
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Không lấy được; DeepSeek đầy đủ cần hai máy ở W8A8, bốn máy ở BF16	Hỏi&Đáp và phân tích đa người dùng toàn nhà máy	● Có (nội địa)
RTX 5090 D V2 (mẫu Trung Quốc)	24 GB	16,499 RMB khi ra mắt (tháng 8 năm 2025), tăng theo giá bộ nhớ	Mô hình thị giác 8B, VLA nhỏ	● Có

Nguồn: (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). Giá đại lục cho Mac Studio 512 GB được ngoại suy từ phụ phí cấu hình Mỹ và có độ tin cậy thấp estimate.

FIG 08 **Tính toán on-prem: bộ nhớ, giá và thứ mua hợp pháp ở Trung Quốc đại lục*Phân loại xuất khẩu quyết định danh sách trước cả benchmark.***



Nguồn: Phân loại xuất khẩu NVIDIA, Apple, RS China, JD.com và giá đánh giá(tháng 11/2026)

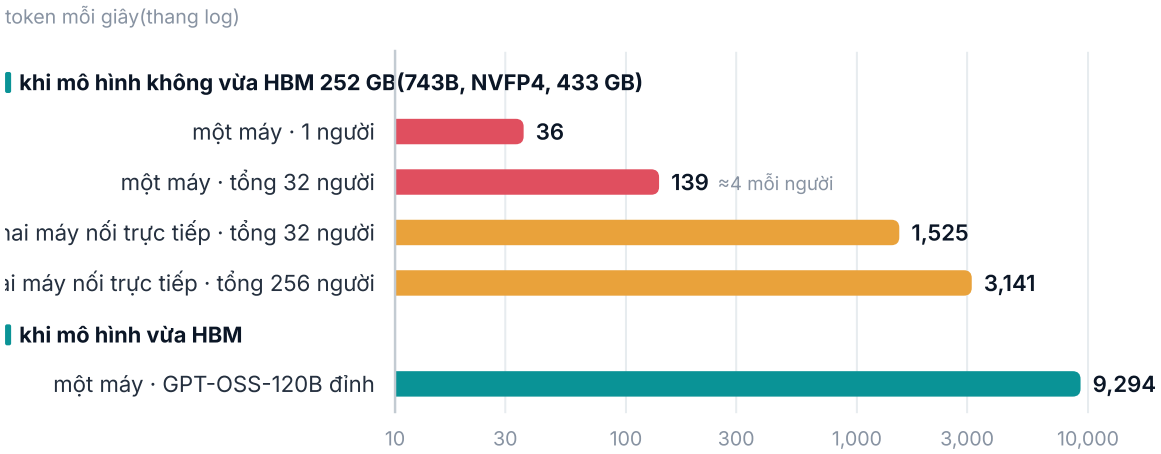
Jetson Thor làm cho triển khai cấp trạm khả thi. Các bài kiểm tra TensorRT Edge-LLM của NVIDIA cho thấy Qwen3-VL-8B (NVFP4) prefill 2,048 token trong 104.5 ms và decode 44.4 token/s trên Thor (TensorRT Edge-LLM) vendor claim; với một ảnh cắt AOI và câu trả lời 20 token, đó là khoảng 0.5 s với mô hình 8B và 0.2 s với 2B, tức 8.5–21 ảnh cắt mỗi giây khi batch estimate. Độ dài đầu ra, không phải số tham số, quyết định độ trễ, nên giải mã có ràng buộc thành một token lớp hoặc JSON ngắn là một phần của thiết kế. GR00T N1.7 (3B) chạy ở 25.1 Hz trên Thor (Jetson AI Lab) vendor claim. Một Orin NX mất khoảng 1.8 s mỗi ảnh cắt và Atlas 300I Duo có p95 trên 23 s, chỉ phù hợp cho batch (CSDN) single source. Ở đại lục, hãy dự toán sản phẩm NVIDIA cao hơn 40–60% so với giá niêm yết Mỹ estimate.

Một Mac Studio 512 GB làm cho mô hình kích thước đầy đủ trên một máy trở nên khả thi, cho vài người dùng. M5 Ultra (tháng 8 năm 2026) cung cấp tới 512 GB ở 1.2 TB/s từ khoảng $22,299 (Apple, TechRadar); DeepSeek 671B ở 4-bit (≈380–404 GB) decode 17–20 token/s trên M3 Ultra (Pinggy) verified và ước tính 25–29 trên M5 Ultra. Prompt dài là điểm yếu: một prompt truy xuất 10,000 token mất khoảng 17–70 s estimate, vượt giới hạn tương tác 10 giây.

DGX Station là máy đơn mạnh nhất, nhưng không khả dụng ở đại lục và đắt gấp năm đến sáu lần một chiếc Mac. GB300 xuất xưởng có **748 GB**, trong đó chỉ 252 GB là HBM nhanh (Wccftech, Tom's Hardware, Aug 2026) verified. Các bài kiểm tra độc lập cho thấy khoảng 9,300 token đầu ra/s trên GPT-OSS-120B, vừa HBM, nhưng một mô hình

743B tràn sang LPDDR5X chỉ cho 32 người dùng tổng 139 token/s, khoảng 4 mỗi người; hai máy liên kết đạt 1,525 (StorageReview, Aug 2026, StorageReview, Oct 2026) single source . Giá thị trường tăng lên $123–140k vào tháng 11 (CyberPowerPC, monitors.com). Ở đại lục, lựa chọn thực tế cho cùng công việc là Mac Studio, máy chủ CPU+GPU kiểu KTransformers (KTransformers) vendor test hoặc máy chủ Ascend nội địa.

FIG 09 **Một máy trạm lớn không thành trợ lý dùng chung** ***Khi mô hình không vừa bộ nhớ băng thông cao, 32 người dùng đồng thời chỉ được ≈4 tokens/s mỗi người.***



Nguồn: StorageReview(tháng 8 và 10/2026); ServeTheHome. Kiểm độc lập một nguồn

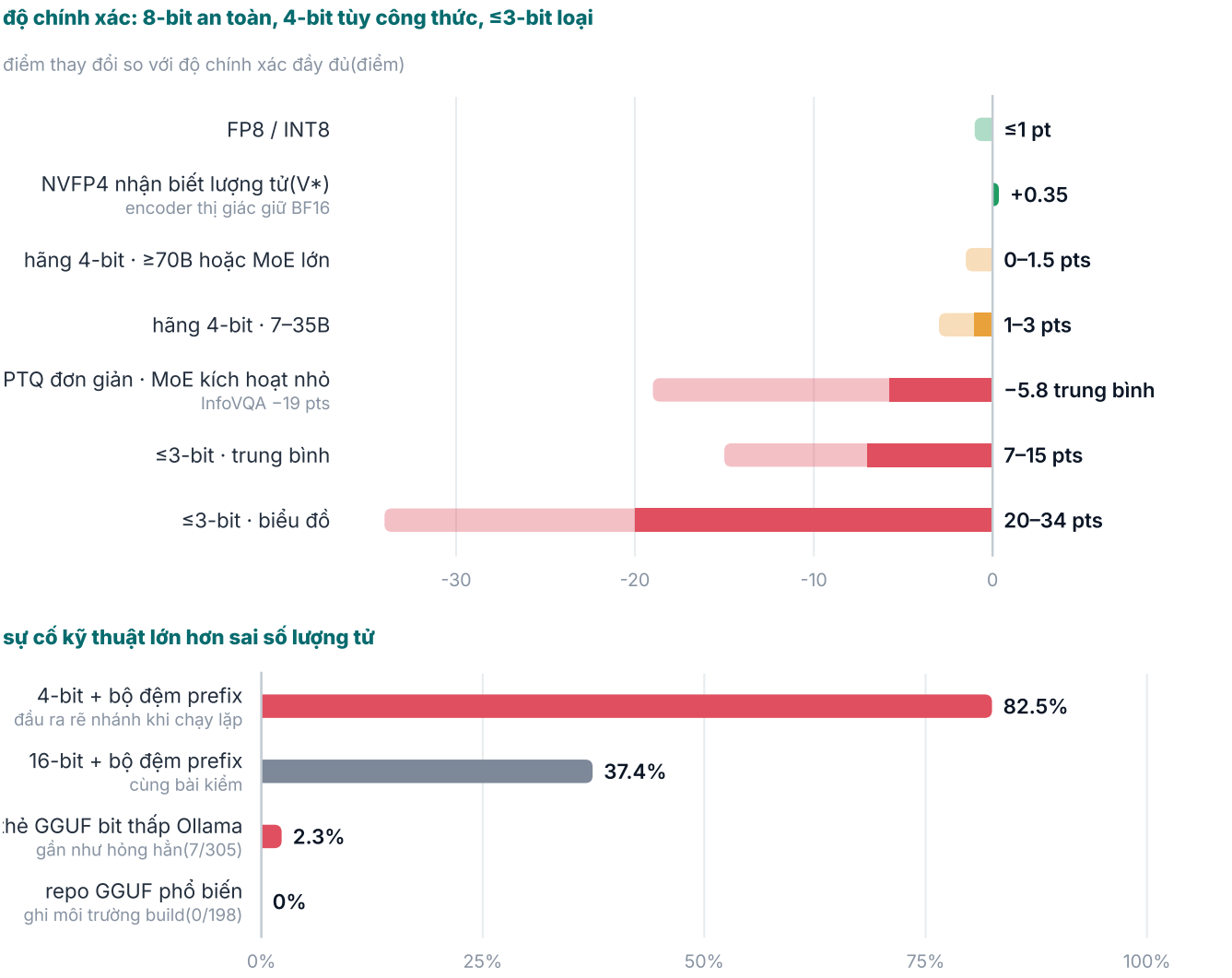
Thiếu hụt bộ nhớ đã đẩy cùng các linh kiện tăng 20–100% trong một năm (DGX Spark từ $3,999 lên $6,950 (TechRadar)), nên các báo giá thiết bị nên được lập chỉ số thay vì cố định.

Các bộ tăng tốc nội địa chạy mô hình kích thước đầy đủ; bảo trì phần mềm là nút thắt. Ascend 910B có 64 GB HBM, nhưng vLLM-Ascend đánh dấu DeepSeek V3/R1 là "không còn bảo trì" và chỉ hỗ trợ vài mô hình trên card nhỏ hơn (vLLM-Ascend) single source, primary , và mỗi mô hình mới mất một đến hai tháng để thích ứng (Caijing/Anue); hãy giao một phiên bản mô hình đã đóng băng, đã xác thực. Phục vụ DeepSeek 671B đầy đủ cho 20–100 người dùng cần ít nhất hai máy chủ Atlas 800I A2 tám card (China Telecom Cloud) verified , nhưng các gói thầu công khai cho thiết bị LLM ở Trung Quốc chỉ đạt tổng 27.5 triệu RMB trong tháng 1–tháng 9 năm 2025 (C114) single source : nhà máy trả tiền cho tích hợp và dịch vụ, không phải cho một hộp tính toán.

Lượng tử hóa: 8-bit an toàn, 4-bit tùy công thức, 3-bit trở xuống là loại

8-bit (FP8 hoặc INT8) về cơ bản không mất mát cho các mô hình ngôn ngữ-thị giác (Red Hat, Hugging Face) verified . Ở 4-bit, mất mát tùy kích thước và công thức: bản AWQ 72B của Qwen mất dưới một điểm trên MMMU trong khi bản 7B mất 2.8 (Qwen AWQ) verified , và các nhiệm vụ nặng chi tiết mất nhiều hơn (GPTQ 4-bit ngây thơ khiến Qwen3-VL-30B-A3B mất 19 điểm trên InfoVQA (VEQ)). Ở 3-bit trở xuống, mất mát trung bình là 7–15 điểm và biểu đồ mất 20–34 (VLMQ) verified ; vượt quá 64k token, 4-bit mất tới 59%, tệ hơn trên đầu vào không phải tiếng Anh (arXiv 2505.20276). Giữ bộ mã hóa thị giác ở FP16/BF16.

FIG 10 **Lượng tử hóa: 8-bit an toàn, 4-bit tùy công thức, 3-bit trở xuống loại** ***Sai sót kỹ thuật trong ngăn xếp phục vụ tốn độ chính xác hơn cả lượng tử hóa.***



Nguồn: NVIDIA Nemotron model card, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 và khác

Hai cảnh báo quan trọng hơn. Thứ nhất, **tính đến ngày 11 tháng 11 năm 2026, không nghiên cứu hay thẻ mô hình nào đo lượng tử hóa trên ảnh khuyết tật PCB/SMT**, nên hãy tái xác thực trên bộ dữ liệu vàng. Thứ hai, lỗi ngăn xếp phục vụ thường lớn hơn lỗi lượng tử hóa: cùng một mô hình đạt thấp hơn khoảng 30 điểm khi chạy dưới Ollama so với triển khai tham chiếu (SiliconBench); 7 trong 305 thẻ mô hình mã chính thức của Ollama về cơ bản hỏng dù tệp còn nguyên (arXiv 2609.05881); không repo GGUF nào trong 198 repo phổ biến ghi lại bản build của nó (arXiv 2610.19498); và với tái sử dụng prefix-cache, 82.5% các episode phân kỳ ở 4-bit so với 37.4% ở 16-bit (arXiv 2609.04748) single sources. 4-bit nhận biết lượng tử hóa của nhà cung cấp gần như không mất mát trên các nhiệm vụ chi tiết (NVIDIA Nemotron 3.5) vendor claim, nhưng VLA 4-bit mất 1.5–4.3 điểm dưới nhiễu loạn và chạy chậm hơn trên Thor (LoopQ) single source. Quản lý một mô hình lượng tử hóa như một thước đo mới: ghim hash trọng số, ghi lại bản build, đặt lại prefix caching và tái xác thực sau mọi thay đổi về độ chính xác, engine hoặc template.

Độ chính xác	≥70B dense hoặc ≥100B MoE	7–35B	≤4B hoặc nhiệm vụ nặng chi tiết	Kết luận
FP8 / INT8	An toàn (≤1 điểm)	An toàn	An toàn; mặc định cho kiểm tra	Dùng

Độ chính xác	≥70B dense hoặc ≥100B MoE	7–35B	≤4B hoặc nhiệm vụ nặng chi tiết	Kết luận
4-bit AWQ/GPTQ (công thức nhà cung cấp, encoder thị giác FP16)	0–1.5 điểm	1–3 điểm; GPTQ ngây thơ trên MoE hoạt động nhỏ −5.8 trung bình, −19 trên InfoVQA	2–3 điểm; xác thực nhiệm vụ chi tiết trước	Có điều kiện
NVFP4 (công thức nhà cung cấp)	≈0.3–1 điểm dưới FP8 (dữ liệu nhà cung cấp)	Bản nhận biết lượng tử hóa gần như không mất mát trên V*, OCRBench	Không dữ liệu công khai	Có điều kiện (ưu tiên bản nhận biết lượng tử hóa)
GGUF Q4_K_M	Chưa kiểm	Phương sai cao trên nhiệm vụ miền	Không cho văn bản nhỏ hoặc khuyết tật	Không cho phán đoán chi tiết
GGUF Q5_K_M trở lên	Được coi là an toàn	Gần như không mất mát trong hai bài kiểm	Sàn cho chi tiết trên llama.cpp	Dùng
≤3-bit (kể cả mô hình 1T ép xuống 1.8–2 bit trên một máy)	Không cho quyết định	Không	Không	Không

Cấu hình nào cho công việc nào

Cấu hình	Kiểm tra nội tuyến	Rà soát AI cảnh báo giả	Lập kế hoạch và điều khiển robot	Hỏi&Đáp và phân tích
Bộ phát hiện chuyên dụng hoặc mô hình dị thường (YOLO, RF-DETR, lớp PatchCore), INT8/FP16, bất kỳ hộp biên NPU hoặc GPU nào	● Có: thẩm phán chính, 2–12 ms mỗi ảnh cắt	● Có (bộ phân loại tinh chỉnh trên dây chuyền)	● Chỉ nhận thức	—
Mô hình thị giác 2–4B, NVFP4/AWQ/FP8 với encoder thị giác FP16, trên Thor/T4000 hoặc DGX Spark	● Có điều kiện: kiểm lại ảnh cắt bị AOI gắn cờ ở 0.2–0.3 s mỗi cái	● Có điều kiện, dưới 1 s sau khi tinh chỉnh	● Xương sống VLA (VLA 2–3B ở ≈40–130 ms mỗi đoạn)	● Yếu
Mô hình thị giác 7–9B (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 hoặc 4-bit nhà cung cấp, trên Thor, Spark hoặc card 24–32 GB	● Có điều kiện: ≈0.5 s mỗi ảnh cắt bị gắn cờ	● Có: 0.5–1 s mỗi lần gọi, giá trị tốt nhất	● Lập kế hoạch và chọn kỹ năng (1–3 s), không trong vòng điều khiển	● Hỏi&Đáp SOP cơ bản cho một hoặc vài người dùng
MoE 26–35B (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● Không (quá nặng)	● Có, như người giải thích hoặc cho ca khó	● Lập kế hoạch	● Có (8 người dùng trên Spark, ≈0.1 s token đầu)
MoE 106–241B ở 4-bit trên Mac Studio, RTX PRO 6000 hoặc Spark	● Không	● Không	● Lập kế hoạch ngoại tuyến	● Có, cho một hoặc vài người dùng
DeepSeek 671B hoặc Kimi K2.x 1T ở 4-bit trên Mac 512 GB, máy chủ KTransformers hoặc DGX Station	● Không	● Không	● Chỉ sinh chương trình ngoại tuyến	● Có điều kiện: 1–3 người dùng và prompt ngắn; prompt dài mất 16–170 s

Cấu hình	Kiểm tra nội tuyến	Rà soát AI cảnh báo giả	Lập kế hoạch và điều khiển robot	Hỏi&Đáp và phân tích
Bất cứ thứ gì ở ≤3-bit; GGUF Q4_K_M cho công việc chi tiết mô hình nhỏ; Ollama cho nhiều người dùng	● Không	● Không	● Không	● Không

Nguồn: các dữ liệu TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research và KTransformers ở trên, và (DGX Spark tests, RTX PRO 6000 tests, Hardware Corner on Atlas, Ultralytics YOLO26, RF-DETR) **cộng đồng và nhà cung cấp, chủ yếu single source** .

Vậy, với hai nỗi lo thường gặp: **không nhất thiết phải chậm** nếu mô hình khớp công việc (mô hình thị giác nhỏ và trung bình trên Thor hoặc Spark đáp ứng mọi ngân sách nội tuyến và rà soát; mô hình 671B–1T ở ngoài dây chuyền), và **lượng tử hóa không nhất thiết làm sai** với FP8/INT8 mặc định, hoặc 4-bit công thức nhà cung cấp trên 7B đã xác thực trên bộ dữ liệu vàng. Rủi ro lớn hơn là độ chính xác zero-shot công nghiệp khoảng 57 điểm của các mô hình tổng quát và lỗi ngăn xếp phục vụ. Hãy cho các yêu cầu kiểm tra kích thước batch cố định và hàng đợi riêng (temperature 0 không tái lập bit chính xác dưới batch (vLLM documentation)), giữ dự phòng nóng N+1, và làm cho mọi bước trên dây chuyền mở khi lỗi.

Thầy và trò

03 Thầy và trò: mô hình frontier có thể gán nhãn trước dữ liệu kiểm tra; với cánh tay robot người thầy thực sự vẫn là con người

Các ứng viên thầy tháng 11 năm 2026: thị giác tổng quát tăng gấp đôi, điểm khuyết tật công nghiệp còn thiếu

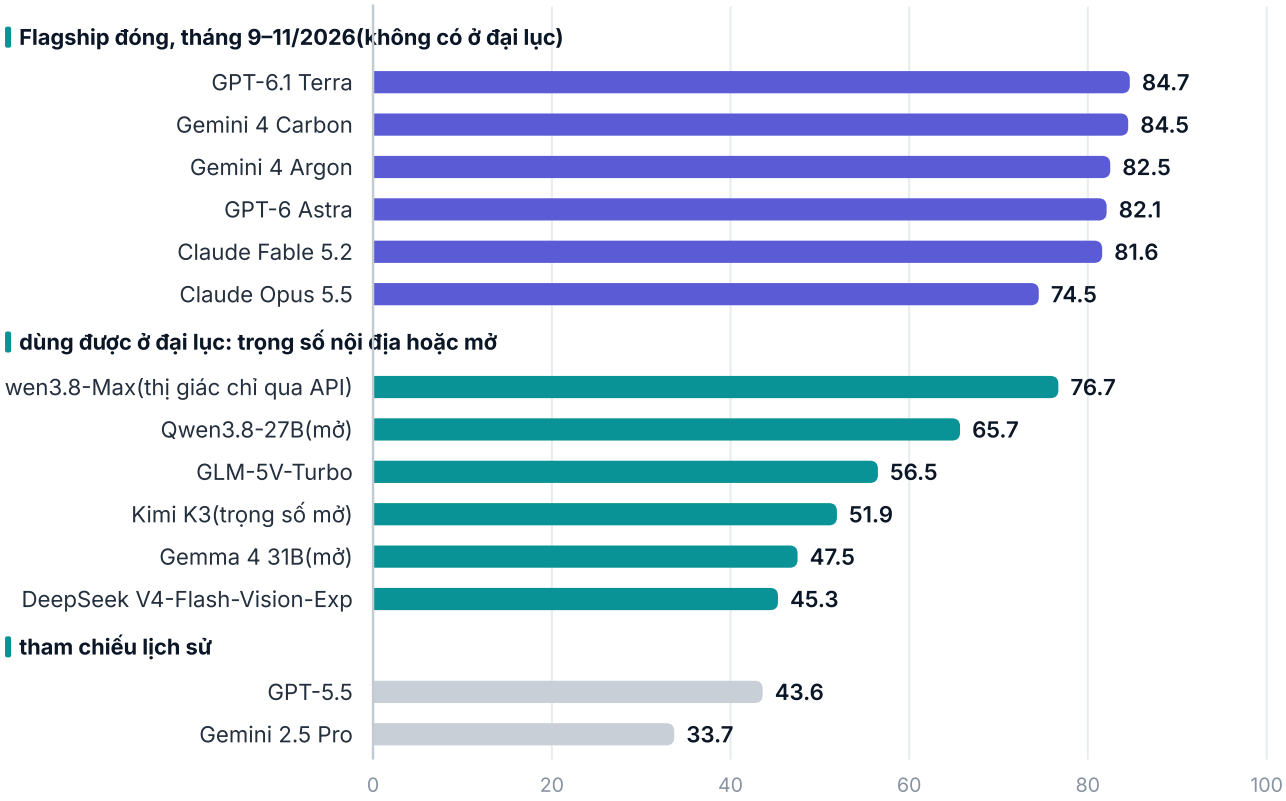
Mọi mẫu flagship đóng có thể làm thầy trong tháng 11 năm 2026 đều nhận ảnh: Claude Fable 5.2 (phổ biến rộng rãi từ ngày 9 tháng 11, $8/$40 mỗi triệu token) và Opus 5.5 ($4/$20) (Anthropic, Claude models overview) verified; GPT-6 Astra ($10/$50) và GPT-6.1 Terra ($1/$5) (OpenAI pricing); và Gemini 4 Argon và Carbon, mở theo từng giai đoạn (Google Carbon). Dùng được ở Trung Quốc đại lục là Qwen3.8-Max (thị giác qua API) và các mô hình mở như Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp và Gemma 4 (Roboflow blog, Hugging Face).

Khoảng trống then chốt phải nói rõ ràng: không mô hình nào trong số này có điểm đã công bố trên MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB hay bất kỳ bộ dữ liệu PCB/SMT nào. MMAD vẫn chưa có bảng xếp hạng (MMAD GitHub), và hai bài báo dị thường công nghiệp mới nhất (tháng 8–tháng 9 năm 2026) vẫn dùng Gemini 2.5 Pro và Claude Sonnet 4 làm đường cơ sở tổng quát mới nhất của chúng (InspectorGPT, Anomaly-LR). Một tìm kiếm toàn văn arXiv ngày 11 tháng 11 tìm thấy không kết quả cho "Fable 5.2" hay "Gemini 4 Carbon" trong ngữ cảnh này (arXiv API) single source, absence of evidence. Các thứ thay thế gần nhất:

- Mẫu frontier mới nhất có điểm MMAD là GPT-5.5: 69.95% ở "có khuyết tật hay không" và 76.20% trung bình trên bảy nhiệm vụ trên một tập con 500 ảnh, so với 95.24% và 86.65% của chuyên gia người (AnomalyClaw) single source; GPT-4o đạt 68.63% và 74.92% vào năm 2024 (MMAD, ICLR 2025), khoảng 1.3 điểm tiến bộ trong hai năm.
- Trên LDU-Bench, ảnh đánh giá SEM thực từ quang khắc IC và benchmark điện tử duy nhất có mô hình 2026, phân loại ba nhánh tốt-xấu khá ổn (macro F1 0.932 cho GPT-5.4, 0.884 cho GLM-5V-Turbo, 0.876 cho Claude Opus 4.6, 0.504 cho Gemma 4 31B), nhưng phân loại 11 hình thái khuyết tật chỉ đạt 0.20–0.41 (LDU-Bench) single source.
- Trên các đánh giá thị giác độc lập của Roboflow (cập nhật ngày 10 tháng 11 năm 2026), mAP@50 phát hiện đối tượng là 84.5 cho Gemini 4 Carbon, 84.7 cho GPT-6.1 Terra và 81.6 cho Fable 5.2, so với 33.7, 43.6 và 38.6 cho Gemini 2.5 Pro, GPT-5.5 và Claude Opus 4.8 (Roboflow Vision Evals) single source, đánh giá độc lập, bộ dữ liệu riêng.

FIG 11 **Ứng viên thầy: điểm chung tăng gấp đôi trong một thế hệ*Nhưng không flagship 2026 nào có điểm công bố trên benchmark lỗi công nghiệp.***

Roboflow phát hiện chung mAP@50



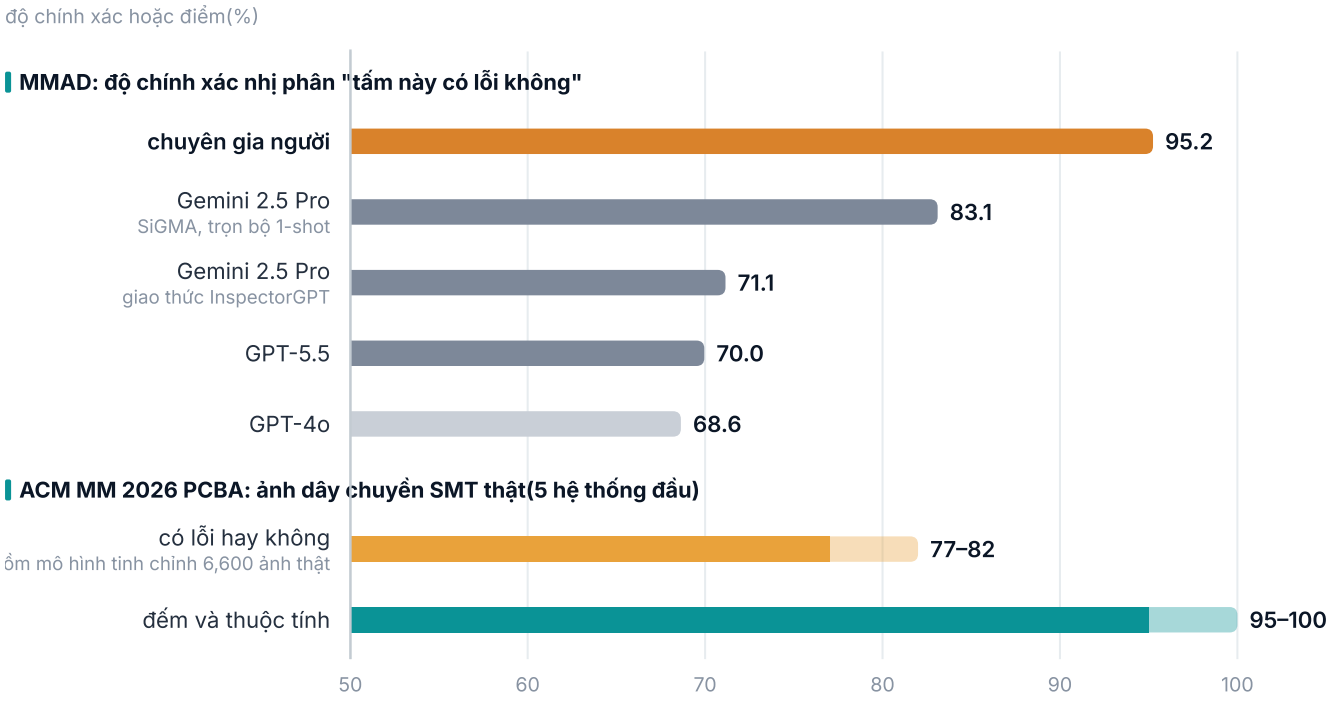
benchmark lỗi công nghiệp(MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): không điểm công bố từ flagship 2026
điểm phát hiện chung tăng gấp đôi trong một thế hệ nhưng không thay được phán đoán lỗi; phải đo trên bộ chuẩn vàng của nhà máy

Nguồn: Roboflow Vision Evals; ghi chú phát hành mô hình; MMAD, MVTec AD 2 và các trang benchmark khác(đến 11/11/2026)

Ứng viên thầy (tháng 11/2026)	Giá mỗi M token (vào/ra)	Roboflow phát hiện mAP@50 / đếm	Benchmark khuyết tật công nghiệp	Dùng được ở Trung Quốc đại lục
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Không	Không
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Không	Không
GPT-6 Astra	$10 / $50	82.1 / 80.2	Không	Không
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Không	Không
Gemini 4 Carbon	$2 / $10 (khuyến mãi)	84.5 / 90.1	Không	Không
Qwen3.8-Max (chỉ thị giác qua API)	$2 / $6	76.7 / 81.1	Không	Có
Qwen3.8-27B (mở)	Nội bộ	65.7 / 64.9	Không	Có
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench triage 0.884	Có
Gemma 4 31B (mở)	Nội bộ	47.5 / 51.4	LDU-Bench triage 0.504	Có (trọng số nội địa)
Tham chiếu: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD nhị phân 69.95% / 68.63%	Không
Chuyên gia người	—	—	MMAD nhị phân 95.24%	—

Các bài báo mới hơn vẫn thử nghiệm thế hệ trước, nhưng chỉ cùng một hướng. Trên tập MMAD đầy đủ, chuyên gia người đạt 95.2% ở hạng mục có-khuyết-tật-hay-không và Gemini 2.5 Pro 83.1% (SiGMA), nhưng cùng mô hình đó chỉ đạt 71.14% trong một bài báo khác (InspectorGPT): giao thức kiểm tra làm dịch chuyển điểm số nhiều ngang một thế hệ mô hình. Trong thử thách CVPR 2026 VAND 4.0, các mô hình ngôn ngữ-thị giác có sẵn đạt đỉnh 61.88 AP so với 90.84 của các phương pháp có giám sát, và ban tổ chức kết luận chúng "hiện chưa thể thay thế các bộ phân loại đã tinh chỉnh" (VAND 4.0) single source. Trong thử thách ACM MM 2026 PCBA, năm hệ thống hàng đầu đạt 76.92–82.47 ở hạng mục có-khuyết-tật-hay-không trong khi loại linh kiện và đếm chân đạt 95–99.5 (arXiv 2609.21276) single source: nhận diện dễ hơn nhiều so với phán đoán.

FIG 12 **"Có lỗi không?" vẫn khó với mô hình chung** ***Mô hình chung tốt nhất kém chuyên gia người hơn 12 điểm trên MMAD; hệ đã tinh chỉnh làm tốt hơn trên ảnh dây chuyền thật.***



Nguồn: MMAD; SiGMA; InspectorGPT; ACM MM 2026 PCBA challenge

Ba kết luận rút ra. Thấy và đếm tăng gấp đôi trong một thế hệ, nhưng phán đoán bất thường tiến chậm, phụ thuộc nặng vào giao thức và mang sai lệch hệ thống; các mẫu mới đạt được bao nhiêu chỉ đo được trên chính ảnh của nhà máy. Bên trong phân phối huấn luyện của chúng, các chuyên gia 7B đã tinh chỉnh vẫn thắng mô hình tổng quát (Anomaly-LR, UniPCB). Và dữ liệu dây chuyền thực làm sụp điểm số: trên các tấm điện tử đã lắp ráp, AUROC cấp ảnh rơi từ 97–98 trên MVTec AD xuống 31.5–78.6 (arXiv 2606.01992) single source. Với các công ty đại lục, thấy mạnh nhất khả dụng, Qwen3.8-Max, thấp hơn frontier toàn cầu khoảng 8 điểm.

Chạy hàng chục hay hàng trăm vòng: tiền không phải vấn đề, chất lượng nhãn mới là

Tiền không phải vấn đề. Ở giá niêm yết tháng 11 năm 2026, 1,000 ảnh một megapixel tốn khoảng $3 (GPT-6.1 Terra) đến $32 (GPT-6 Astra) mỗi vòng, và một trăm vòng trên 100,000 ảnh khoảng $32,000–135,000 trước chiết khấu batch estimate (Claude vision docs, OpenAI vision guide, Roboflow Vision Evals). Với nhãn đơn giản, một người tốn ngang Terra; lợi thế của thấy là tốc độ và quy mô, không phải giá. NDA của khách hàng thường giữ ảnh sản phẩm trong nhà máy, nên bộ dữ liệu vàng được gán nhãn tại chỗ.

Trở ngại thực sự là chất lượng, và nhiều vòng hơn không sửa được lỗi hệ thống. Roboflow chạy mỗi nhiệm vụ ba lần và điểm số chỉ dịch ±0.1–1.4 điểm (Roboflow), nên một mô hình hầu như lặp lại lỗi của nó và bỏ phiếu đa số không cứu được. Các vòng chỉ giúp khi mỗi vòng thêm thông tin: một hoặc hai ảnh tham chiếu tốt thêm khoảng 3–4 điểm, nhiều hơn thì không thêm gì (AD-Copilot); để một mô hình tranh luận với một tham chiếu tốt qua vài lượt thêm 6.23

điểm AUROC xuyên miền cho GPT-5.5 và 2.95 trên MMAD, đạt đỉnh sau khoảng ba lần gọi (AnomalyClaw) single sources. "Một học trò huấn luyện trên nhãn của thầy tiệm cận nhãn người" mới chỉ được chứng minh trên ảnh tự nhiên: một ResNet-101 huấn luyện trên nhãn Qwen3-VL của COCO và Objects365 đạt 95.7–99.4% kết quả nhãn người, và các tác giả nói dữ liệu miền chưa được kiểm chứng (arXiv 2602.20972) single source.

FIG 13 **Hai pipeline thầy–trò** ***Với kiểm tra, mô hình frontier gán nhãn trước và người quyết định; với cánh tay robot, thầy thật vẫn là demo của người.***



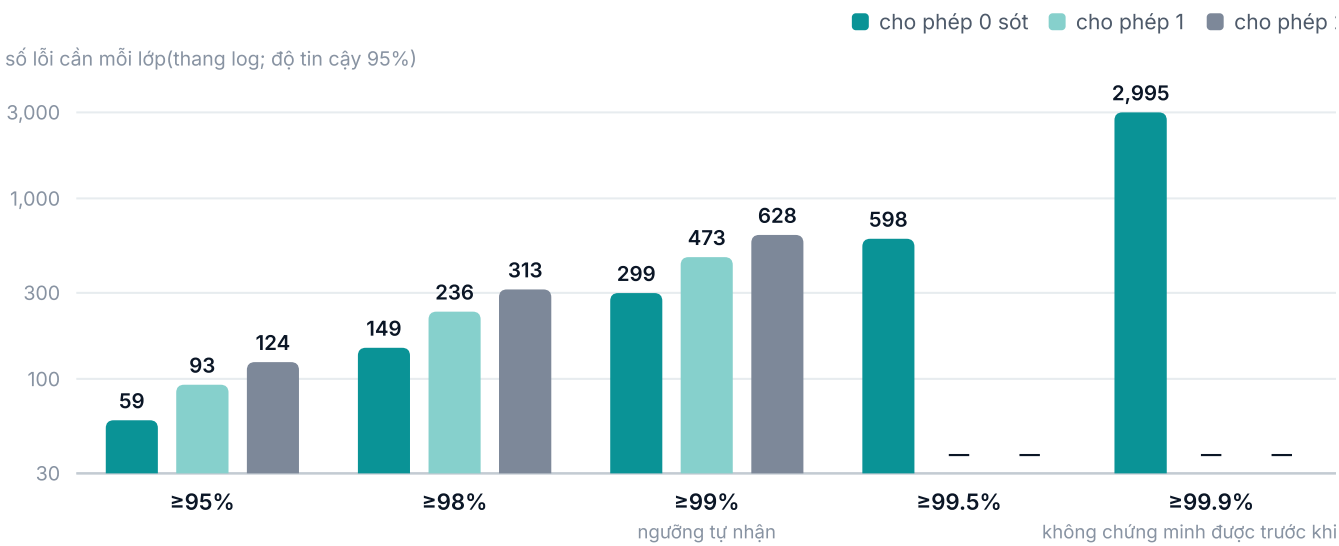
Nguồn: Tổng hợp của tác giả(Chương 3); chi phí và cỡ mẫu là ước tính

Một giao thức khả thi estimate:

1. **Thu thập bộ dữ liệu vàng từ chính kho ảnh AOI của nhà máy và các tấm mạch NG được giữ lại**, phân tầng thành 12 lớp khuyết tật thật (thiếu, lệch, tombstone, đảo cực, sai linh kiện, cầu/chập, thiếu thiếc, thừa thiếc, bi thiếc, chân nhấc/hở, linh kiện hỏng, vật lạ) và bốn nguyên nhân cảnh báo giả (bóng, flux, silkscreen, dung sai), cộng một "lát trôi" ít nhất 300 khuyết tật từ dây chuyền thứ hai hoặc chiếu sáng, màu mạch hay mã linh kiện khác.
2. **Gán nhãn mù với hai hoặc ba người kiểm tra đã đào tạo IPC-A-610**, với một kỹ sư cấp cao quyết định và X-ray, kiểm tra điện hoặc cắt ngang giải quyết các mẫu mơ hồ. Chỉ chấp nhận một định nghĩa lớp nếu mức đồng thuận giữa người kiểm tra đạt κ ≥0.75 (≥0.90 cho các lớp trọng yếu); đồng thuận của con người trên 11 hình thái khuyết tật điện tử trong LDU-Bench chỉ là 0.727, nên một số lớp sẽ cần gộp hoặc định nghĩa lại. Đồng thuận thuộc tính kiểu AIAG (hiệu quả ≥90%, bỏ sót ≤2%, báo động giả ≤5%, κ >0.75; >0.90 để phát hành) là mức tối thiểu hợp lý (Fabrico, iFactory).
3. **Định cỡ bằng cận 95% một phía nhị thức chính xác (Clopper–Pearson)** (binomial confidence intervals, rule of three:

Recall cần chứng minh	0 lần bỏ sót cho phép	1 cho phép	2 cho phép
≥95%	59 khuyết tật	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **Bộ chuẩn vàng phải lớn cỡ nào** ***Chứng minh recall 99% không sót cần 299 lỗi mỗi lớp; 99.9% không chứng minh được trước khi mua.***



Nguồn: Khoảng tin cậy nhị thức chính xác(95%)

Sáu mươi khuyết tật mỗi lớp là đủ cho một màn lọc đạt/không đạt; 300 khuyết tật là cần để quyết định chấp nhận tự động. Tập đầy đủ, khoảng 3,600 khuyết tật cộng 3,500 ảnh cắt cảnh báo giả, mất khoảng 37–57 giờ công để gán nhãn và phân xử, và chạy một ứng viên thầy trên đó tốn $23–196; chi phí thực là hai đến sáu tuần thu thập, do lớp hiếm nhất quyết định estimate. Hãy gửi các ảnh cắt AOI đơn linh kiện, không phải cả tấm mạch, với mục tiêu rộng ít nhất 42–48 pixel (arXiv 2610.09313).

1. **Chỉ tự động chấp nhận nhãn của thầy cho một lớp có không lần bỏ sót trên ≥299 khuyết tật vàng**, đồng thuận mô hình-thật không tệ hơn đồng thuận người-người, và ít nhất 50% cảnh báo giả được khép lại. Các lớp trọng yếu như chập mạch, thiếu linh kiện, đảo cực và sai linh kiện cần 99.5–99.9% (598–2,995 khuyết tật); nếu không, bất cứ thứ gì AI không thể khép lại với độ tin cậy cao đều chuyển cho người. Tỷ lệ lọt lưới dưới 0.1% khuyết tật không thể chứng minh trước khi mua (2,995 khuyết tật với không lần bỏ sót; 0.01% sẽ cần 29,956), nên dây chuyền giữ lấy mẫu thủ công các ca AI đã khép lại, ICT/FCT hạ nguồn làm lưới chặn và SPC trên tỷ lệ lọt lưới. Các nhãn đã xác thực sau đó được chưng cất vào một bộ phát hiện hoặc phân loại truyền thống và tái xác thực trên mọi sản phẩm mới.

Điều khoản và quy tắc dữ liệu cũng quyết định người thầy. Anthropic cấm dùng dịch vụ của họ để huấn luyện mô hình cạnh tranh (Anthropic commercial terms); OpenAI cho phép dùng đầu ra để huấn luyện bộ phân loại nội bộ không bán cho bên thứ ba (OpenAI business terms); DeepSeek cho phép chưng cất một cách rõ ràng (DeepSeek terms) primary terms. Không bên nào định nghĩa "mô hình cạnh tranh", nên dùng thương mại cần rà soát pháp lý, và quy tắc dữ liệu của khách hàng thường giữ ảnh tại chỗ. Mặc định an toàn nhất là một thầy trọng số mở chạy nội bộ, chỉ dùng mô hình frontier đám mây khi khách hàng cho phép bằng văn bản.

Học trò kiểm tra: nhỏ và nhanh; nút thắt là quang học, dữ liệu và xác thực

Học trò không phải nút thắt: YOLO26n mất 3.80 ms mỗi ảnh ở INT8 trên một Orin Nano Super $249–399 (Ultralytics) vendor test, NVIDIA báo cáo 98.51% trên ba lớp khuyết tật PCB sau khi thích ứng một mô hình nền trên khoảng 600 ảnh (NVIDIA) vendor claim, và khuyết tật tổng hợp chỉ thêm 0.9–2.5 điểm mAP (arXiv 2605.04635). Các ràng buộc là quang học, dữ liệu dây chuyền, xác thực và giấy phép: Ultralytics YOLO là AGPL-3.0 hoặc bản quyền doanh nghiệp (Ultralytics), nên sản phẩm đóng cần giấy phép hoặc một bộ phát hiện Apache-2.0 như RF-DETR.

Học trò robot: trạm có cấu trúc tiệm cận cấp sản xuất; công việc linh hoạt và động thì không

Trong tháng 11 năm 2026 có ba loại thầy robot, và chỉ một loại có thể trở thành học trò nội bộ: các VLA mở, tinh chỉnh được (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Các mô hình robot đóng (π0.6/π0.7 qua API beta, Gemini Robotics 2 cho người thử nghiệm tin cậy) không thể triển khai tại chỗ (The AI Insider, DeepMind). Các mô hình tổng quát frontier chậm và kém ở thao tác tinh: trong bài kiểm tra của Artificial Analysis, chúng đạt 46–69 với 28–45% thành công cầm nắm và khoảng 30 phút mỗi nhiệm vụ, "quá chậm và đắt cho bất kỳ mục đích sản xuất nào" (AA-Robotics); GPT-6 Astra chỉ cắm được khe chính xác 2 lần trong 20 (RoboCurve); và điều khiển trực tiếp khớp đạt 0–5.5% trên LIBERO ở 0.2–0.4 Hz so với khoảng 83 Hz cần thiết (Anthropic) verified trên ba nguồn. Con số 64.9% của Fable 5.2 trên benchmark mô phỏng RoboDojo là một lần chạy nội bộ (Anthropic) vendor claim.

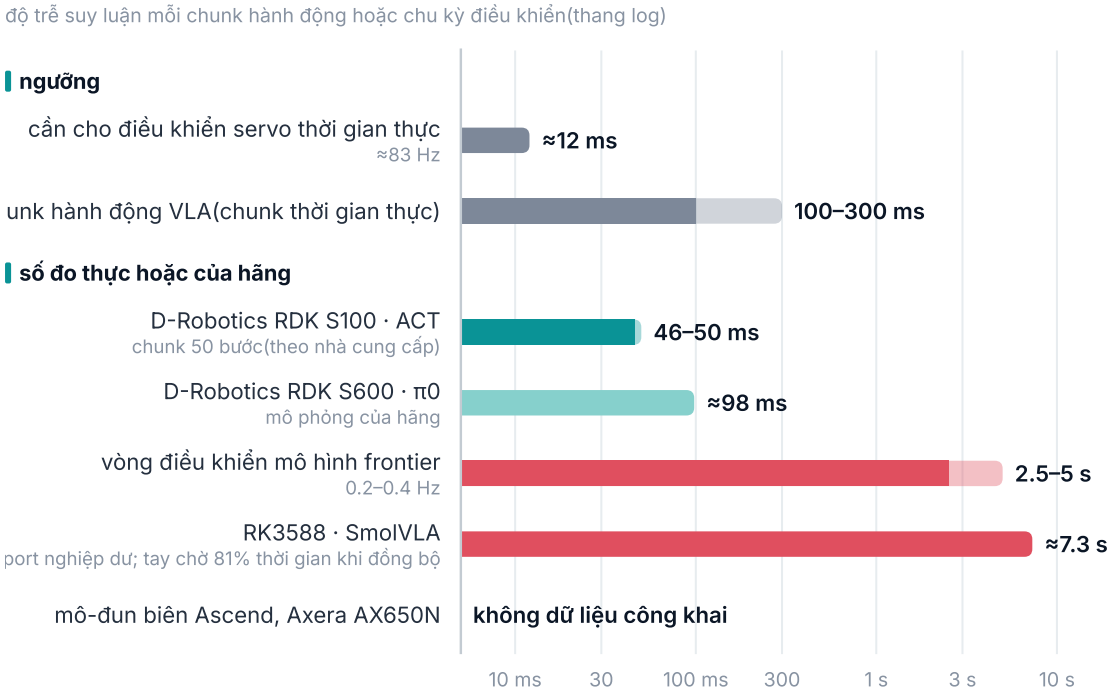
Các nghiên cứu bên thứ ba từ tháng 10 đồng thuận: khoảng 33–44% trên LIBERO-10 cho ba tác tử dẫn đầu (arXiv 2610.30051); các tác tử tuyên bố sai thành công trên 24–39% nhiệm vụ con (arXiv 2611.08209); và 104 trong 108 nhiệm vụ robot bán tĩnh trên robot thật cho Opus 5.5, nhưng ở 280 giây mỗi nhiệm vụ và không có cắm hay lắp ráp (AMP1) single sources. Các mô hình frontier thuộc về ngoại tuyến: viết chương trình, điều phối nhiệm vụ và phác thảo tiêu chí thành công.

Chất lượng học trò có một lợi thế rõ: tốt bên trong phân phối huấn luyện, sụp đổ bên ngoài nó. π0.7 "thường vượt 90%" trên các nhiệm vụ đã thấy và đạt 60–80% trên nhiệm vụ chưa thấy (π0.7 paper) vendor research; ContactDP đạt 97.8% trên đầu nối đã huấn luyện nhưng 28.6–66.7% trên đầu nối chưa thấy (ContactDP); thay đổi nhẹ camera hoặc tư thế làm rơi 95% xuống dưới 30% trong LIBERO-Plus (LIBERO-Plus, CVPR 2026); cầm nắm học được từ băng chuyền đang chuyển động đạt đỉnh 58.89% (ForeTime-VLA) single sources, một công việc mà theo dõi bằng thị giác cổ điển đã làm trong sản xuất nhiều năm.

Bằng chứng gần-sản-xuất đều đến từ trạm có cấu trúc, trình diễn của con người và hiệu chỉnh robot thật. Neuromeka giữ máy trạng thái và chuyển động dạy trước, chỉ học cầm nắm, cắm cáp và hàn, mỗi thứ từ dưới 20 phút dữ liệu: 322 trong 324 trong sản xuất, chậm hơn người 12.8% (Neuromeka) single source, vendor. Fraunhofer đạt 92.8% trên bốn đầu nối từ 68 phút trình diễn (Fraunhofer); HIL-SERL của Berkeley thành công 100 trong 100 lần cắm RAM, SSD và USB sau 1–2.5 giờ RL robot thật (HIL-SERL); AgiBot chạy 18–20 giây mỗi bước trong takt 20 giây trên một dây chuyền kiểm tra máy tính bảng (IThome) vendor claim.

Chưng cất mua được tốc độ, không mua được khái quát hóa: π0.5 chạy ở 0.70 Hz trên Jetson Orin thô và 10.30 Hz sau lập lịch bất đồng bộ (EagleVLA); chưng cất lớp đưa nó từ khoảng 3 Hz lên khoảng 10 Hz (Shallow-π) verified. Một học trò chỉ thừa hưởng những gì thầy bao phủ; mỗi sản phẩm mới vẫn cần dữ liệu nhiệm vụ.

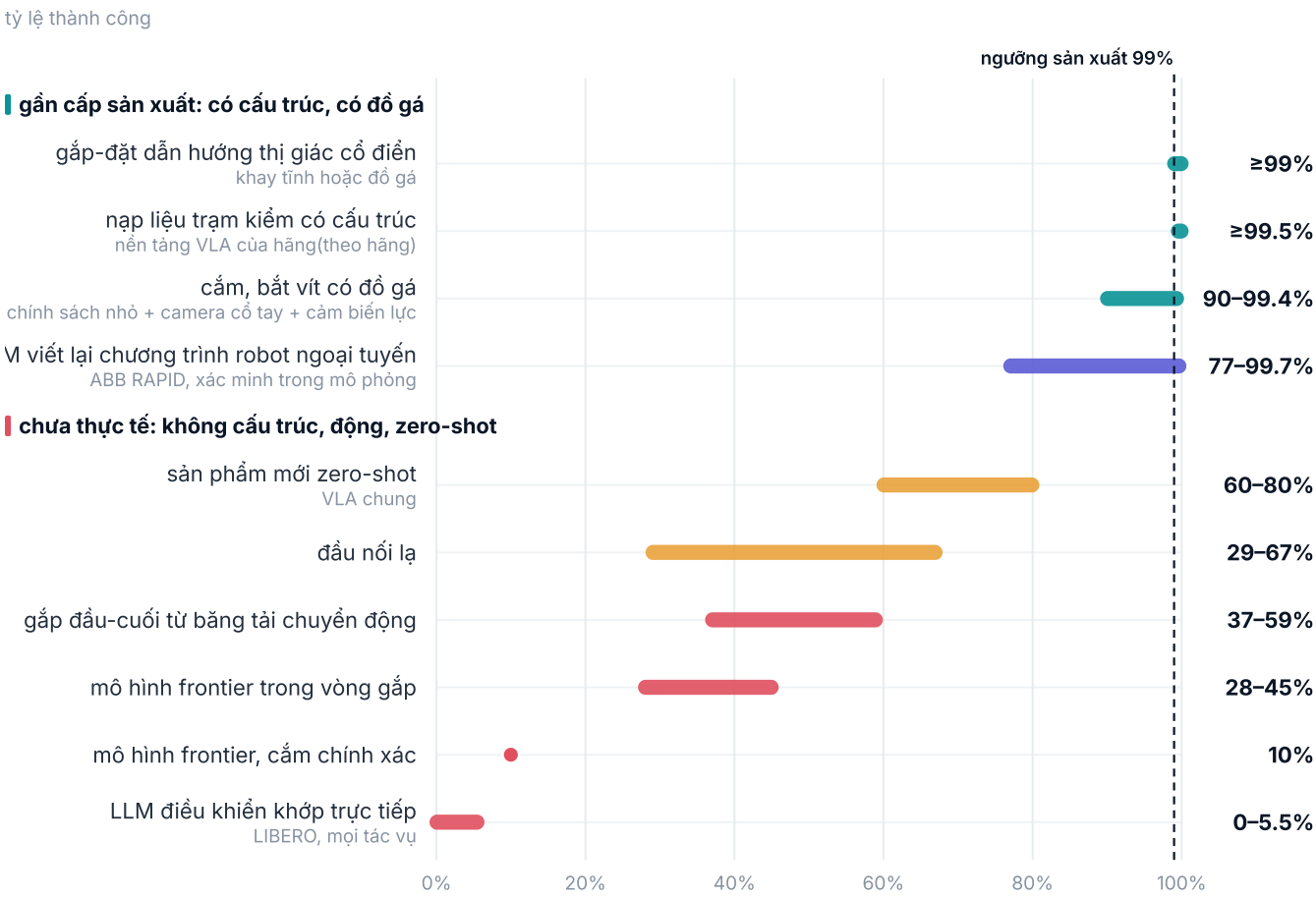
FIG 15 **Độ trễ chính sách biên** ***Chính sách ACT nhỏ chạy trong ≈50 ms trên bo biên; VLA port sang chip nghiệp dư mất 7.3 s.***



Nguồn: Tài liệu và diễn đàn D-Robotics(số của hãng); port cộng đồng GitHub; nghiên cứu Anthropic

Trên chip biên Trung Quốc, chính sách học được duy nhất được đo là một chính sách ACT trên RDK S100 của D-Robotics ở 46–50 ms mỗi đoạn 50 bước, không công bố tỷ lệ thành công (D-Robotics RDK S docs) vendor claim ; một bản port nghiệp dư của SmolVLA trên RK3588 mất khoảng 7.3 s mỗi đoạn (GitHub). Lượng tử hóa không mua tốc độ miễn phí: không bản π0.5 4-bit nào nhanh hơn 16-bit ở batch size 1 (LoopQ). Trình diễn thì đắt (Reuters báo cáo khoảng 300 lần thử mỗi chuyển động dùng được cho một người vận hành từ xa mới (Notebookcheck)), và xác thực tốn nhiều hơn: chứng minh tỷ lệ lỗi ≤0.1% ở độ tin cậy 95% mất khoảng 3,000 lần thành công liên tiếp, 17–22 giờ robot estimate . Không chính sách thao tác học được nào có chứng nhận an toàn.

FIG 16 **Robot trò: gần cấp sản xuất chỉ khi trạm có cấu trúc** ***Trạm có đồ gá đạt 90–99%+; việc không cấu trúc, động và zero-shot chưa thực tế.***



Nguồn: Bài báo và tài liệu hãng trích trong Chương 3; một số số liệu theo hãng

Nhiệm vụ điện tử	Học trò khả thi	Thành công kỳ vọng	Tốc độ so với người	Chi phí mỗi lần chuyển đổi	Kết luận 2026–2028
Pick-and-place nhiều chủng loại, kitting, phục vụ máy (khay hoặc đồ gá tĩnh)	Nhận thức học được cộng chuyển động được lập trình hoặc lập kế hoạch	≥99% với dẫn hướng thị giác cổ điển	Như trạm truyền thống	Vài giờ	Khả thi ngay
Đầu nối cố định, cắm FFC và cáp, bắt vít, hàn	Máy trạng thái + chính sách BC/diffusion/ACT nhỏ + camera cổ tay + lực/mô-men (lý tưởng là cảm ứng) + hiệu chỉnh người	90–99.4%	Chậm hơn 10–15%	10–300 lần trình diễn (≈15–70 phút) + 1,000–3,000 lần chạy xác thực (6–22 giờ robot)	Khả thi, nặng R&D vài tháng cho trạm đầu tiên
Phục vụ trạm kiểm tra có cấu trúc và phân loại OK/NG	Nền tảng VLA của nhà cung cấp + bản sao số + RL robot thật	≥99.5% vendor claim, mâu thuẫn	18–20 s mỗi bước, trong takt 20 s	≤4 h huấn luyện lại + 15 phút hiệu chuẩn	Khả thi trên nền tảng nhà cung cấp
Lập trình chuyển đổi	LLM viết hoặc sửa chương trình robot ngoại tuyến, xác thực trong mô phỏng	77–99.7% trên viết lại ABB RAPID	Không ảnh hưởng takt	Phút đến giờ mỗi chương trình	Khả thi như công cụ kỹ thuật

Nhiệm vụ điện tử	Học trò khả thi	Thành công kỳ vọng	Tốc độ so với người	Chi phí mỗi lần chuyển đổi	Kết luận 2026–2028
Cầm nắm học đầu-cuối từ băng chuyền đang chuyển động	Học trò VLA	37–59%; ≤37% trên băng nhanh	—	≥458 lần trình diễn mỗi thiết lập	**Không trước 2028; dùng theo dõi cổ điển cộng nhận thức học được**
Sản phẩm mới không dữ liệu mới (zero-shot)	VLA tổng quát	60–80%; đầu nối chưa thấy 29–67%	—	—	**Không thực tế**
LLM frontier trong vòng lặp điều khiển cánh tay	Tác tử GPT-6, Claude, Gemini	Cầm nắm 28–45%; cắm chính xác 10%	≈30 phút hoặc 280 s mỗi nhiệm vụ so với ≈20 s	—	**Không thực tế**
Bó dây định tuyến tự do và lắp ráp linh hoạt nhiều bước	—	Không bằng chứng sản xuất 2026	—	—	**Không trước 2028**

Nguồn: các nghiên cứu ở trên và (University of Passau on RAPID rewriting, GraspVLA, Komax) **chủ yếu single source và dữ liệu nhà cung cấp** . Trong 2026–2028, giá trị của một mô hình lớn với robot là như một prior đã huấn luyện trước và một lập trình viên ngoại tuyến, không phải thứ "dạy" một học trò. Mọi ca gần-sản-xuất tôi tìm thấy (Neuromeka, AgiBot tại Longcheer, Fraunhofer, ContactDP) đều dựa vào trình diễn của con người, cảm biến lực hoặc cảm ứng, mô phỏng hoặc bản sao số và RL có giám sát. Không ca nào dựa vào nhãn từ một mô hình frontier.

Chi phí và hoàn vốn

04 Chi phí và hoàn vốn: rà soát AI hoàn vốn trong vài tháng, phục vụ máy trong vài năm, cánh tay gắp bỏ thì không bao giờ

Chi phí lao động: Mỹ cao gấp khoảng 5.5 lần Thâm Quyến, Đài Loan khoảng 1.5 lần, phần lớn Đông Nam Á dưới một nửa

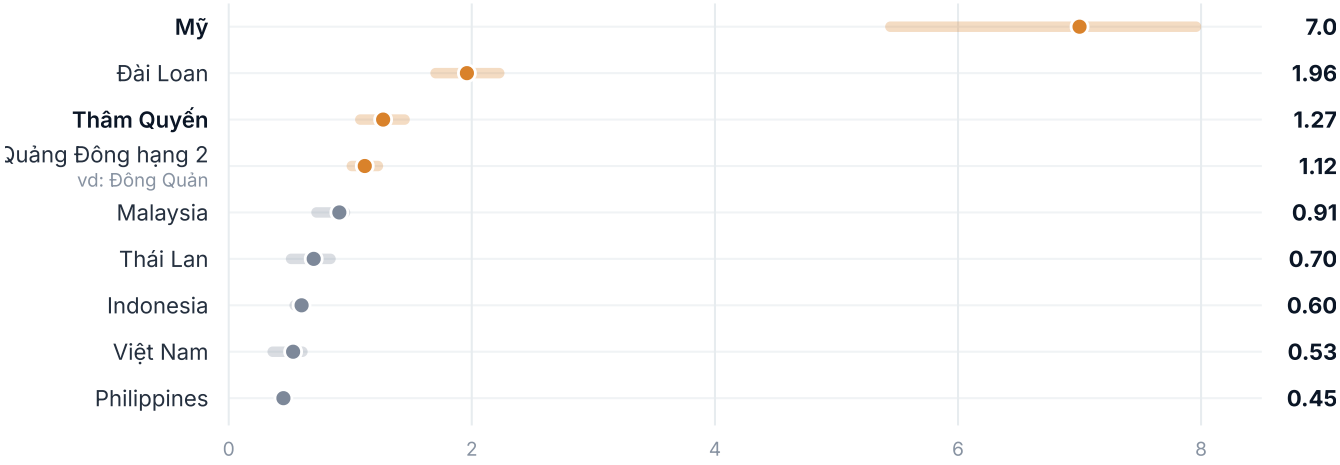
Lợi tức khi thay lao động bắt đầu từ chi phí của lao động đó. Hướng dẫn lương chính thức của Thâm Quyến đặt trung vị 2024 cho một người kiểm tra chất lượng sản xuất ở mức **74,074 RMB** một năm; lương chiếm 86.92% tổng chi phí lao động trong sản xuất điện tử, nên chi phí đầy đủ là khoảng 85,000 RMB. Khảo sát JETRO 2025 đặt chi phí hằng năm của một công nhân nhà máy Trung Quốc ở $12,667, khoảng 85,100 RMB, chênh trong 0.2% (Shenzhen wage guidance, JETRO 2025) verified .

Khu vực	Chi phí đầy đủ hằng năm của một người kiểm tra hoặc xử lý (thấp / cơ sở / cao)	Căn cứ
Trung Quốc đại lục, Thâm Quyến	70k / **85k** / 100k RMB ($10.4k / **12.7k** / 14.9k)	Trung vị chính thức × 1.15; khớp JETRO verified
Vùng lương bậc 2 Quảng Đông (Đông Quản v.v.)	65k / **75k** / 85k RMB	Trung vị Quảng Đông 2023 × 1.15 cộng hai năm tăng lương estimate
Đồng bằng Dương Tử (Tô Châu, Côn Sơn, Vô Tích)	Như Thâm Quyến	Không căn cứ cho lao động kiểm tra rẻ hơn; lương tối thiểu cao nhất Giang Tô cao hơn Thâm Quyến khoảng 5.6% estimate
Đài Loan	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	Khảo sát lương nghề của Bộ Lao động cộng ≈1.18 phụ phí luật định; JETRO là đầu cao estimate
Malaysia	$6.8k / **9.1k** / 10.0k	JETRO và lương tối thiểu estimate
Thái Lan	$4.7k / **7.0k** / 8.8k	Tương tự
Indonesia	$5.0k / **6.0k** / 6.3k	Tương tự
Việt Nam	$3.2k / **5.3k** / 6.5k	Tương tự
Philippines	$4.0k / **4.5k** / 4.9k	Tương tự
Hoa Kỳ	$54k / **70k** / 80k	Trung vị người kiểm tra BLS tháng 5/2025 $48,570 ÷ lương ở 65.5% tổng thù lao estimate

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

FIG 17 **Chi phí năm của một công nhân kiểm hoặc vận chuyển *Trọn gói, Mỹ ≈5.5× Thâm Quyến và Đài Loan ≈1.5×.***

chi phí trọn gói năm mỗi người QC/vận chuyển($10k; chấm=chuẩn, thanh=thấp–cao)



Nguồn: Hướng dẫn lương chính thức Thâm Quyến, Bộ Lao động Đài Loan, JETRO, BLS; ước tính của tác giả

Ba hiệu chỉnh đặt căn cứ. Thứ nhất, Thâm Quyến đã nâng sàn đóng góp hưu trí từ 4,775 lên 5,521 RMB một tháng vào tháng 7 năm 2026 (Shenzhen Bendibao) single source, nên một người kiểm tra Thâm Quyến năm 2026 nhiều khả năng tốn 80–90k RMB. Thứ hai, một vị trí không phải một người: hai ca 12 giờ cần khoảng 2.0 người (sáu ngày một tuần) đến 2.4 (cả năm) ở Đồng bằng Châu Giang, khoảng 2.7–3.1 ở Đông Nam Á dưới trần làm thêm giờ, và 4.2–4.5 cho 24×7 ở Mỹ và Đài Loan estimate. Rà soát AI thường cho phép một người rà soát phụ trách hai đến bốn dây chuyền, nên tiết kiệm thực tế là **0.5–1.5 người toàn thời gian mỗi dây chuyền, không phải một người mỗi dây chuyền mỗi ca**. Thứ ba, người mua ở Đài Loan, Việt Nam và Malaysia không chỉ đuổi theo chi phí: trong khảo sát của JETRO, 61.1% doanh nghiệp ở Đài Loan, 48.2% ở Việt Nam và 44.9% ở Malaysia nói tuyển dụng đã khó hơn trong hai năm qua, so với 18.6% ở Trung Quốc đại lục (JETRO 2025) single source.

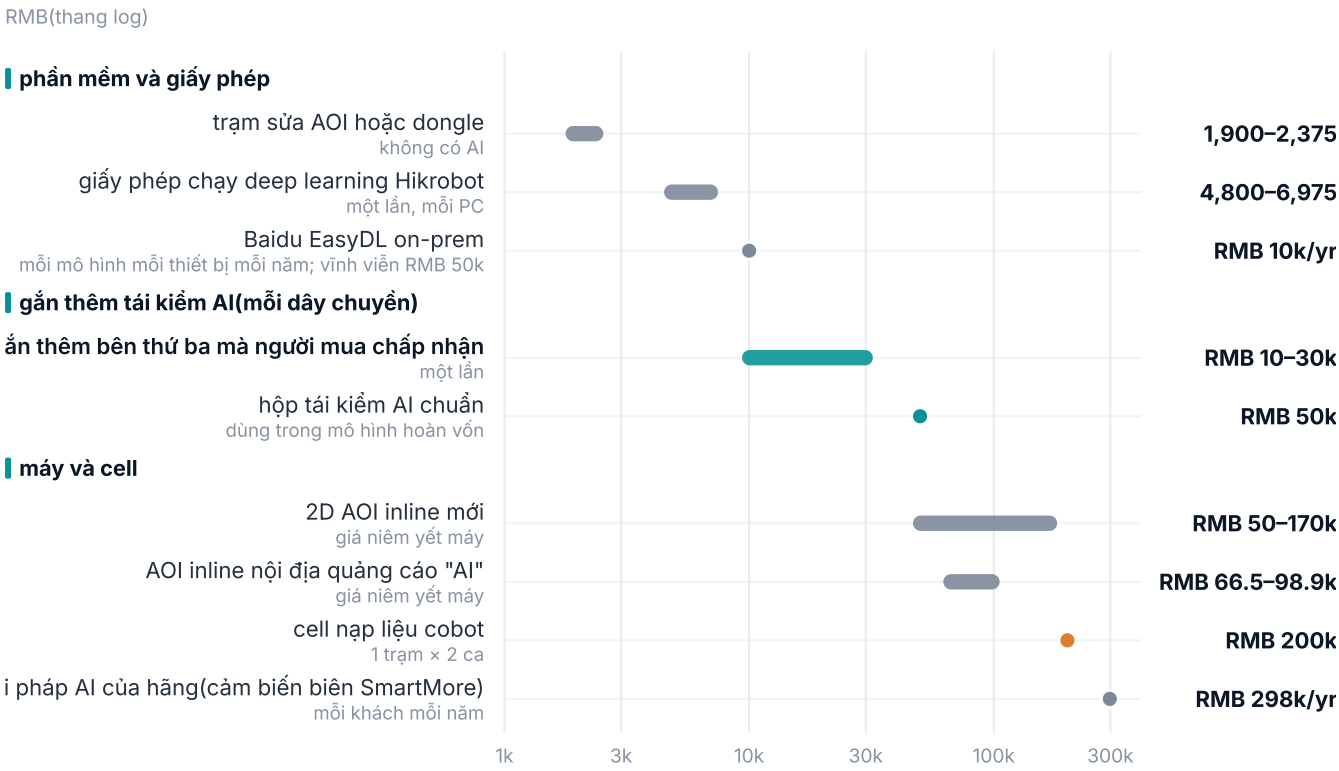
Giá thiết bị: bốn căn cứ không nên lấy trung bình

Giá AOI khác nhau chủ yếu vì chúng đo những thứ khác nhau. Niêm yết trực tuyến (tháng 11 năm 2026, lẫn mới, cũ và cho thuê) chạy khoảng 17–80k RMB cho 2D ngoại tuyến hoặc để bàn, 50–170k RMB cho 2D nội tuyến và 80–475k RMB cho 3D nội tuyến, lên tới 700k RMB cho 3D Hàn Quốc (1688 inline AOI, 1688 offline AOI, 1688 3D AOI) single source, list prices. Một gói thầu tháng 11 năm 2025 chốt trần một AOI 3D cộng một SPI 3D ở 800k RMB (tender notice). Bản cáo bạch của Sinic-Tek ngụ ý 428–512k RMB mỗi AOI 3D trong 2019–2022, khoảng 340–470k RMB theo báo cáo 2025 (Sinic-Tek prospectus, Sinic-Tek 2025 annual report) primary + estimate. Một nhà cung cấp nước ngoài ước tính $3,000–20,000 cho 2D để bàn, $30–70k 2D nội tuyến, $50–90k 3D nhập môn và $100–200k+ 3D cao cấp, với cùng máy nội tuyến ở $30–50k tại Trung Quốc (Averroes) vendor estimate từ một người bán phần mềm AI cạnh tranh.

Bản thân rà soát AI không có giá RMB công khai; nhà cung cấp báo giá từng ca sau một đánh giá tại chỗ miễn phí (Sohu on Hengtianyi), và các tìm kiếm niêm yết thương mại điện tử, gói thầu, bài viết nhà cung cấp và truyền thông Đài Loan không tìm thấy báo giá công khai (1688, 51SMT). Các mốc neo: khoảng €27k mỗi dây chuyền ngụ ý từ ca Siemens estimate; $21,924 giá niêm yết cho một module mạng rà soát từ xa của nhà phân phối Mỹ (Manncorp) vendor list price; khoảng 298k RMB mỗi khách hàng mỗi năm cho cảm biến AI biên của SmartMore năm 2025 (36Kr) single source. Một AOI 2D nội tuyến nội địa bán là "AI" niêm yết ở 66.5k RMB (1688 inline AOI), và một bộ đệm NG để tái kiểm tra thủ công khoảng 50k RMB (1688 AOI review). Vậy trần cho một tiện ích rà soát bổ sung của bên thứ ba là khoảng 10–30k RMB mỗi dây chuyền, hoặc một phí hằng năm tương tự; trên mức đó, người mua thà thay máy bằng một AOI nội địa đã tích hợp AI estimate.

Mốc neo	Giá (RMB)	Mua được gì	Nguồn
Trạm sửa chữa AOI hoặc khóa phần mềm (TRI, Zhenhuaxing cũ)	≈1,900–2,375	Khóa giấy phép, không AI	1688 list price
Runtime thị giác học sâu tổng quát (khóa Hikrobot VisionMaster)	≈4,800–6,975 một lần mỗi PC	Nền tảng; khách hàng tự xây và bảo trì mô hình	1688 list price
Triển khai tại chỗ Baidu EasyDL	10,000 mỗi mô hình mỗi thiết bị mỗi năm, hoặc 50,000 vĩnh viễn	Dịch vụ suy luận đóng gói	Baidu AI docs [trang giá 2022]
AOI nội tuyến nội địa bán là "AI"	66,500–98,900 mỗi máy	Thay máy	1688 list price
Cảm biến AI biên SmartMore	≈298,000 mỗi khách hàng mỗi năm	Giải pháp nhà cung cấp trên vài máy	36Kr single source

FIG 18 **Bậc thang giá** ***Từ vài nghìn RMB giấy phép phần mềm đến RMB 200–300k cho cell robot hoặc đề xuất AI của hãng.***



Nguồn: Danh sách 1688, tài liệu Baidu AI(bảng giá 2022), 36Kr, phỏng vấn ngành(chưa xác minh)

Một phí hằng năm kiểu Baidu, với một máy chủ suy luận cho một đến bốn dây chuyền và một đến bốn mô hình mỗi cái, ngụ ý một thuê bao 2.5–40k RMB mỗi dây chuyền mỗi năm, thực tế khoảng 5–10k RMB; ở 20% giá giấy phép vĩnh viễn, thuê bao và mua đứt hòa vốn ở năm thứ năm, một cấu trúc đáng sao chép estimate. Người mua quen với phí phần mềm gần bằng không: toàn bộ doanh thu "khác" của Sinic-Tek (phụ tùng, dịch vụ, cho thuê, phụ kiện) là 7.65% doanh số 2025, khoảng 2,300–3,100 RMB mỗi máy đã lắp mỗi năm (Sinic-Tek 2025 annual report) primary + estimate. Một thuê bao AI 5–10k RMB mỗi dây chuyền mỗi năm gấp hai đến năm lần số tiền người mua hiện chi cho phần mềm AOI, nên nó phải được biện minh bằng số giờ rà soát nó giải phóng (khoảng 75–90k RMB mỗi người toàn thời gian một năm), không phải bằng so sánh với phí phần mềm hiện có.

Với cánh tay robot, giá cobot trung bình của Trung Quốc giảm xuống khoảng 58,600 RMB vào năm 2025 (GGII) estimate, và giá trung bình sáu trục của Dobot giảm 19.1% trong năm đó khi các mẫu rẻ hơn giành thị phần (Huxiu) verified, dù giá niêm yết của cùng các mẫu tăng 5–15% trong 2026 (Sina Finance). Tại Mỹ, một FANUC CRX-10iA

niêm yết ở $51,434 (Vention) và các tổng hợp đặt một UR7e ở $30–48k (cobotcost) estimate. Cánh tay chỉ là 40–50% của một cụm Mỹ (Standard Bots); tại Trung Quốc, camera, kẹp, an toàn và kỹ thuật dễ chạy gấp hai đến bốn lần một cánh tay 40k RMB, nên hãy lập ngân sách chi phí kỹ thuật tuyệt đối thay vì bội số estimate.

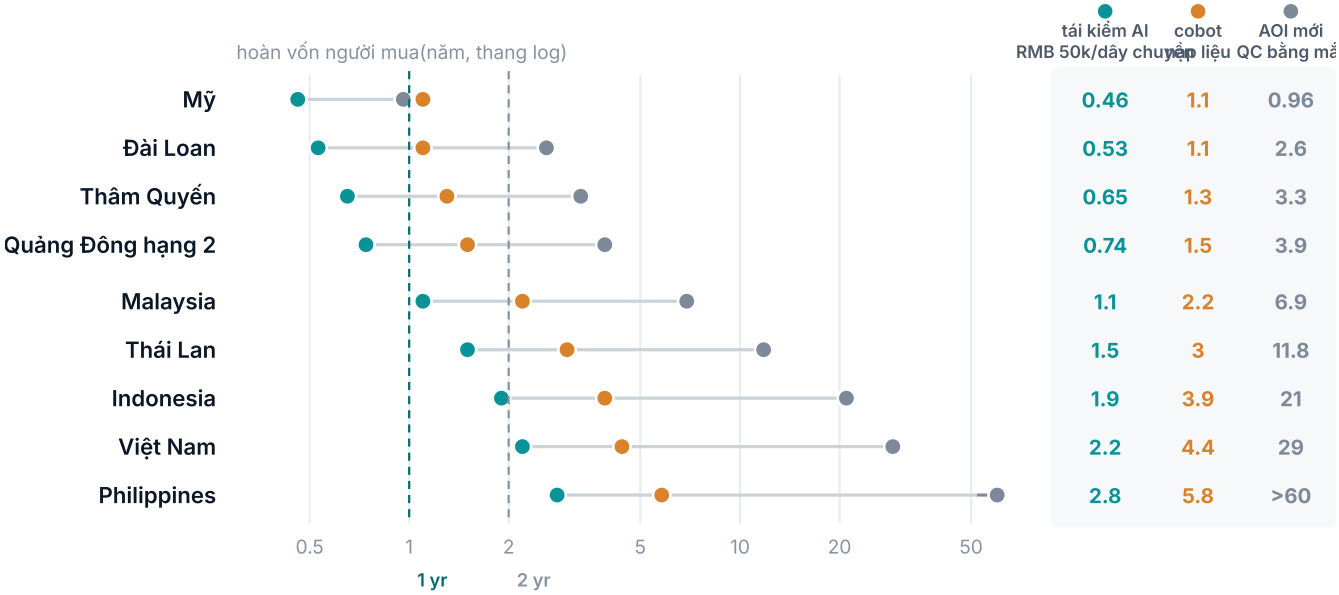
Hoàn vốn: rà soát trong vài tháng, phục vụ trong vài năm, gắp bỏ không bao giờ

Hoàn vốn đơn giản = đầu tư ÷ (người tiết kiệm được × chi phí lao động đầy đủ − chi phí vận hành hằng năm), cho một dây chuyền hai ca không trợ cấp hay thuế quan. Thiết bị Đài Loan tính giá gấp 1.3 lần Trung Quốc, Việt Nam, Malaysia và Thái Lan gấp 1.15, Indonesia và Philippines gấp 1.2, và Mỹ theo giá nước ngoài estimate.

Kịch bản	Đầu tư (Trung Quốc / Mỹ)	Người tiết kiệm	Chi phí vận hành hằng năm
(a) Hộp rà soát AI hoặc hộp camera AI	50k RMB / $30k mỗi dây chuyền	1.0 mỗi dây chuyền	15%
(b) AOI 2D nội tuyến mới thay kiểm tra thị giác thủ công (còn 0.5 người rà soát mỗi ca)	200k RMB / $60k	1.0	12%
(c) Cụm nạp cobot (một trạm × hai ca)	200k RMB / $140k	2.0	8%
(d) Cụm phát hiện-và-gắp tích hợp (kiểm tra 1.0 + gắp 0.07)	300k RMB / $200k	1.07	8%
(e) Một cánh tay robot chỉ để gắp ra các tấm mạch NG	150k RMB / $100k	0.07	8%

Khu vực	(a) Rà soát AI	(b) AOI mới	(c) Cụm cobot	(d) Phát hiện-và-gắp	(e) Cánh tay gắp bỏ
Thâm Quyến	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)	Không bao giờ
Quảng Đông bậc 2	0.74	3.9	1.5	5.3	Không bao giờ
Đài Loan	0.53	2.6	1.1	3.6	Không bao giờ
Malaysia	1.1	6.9	2.2	9.2	Không bao giờ
Thái Lan	1.5	11.8	3.0	15	Không bao giờ
Indonesia	1.9	21	3.9	25	Không bao giờ
Việt Nam	2.2 (1.6–4.5)	29	4.4	34	Không bao giờ
Philippines	2.8	>60 (chi phí vận hành ≈ tiết kiệm)	5.8	≈100	Không bao giờ
Hoa Kỳ	0.46	**0.96**	**1.1**	3.4	Không bao giờ

FIG 19 **Hoàn vốn theo vùng** ***Tái kiểm AI hoàn vốn trong vài tháng ở Mỹ, Đài Loan và Thâm Quyến và 1–3 năm ở Đông Nam Á; nạp liệu cobot tính bằng năm.***



cánh tay chỉ để nhặt tấm lỗi: không hoàn vốn ở mọi vùng
khi tỷ lệ đạt lần đầu trên 90%, nhặt chỉ chiếm 0.003–0.14 FTE mỗi ca

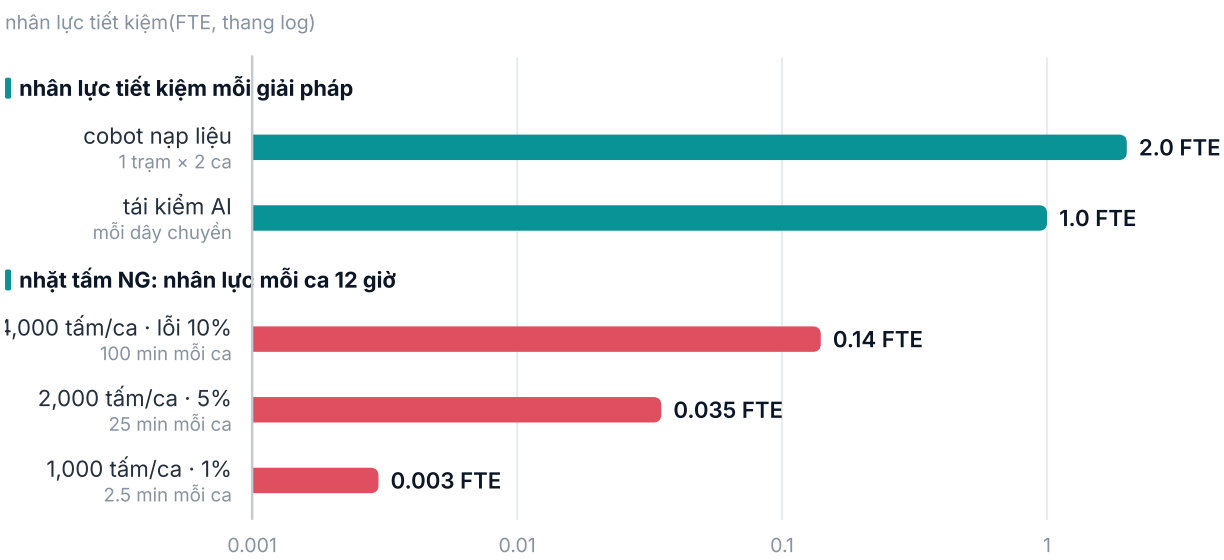
Nguồn: Mô hình tác giả: hộp tái kiểm AI RMB 50k/dây chuyền; cell nạp liệu cobot RMB 200k hoặc $140k, hai ca

Mọi giá trị đều là ước tính từ các đầu vào ở trên. Độ nhạy: giảm một nửa số người tiết kiệm thì mọi thời gian hoàn vốn gần như gấp đôi; chạy một ca thì (b)–(d) gấp đôi; khấu hao 100% năm đầu ở Mỹ rút ngắn hoàn vốn khoảng 20%, trong khi thuế quan với thiết bị Trung Quốc, Nhật Bản hoặc Hàn Quốc kéo dài khoảng 10–40%; trợ cấp địa phương 15% ở Trung Quốc rút ngắn khoảng 15%. Tôi không tìm thấy khảo sát độc lập nào về ngưỡng hoàn vốn của người mua; nhà cung cấp nói AOI hoàn vốn "thường trong 9–18 tháng" (Averroes) và ca Siemens tuyên bố tám tháng vendor claims. Tôi dùng một và hai năm làm mức sàng lọc. Mô hình bỏ qua lợi ích chất lượng (bảo hành, hàng trả) và chi phí nghỉ việc và đào tạo, nên nó là thận trọng.

Phần phản trực giác của "phát hiện và gắp" là lao động nằm ở rà soát, không phải ở việc lấy vài tấm mạch NG thật ra khỏi dây chuyền. Ở 15 giây xử lý mỗi tấm NG estimate:

Tấm mỗi ca	Tỷ lệ khuyết tật	Thời gian gắp mỗi ca 12 giờ	Người
1,000	1%	2.5 phút	0.003
2,000	5%	25 phút	0.035
4,000	10%	100 phút	0.14

FIG 20 **Vì sao nhặt tấm lỗi không hoàn vốn** ***Trên 90% tỷ lệ đạt lần đầu, nhặt chỉ chiếm 0.003–0.14 người mỗi ca.***



Nguồn: Ước tính của tác giả

Trường hợp cơ sở (2,000 tấm mỗi ca, 5% khuyết tật, hai ca) là 0.07 người, đáng giá khoảng 6,000 RMB một năm ở Thâm Quyến, thấp hơn chi phí vận hành hằng năm khoảng 12,000 RMB của cánh tay, nên nó không bao giờ hoàn vốn ở bất kỳ đâu; ngay cả trường hợp tệ nhất (4,000 tấm mỗi ca, 10% khuyết tật) cũng mất khoảng 13 năm ở Thâm Quyến và 9 năm ở Mỹ estimate. Nếu nỗi lo là dừng dây chuyền để gắp, một bộ tách NG hoặc băng tải đệm tiêu chuẩn giải quyết mà không cần cánh tay. Người trong ngành nói "trên 90% tỷ lệ đạt ngay lần đầu, hãy dừng dây chuyền và lấy tấm mạch ra bằng tay", và **điều đó đúng cho việc gắp, nhưng không đúng cho kiểm tra**. Thiết kế sản phẩm trên một dây chuyền SMT do đó là "kiểm tra cộng đánh dấu hoặc tách dòng", với cánh tay robot dành cho những nơi mọi linh kiện đều được xử lý (nạp đồ gá kiểm tra, xử lý khay và băng đạn, đóng gói), nơi kiểm tra trở thành một tiện ích bổ sung gần như miễn phí và kịch bản (c) áp dụng.

Trợ cấp, thuế và thuế quan dịch chuyển hoàn vốn 15–25%, nhưng không đổi thứ hạng

Trung Quốc đại lục: hỗ trợ chủ yếu hạ chi phí tài chính; máy đơn hiếm khi đủ điều kiện trực tiếp. Một trợ cấp lãi vay quốc gia trả 1.5 điểm lãi suất cho tối đa hai năm trong 14 chuỗi ưu tiên (Yicai) verified, khoảng 9,000 RMB trên một máy 300k RMB; thưởng nâng cấp tới 20% của Quảng Đông cần một dự án đã đăng ký với ít nhất 7 triệu RMB thiết bị mới và loại trừ Thâm Quyến (Guangdong rules), và Đông Quản thực chi 4.42% vào tháng 7 năm 2026 (Dongguan allocation) single source, primary. Một thiết bị kiểm tra AI chỉ được phê duyệt là thiết bị "có phần mềm đi kèm" bên trong dự án đã đăng ký của người mua, nên trường hợp cơ sở giả định không trợ cấp.

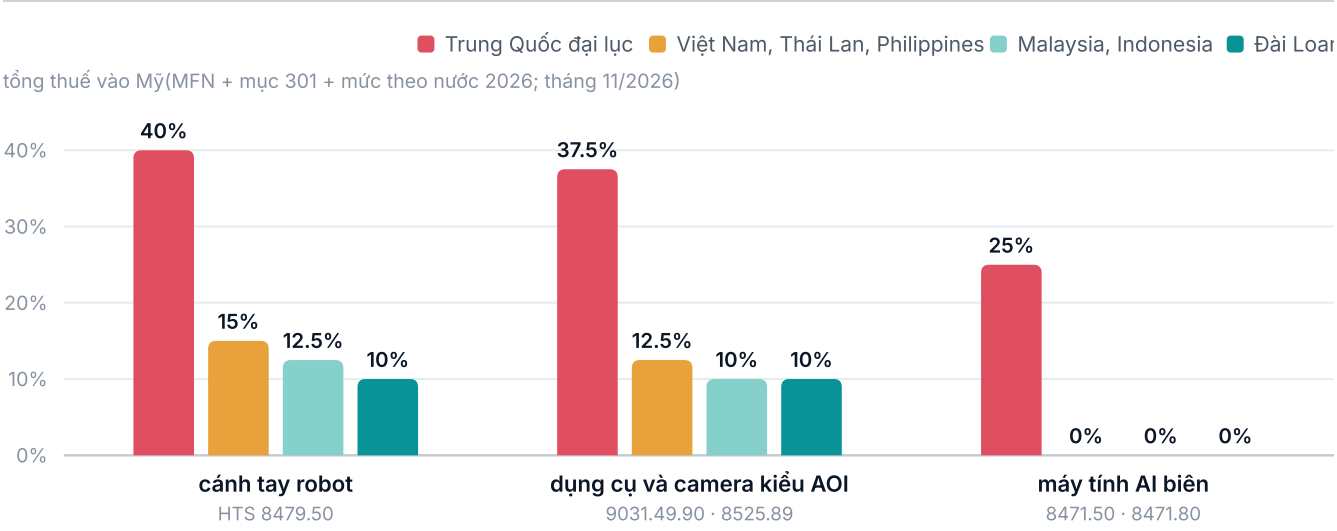
Đài Loan: ưu đãi dành cho người mua. Theo điều 10-1 của Đạo luật Đổi mới Công nghiệp, sửa đổi tháng 5 năm 2025 để bao gồm AI, người mua có thể khấu trừ 5% chi tiêu 2025–2029 cho máy móc thông minh mới hoặc AI (bao gồm kiểm tra dựa CNN/DNN) vào thuế thu nhập năm đó, hoặc 3% mỗi năm trong ba năm (Statute, credit rules) verified; các khoản tài trợ phía nhà cung cấp đóng với vốn đại lục (MOEA) single source.

Đông Nam Á: chỉ Malaysia và Thái Lan thường cho việc mua tự động hóa. Phụ cấp vốn tự động hóa 200% của Malaysia trên 10 triệu RM đầu tiên đáng giá khoảng 24% capex cho một người mua có lãi (MIDA, PwC Malaysia) verified; BOI của Thái Lan miễn thuế ba năm tới 50–100% khoản đầu tư nâng cấp, trần cao hơn khi có nhà cung cấp tự động hóa Thái Lan tham gia (BOI) single source, primary, nên vào Thái Lan tốt nhất là với một nhà tích hợp địa phương. Việt Nam, Indonesia và Philippines thường cho việc xây nhà máy hơn là mua tự động hóa (Vietnam Law Magazine, PwC Philippines).

Hoa Kỳ: khấu hao rút ngắn hoàn vốn; thuế quan kéo dài. Tài sản mua sau ngày 19 tháng 1 năm 2025 được khấu hao 100% vĩnh viễn năm đầu, khoảng 21% lá chắn thuế liên bang cho người mua có lãi (IRS Notice 2026-11) single source, primary ; điều tra Section 232 về robot và máy móc công nghiệp kết thúc ngày 28 tháng 10 năm 2026 bằng đàm phán thay vì thuế mới, sẽ đánh giá lại trong 180 ngày (White House, Mayer Brown) verified . Lắp ráp cuối tại Đông Nam Á cắt phụ phí trên cánh tay và thiết bị dạng AOI từ khoảng 37.5–40% xuống 10–15% (Chương 1).

Khu vực	Công cụ chính	Ảnh hưởng đến hoàn vốn của người mua estimate
Trung Quốc đại lục	Trợ cấp lãi vay 1.5 điểm tối đa hai năm; thưởng nâng cấp cấp tỉnh và thành phố (Quảng Đông cần dự án đăng ký ≥7 triệu RMB; Đông Quản thực chi 4.42%); biện pháp Thâm Quyến 2026–2029 (mức chưa công khai)	Khoảng 0–5% cho máy đơn; 4–15% trong dự án lớn
Đài Loan	Khấu trừ Đạo luật §10-1, 5% hoặc 3% × 3 năm	Ngắn hơn khoảng 5%, nếu người mua nộp đủ thuế
Malaysia	Phụ cấp vốn tự động hóa 200% (10 triệu RM đầu)	Ngắn hơn khoảng 15–25% cho người mua có lãi
Thái Lan	Miễn thuế ba năm BOI tới 50–100% khoản đầu tư	Ngắn hơn khoảng 15–25% cho người mua có lãi ngoài kỳ miễn thuế
Việt Nam, Indonesia, Philippines	Ưu đãi thuế thu nhập cho xây nhà máy; khấu hao thiết bị thêm 20% ở Philippines	Gần như không có cho mua tự động hóa đơn lẻ
Hoa Kỳ	Khấu hao 100% năm đầu; thuế tổng hợp ≈37.5–40% trên thiết bị sản xuất tại đại lục	Khấu hao ngắn hơn ≈20%; thuế quan dài hơn ≈10–40%

FIG 21 **Thuế Mỹ với robot, AOI và máy tính biên** ***Tay sản xuất ở Trung Quốc đại lục chịu ≈40%; xuất từ Đông Nam Á hoặc Đài Loan 10–15%.***



Nguồn: USITC biểu thuế chương 99; Federal Register; USTR(tháng 11/2026)

Theo khu vực: ở Trung Quốc đại lục, một hộp rà soát AI hoàn vốn trong vài tháng và một cụm cobot trong 1.3–1.5 năm, trong khi một AOI mới mua chỉ để cắt lao động mất 3–4 năm và phát hiện-và-gắp tích hợp 4.5–5.3, cả hai đều vượt mức sàng lọc hai năm. Đài Loan làm tốt hơn Thâm Quyến đôi chút (0.53 năm cho rà soát, 1.1 cho cụm cobot), và tuyển dụng ở đó khó hơn. Ở Malaysia và Thái Lan, một hộp AI mất 1–1.5 năm và một cụm cobot 2–3, nhanh hơn khoảng 15–25% với ưu đãi thuế địa phương cho người mua có lãi. Ở Indonesia, Việt Nam và Philippines, chỉ một hộp AI rất rẻ mới hoàn vốn trong 2–3 năm; nhu cầu ở đó đến nhiều hơn từ yêu cầu chất lượng của khách hàng và thiếu

hụt lao động. Ở Mỹ, mọi lựa chọn trừ cánh tay gắp bỏ thuần túy đều hoàn vốn trong 0.2–3.4 năm, với thuế quan, đánh giá lại Section 232 và chứng nhận NRTL là các biến chính.

05

Thị trường và cạnh tranh

Thị trường, cạnh tranh và cách người mua thanh toán

Quy mô thị trường: các con số công bố chênh nhau nhiều lần; chỉ đếm từ dưới lên là hữu ích

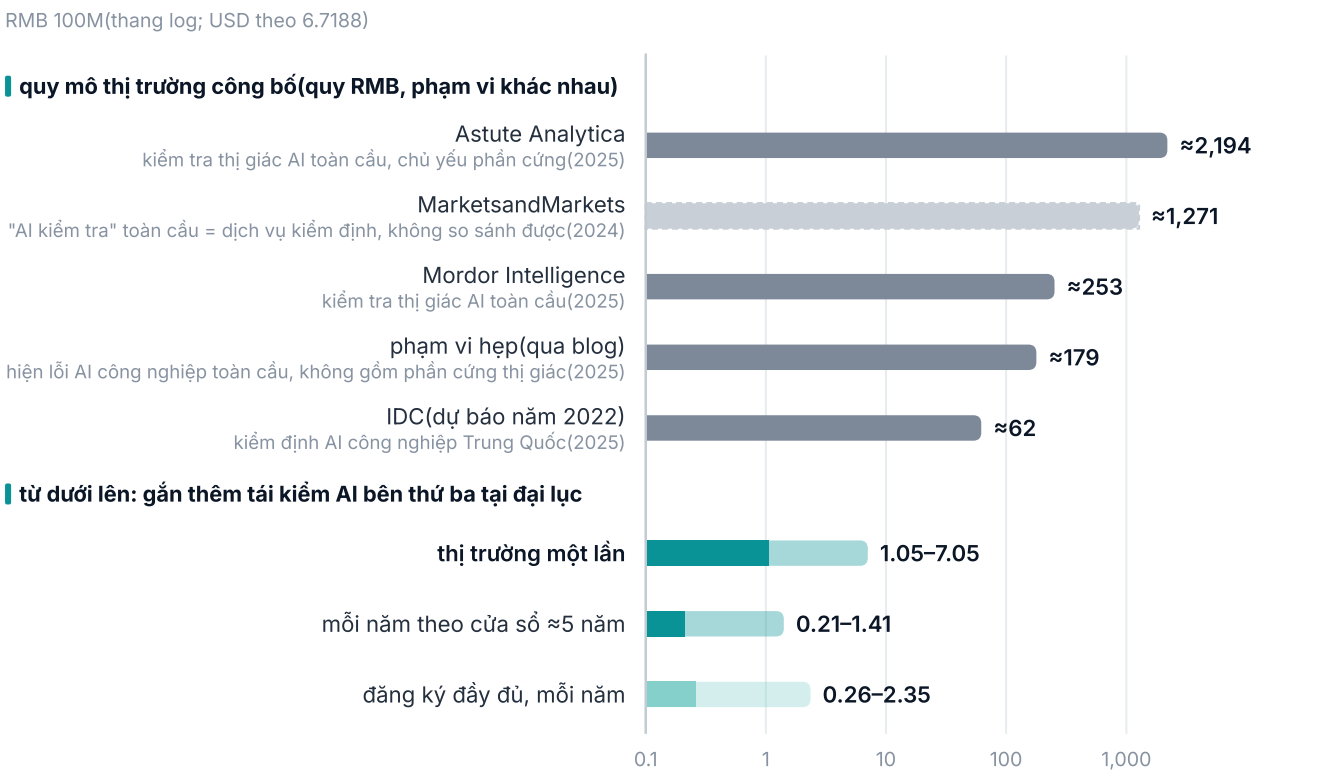
Mọi con số thị trường phải được đọc cùng phạm vi của nó. Thị trường thị giác máy của Trung Quốc là 18.15 tỷ RMB vào năm 2024 (Zhiyan) verified, với điện tử 3C là người mua lớn nhất; các ước tính toàn cầu về thị giác máy chênh nhau khoảng ba lần và về AOI khoảng 6.5 lần (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Tính từ dưới lên từ doanh thu của Koh Young, TRI, Jutze, Sinic-Tek và ViTrox, thiết bị kiểm tra SMT toàn cầu ở mức $1–2 tỷ estimate.

Không có thị trường độc lập cho "phần mềm kiểm tra thị giác AI". Các con số công bố chênh nhau hơn mười lần và đều bao gồm phần cứng hoặc dịch vụ:

Phạm vi	Đơn vị công bố	Con số
"Phát hiện khuyết tật công nghiệp AI" toàn cầu (hẹp, loại trừ phần cứng thị giác)	Công ty không nêu tên, qua blog iFactory	$2.66 tỷ năm 2025, CAGR 8.6% đến 2035, điện tử 34%
"Kiểm tra thị giác AI" toàn cầu	Mordor Intelligence	$3.76 tỷ năm 2025, $12.8 tỷ năm 2031, CAGR 22.67%
"Kiểm tra thị giác AI" toàn cầu (chủ yếu phần cứng)	Astute Analytica	≈$32.66 tỷ năm 2025
"Kiểm tra AI" toàn cầu (dịch vụ kiểm thử và chứng nhận; không so sánh được)	MarketsandMarkets	$18.92 tỷ năm 2024
"Kiểm tra chất lượng AI công nghiệp" Trung Quốc	IDC, dự báo năm 2022	$958 triệu năm 2025 (≈6.2 tỷ RMB), 91.5% trong truyền thông, điện tử và hai ngành khác

Nguồn: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) **mỗi nguồn single source, chủ yếu là tóm tắt báo cáo trả phí**.

FIG 22 **Quy mô thị trường chênh nhau hơn mười lần** ***Chỉ phạm vi hẹp và đếm từ dưới lên mới hữu ích cho quyết định.***



Nguồn: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC(2022), iFactory; ước tính từ dưới lên của tác giả

Tính từ dưới lên, chỉ có một đếm công khai về dây chuyền SMT: New Venture Research ước tính khoảng 73,231 dây chuyền toàn cầu năm 2020, 47% trong số đó ở Trung Quốc (19% năm 2000), khoảng 70% cao cấp và 30% cấp thấp (EMSNow, NVR SMT2020 summary) single source, 2020 ; NVR nay thuộc IPC và dữ liệu mới hơn là trả phí (NVR). Tăng trưởng 3–5% một năm với tỷ trọng Trung Quốc 40–47% cho khoảng 87,000–98,000 dây chuyền toàn cầu năm 2026 và 35,000–46,000 ở Trung Quốc đại lục (cơ sở 40,000) estimate .

Đầu vào	Thấp	Cơ sở	Cao	Căn cứ
Dây chuyền SMT Trung Quốc đại lục (2026)	35,000	40,000	46,000	Cơ sở NVR 2020 × tăng trưởng × tỷ trọng
Tỷ trọng chưa có rà soát AI	75%	80%	85%	Cửa sổ nâng cấp, bên dưới
Tỷ trọng bên thứ ba chạm tới được (không nội bộ, không bị ràng buộc với AI OEM, trả được tiền)	40%	50%	60%	Giả định: ≈30% dây chuyền cấp thấp cộng một số dây chuyền cao cấp đa thương hiệu, loại trừ tự xây kiểu DBG
Dây chuyền có thể phục vụ	**10,500**	**16,000**	**23,500**	Tích
Giá một lần mỗi dây chuyền	10k RMB	20k RMB	30k RMB	Bậc thang giá và trần (Chương 4)
Thị trường một lần có thể phục vụ (cơ sở lắp đặt)	**105 triệu RMB**	**320 triệu RMB**	**705 triệu RMB**	Tích
Mỗi năm trong cửa sổ ≈5 năm	21 triệu RMB	64 triệu RMB	141 triệu RMB	Cửa sổ thu hẹp 8–12% một năm
Thuê bao (2.5k / 5k / 10k RMB mỗi dây chuyền mỗi năm, áp dụng đầy đủ)	26 triệu RMB một năm	80 triệu RMB một năm	235 triệu RMB một năm	Mốc neo phí hằng năm kiểu Baidu

FIG 23 **Từ dưới lên: 40,000 dây chuyền → 16,000 phục vụ được → ≈RMB 320M một lần*Ngách tái kiểm AI bên thứ ba ở Trung Quốc đại lục có thật nhưng nhỏ.***



người mới giành 2–5% dây chuyền phục vụ được trong 3 năm: ≈320–800 dây chuyền, ≈RMB 2–5M/năm
theo đăng ký(RMB 5k mỗi dây chuyền mỗi năm): doanh thu định kỳ RMB 1.6–4M, chỉ 0.3–2.3% dự báo IDC

Nguồn: NVR(cơ sở 2020); ước tính của tác giả

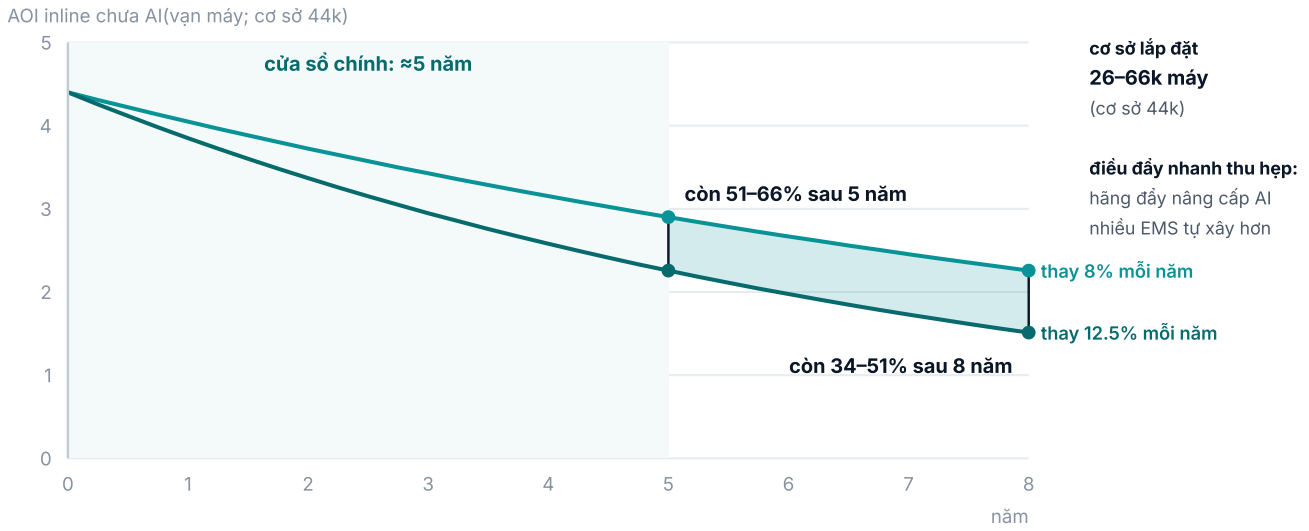
Tất cả đều là ước tính; cả tỷ trọng lẫn giá đều chưa được đo công khai. Mảnh đất ngách chỉ đáng 0.3–2.3% một năm so với dự báo của IDC cho toàn bộ kiểm tra chất lượng AI công nghiệp của Trung Quốc năm 2025 (khoảng 6.2 tỷ RMB). Một người mới giành 2–5% số dây chuyền có thể phục vụ cơ sở trong ba năm sẽ có khoảng 320–800 dây chuyền: tổng 6.4–16 triệu RMB ở 20,000 RMB mỗi dây chuyền, khoảng 2–5 triệu RMB một năm, hoặc 1.6–4 triệu RMB doanh thu định kỳ một năm ở 5,000 RMB mỗi dây chuyền estimate. Đủ cho một đơn vị kinh doanh bên trong một công ty, không đủ cho một công ty độc lập quy mô venture. Mỗi 1,000 dây chuyền lắp đặt ở Đài Loan, Đông Nam Á hoặc Mỹ đáng khoảng 3–15 triệu RMB một lần ở giá đại lục; giá cao hơn ở Đài Loan và Mỹ, nhưng người mua ở đó thường dùng AOI 3D OEM có tùy chọn AI estimate.

Cửa sổ nâng cấp là khoảng năm năm

Không hãng AOI nào công bố tuổi thọ máy; các công ty EMS khấu hao thiết bị trong 3–10 năm (DBG 2025 annual report), và AOI cũ mất cập nhật phần mềm và phụ tùng sau khi hết vòng đời (I-Connect007) single source. Tôi dùng 8–12 năm (cơ sở 10).

Ở 1.0–1.65 AOI nội tuyến mỗi dây chuyền, đại lục có khoảng 35,000–78,000 AOI nội tuyến đang hoạt động (cơ sở khoảng 55,000), nhất quán với các lô hàng đã biết (Circuits Assembly, Sinic-Tek 2025 annual report). Không nguồn nào đo bao nhiêu chạy rà soát AI; giả định máy nội địa đã cung cấp tính năng này từ khoảng 2023 và 40–60% lô hàng gần đây dùng nó, **khoảng 26,000–66,000 AOI nội tuyến ở Trung Quốc đại lục (cơ sở khoảng 44,000) không chạy AI** estimate. Nếu mọi máy mới giờ đều xuất xưởng có AI, nhóm đó thu hẹp 8–12.5% một năm: 51–66% còn lại sau năm năm và 34–51% sau tám năm, nhanh hơn nếu các hãng AOI đẩy nâng cấp AI cho máy mới hơn. **Phần lớn cơ hội rơi vào khoảng năm năm tới**, với các EMS quy mô trung bình chạy AOI 2D nội địa cũ, đa thương hiệu và không có lộ trình nâng cấp AI ở phần lõi.

FIG 24 **Cửa sổ cải tạo khoảng năm năm*AOI inline không AI thu hẹp 8–12.5% mỗi năm khi bị thay hoặc nâng cấp.***

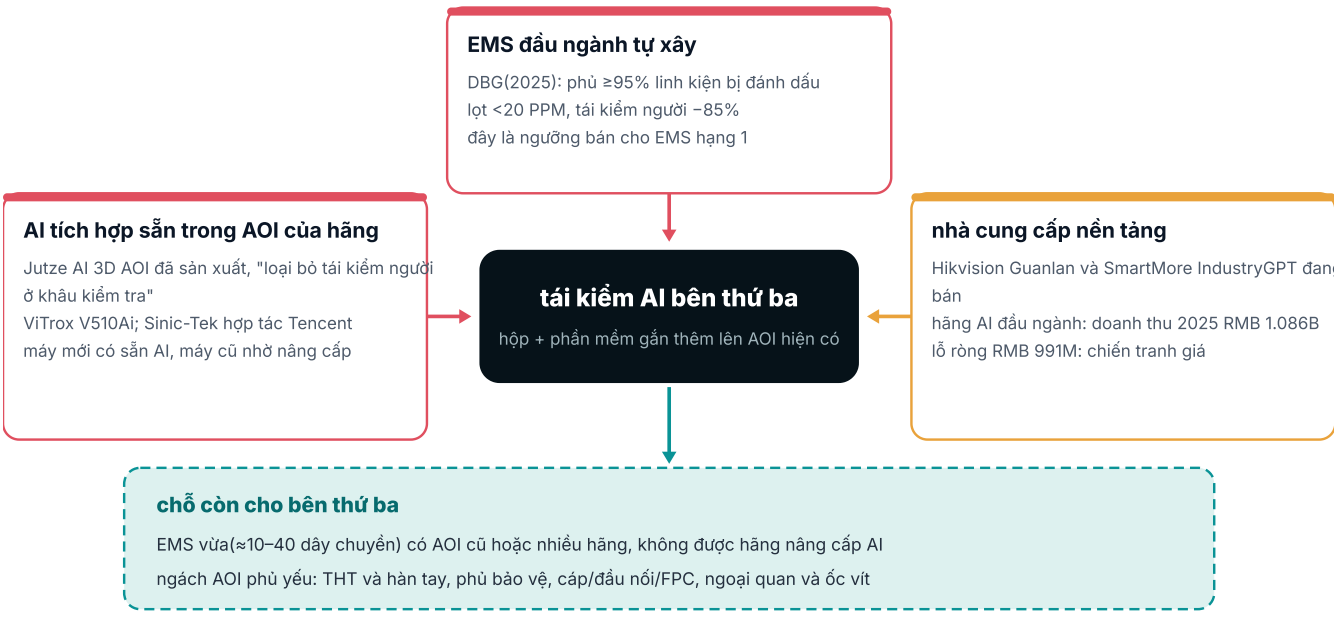


Nguồn: Ước tính của tác giả; cơ sở lắp đặt và thị phần chưa xác minh độc lập

Đối thủ đã mạnh: mảnh đất rà soát đã bị bao vây

Mảnh đất rà soát AI đang bị thu hẹp từ ba hướng. Các hãng dẫn đầu AOI và thị giác đang xây AI vào các nền tảng mới: Jutze đã "hoàn toàn áp dụng thuật toán AI" (Jutze investor relations, June 2026), doanh thu AI công nghiệp của OPT tăng 96.77% năm 2025 (OPT 2025 annual report) single source, primary, và ViTrox tuyên bố tới 95% độ chính xác cho rà soát tự động (ViTrox) vendor claim. Các nhà sản xuất lớn tự xây hệ thống (iThome). Và các công ty AI được đầu tư mạo hiểm bán ở quy mô lớn: khách hàng cảm biến biên của SmartMore tăng từ 65 lên 251 trong hai năm (36Kr) single source, Hikvision bán mô hình Guanlan để kiểm tra (Xinhua), và tích hợp "mắt, não và tay" đã được SmartMore, Hikrobot và Mech-Mind chào bán (36Kr, Securities Star).

FIG 25 **Nêm tái kiểm bị bao vây ba mặt*Chỗ cho bên thứ ba là EMS vừa và ngách kiểm mà AOI phủ yếu.***



Nguồn: Báo cáo năm Jutze, Sinic-Tek và DBG; tài liệu công khai Hikvision và SmartMore; tác giả tổng hợp

Không gian còn lại cho người mới thì hẹp: EMS quy mô trung bình với AOI 2D/3D cũ và không có nâng cấp AI từ hãng; dây chuyền lẫn nhiều thương hiệu AOI (một nhà cung cấp nâng cấp đã hỗ trợ Omron, VI, Koh Young và Mirtec (Sohu on Hengtianyi)); các ngách mà AOI SMT bao phủ kém, như mối hàn xuyên lỗ và hàn tay, phủ conformal, cáp, đầu nối và FPC, cùng ngoại quan cuối cùng và ốc vít; và các nhà máy nhỏ không đủ sức mua một AOI 230–280k RMB.

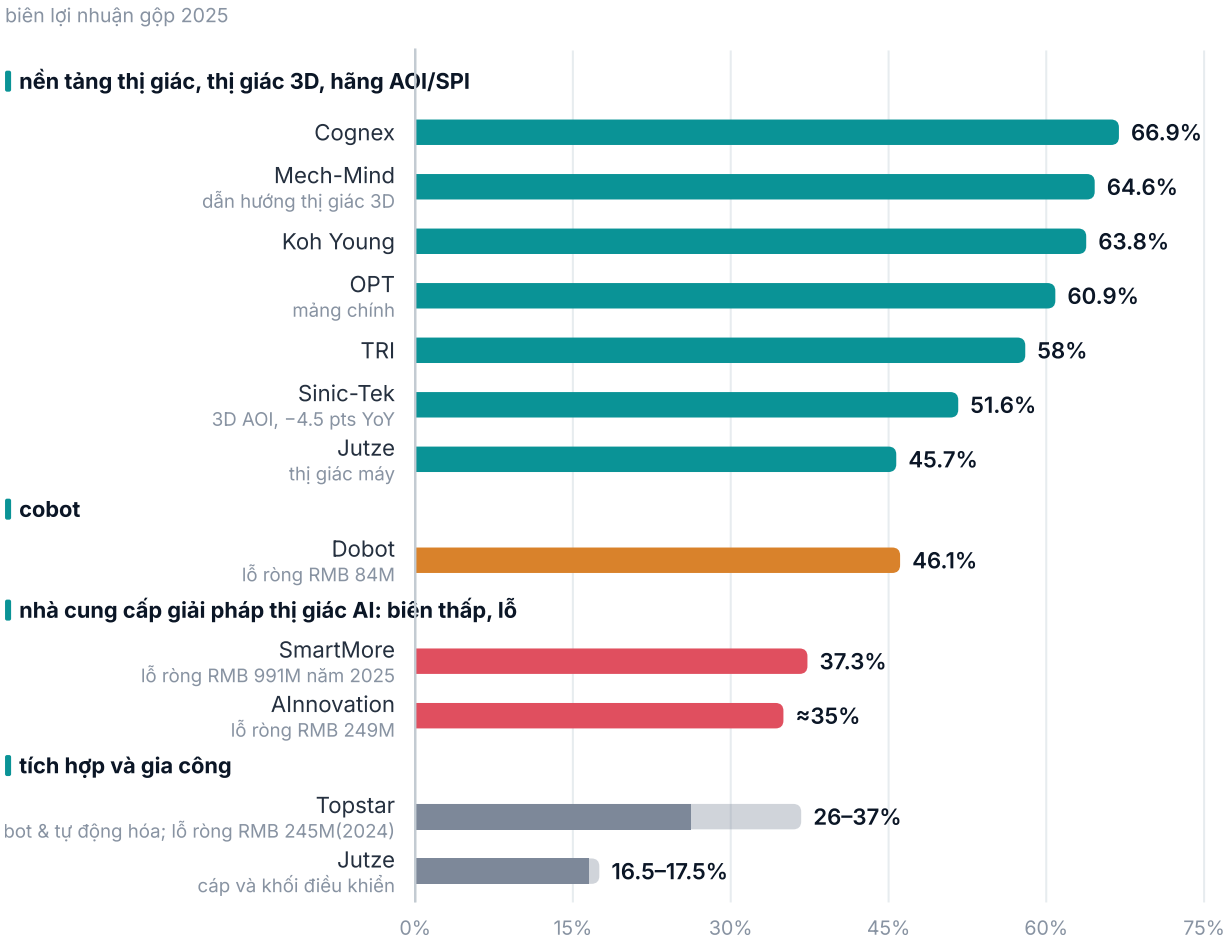
Công ty niêm yết tương đương gần nhất đưa ra cảnh báo. Weiyi Zhizao, bán thiết bị kiểm tra AI và robot, chứng kiến doanh thu 3C giảm 79.9% so với cùng kỳ trong tháng 1–tháng 5 năm 2026, thừa nhận "cạnh tranh gay gắt và áp lực giá", hoãn niêm yết Hồng Kông vào ngày 25 tháng 10 năm 2026, và tiết lộ 15 dự án chấm dứt trước khi hoàn thành, 12 trong số đó là dự án robot hiện thân (Weiyi Zhizao prospectus, postponement notice) single source, primary. AInnovation kết thúc 2025 với 29.8% khoản phải thu bị trích lập dự phòng (AInnovation 2025 annual report). Mảnh đất kiểm tra 3C đông đúc, và quyền định giá theo từng dây chuyền thì yếu.

Lợi nhuận nằm ở phần cứng được sản phẩm hóa

Vị trí trong chuỗi	Công ty tiêu biểu và biên gộp 2025	Lợi nhuận và thu hồi công nợ
Nền tảng thị giác	Cognex 66.9%	Có lãi; đang thoái vốn ≈$22M doanh thu biên thấp
Thiết bị AOI/SPI	Koh Young 63.8%; TRI 58%; mảng lõi OPT 60.85%; Sinic-Tek AOI 3D 51.62% (−4.53 điểm YoY); thị giác máy Jutze 45.74%	Có lãi; ngày phải thu 139 (Jutze) đến 247–274 (OPT)
Dẫn hướng thị giác 3D	Mech-Mind 64.6%	—
Nhà cung cấp giải pháp thị giác AI	SmartMore 37.3%; AInnovation ≈35%	Thua lỗ: SmartMore lỗ ròng 991 triệu RMB năm 2025, AInnovation 249 triệu RMB
Cobot	Dobot 46.1%	Dobot lỗ ròng 84 triệu RMB; doanh thu UR từ $326M (2022) xuống ≈$232M (2025, dự báo)
Tích hợp hệ thống	Topstar robotics và automation 26.2–36.7%	Lỗ ròng 245 triệu RMB (2024); ngày phải thu 180–240
Phần cứng gia công	Cáp và bộ điều khiển Jutze 16.5–17.5%	—

Nguồn: (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 **Biên lợi nhuận nằm ở phần cứng sản phẩm hóa*Nền tảng thị giác và hãng thiết bị đạt biên gộp 46–67%; hãng giải pháp AI 35–37% và lỗ.***



Nguồn: Báo cáo năm 2025 hoặc hồ sơ phát hành

Một người mới chỉ giữ được biên lợi nhuận bằng cách biến kiểm tra thành sản phẩm, thiết bị hoặc phần mềm; tinh chỉnh từng dự án đẩy biên xuống khoảng 35% với thua lỗ dai dẳng. Cơ chế cũng chống lại công ty nhỏ: doanh thu chỉ được ghi nhận sau nghiệm thu tại chỗ, khoản trả trước nhỏ (Jutze, OPT), thí điểm thường miễn phí (A3/Mindtrace), và chủ thương hiệu thường chỉ định thiết bị kiểm tra trên dây chuyền của nhà sản xuất gia công của họ.

Cách người mua thanh toán: 30% khi ký, 60% sau nghiệm thu, 10% giữ một năm

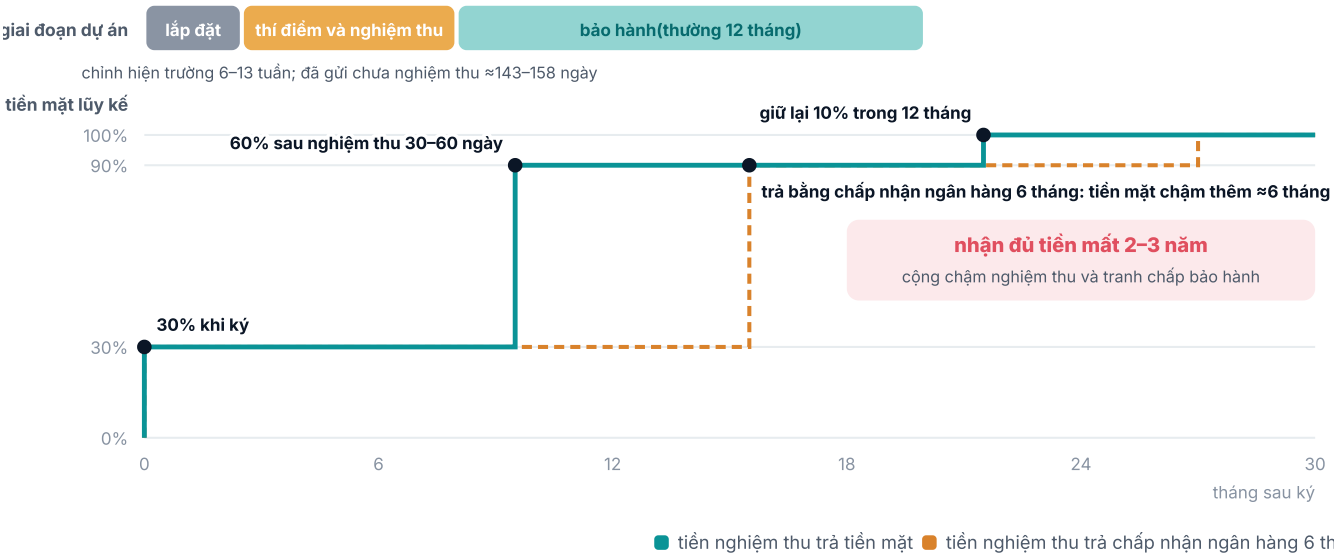
Điều khoản thanh toán mà các nhà máy điện tử đại lục dùng cho thiết bị kiểm tra được nêu trong 13 hợp đồng mua có tên được tiết lộ trong thư trả lời thẩm vấn niêm yết của Sinic-Tek (Sinic-Tek inquiry reply 8-1-1) single source: một hồ sơ gốc, nhiều người mua độc lập, 2020–2022 :

Người mua (năm)	Ký	Giao hàng	Nghiệm thu	Giữ lại hoặc số dư
Một đơn vị Hikvision (2021)	30% trong 15 ngày	—	65% trong 30 ngày sau nghiệm thu cuối	5% bảo hành giữ lại trả 12 tháng sau nghiệm thu; một phần trả bằng hối phiếu chấp nhận ngân hàng 6 tháng+
Nanchang Huaqin (2020–2021)	30% trong 10 ngày làm việc	—	60% ở 60 ngày sau nghiệm thu	10% sau bảo hành một năm

Người mua (năm)	Ký	Giao hàng	Nghiệm thu	Giữ lại hoặc số dư
Guangdong Yinuo (2021, AOI 3D)	30% trong 10 ngày làm việc	—	60% trong 30 ngày	10% sau ba tháng chạy không lỗi
Xiamen Hongxin, Jingmen Hongyi (2022)	30% trong 10 ngày làm việc	—	60% trong 30 ngày làm việc sau nghiệm thu cuối, hối phiếu 6 tháng	10% sau bảo hành một năm
AKM Panyu, Tô Châu (2021–2022)	30% trong 7–10 ngày làm việc	—	60% một tháng sau nghiệm thu và xuất hóa đơn, hối phiếu 180 ngày	10% sau 360 ngày
BYD Precision, Thâm Quyến, Huệ Châu (2020)	—	30% khi đến	70% sau nghiệm thu và xuất hóa đơn	Toàn bộ bằng hối phiếu thương mại 6 tháng
Shenzhen H&T (2020)	—	—	90% sau nghiệm thu và xuất hóa đơn	10% sau sáu tháng sử dụng không lỗi

Các hồ sơ gốc khác cho thấy cùng cấu trúc: các dự án lớn của Mech-Mind trả 30% khi ký, 30% trước giao hàng, 30% sau nghiệm thu và 10% giữ bảo hành trong 12–15 tháng (Mech-Mind prospectus); GKG nói khoản giữ lại trong ngành thiết bị thường là 5–10% hợp đồng, trả một năm sau nghiệm thu (GKG inquiry reply) single sources, primary. Chuẩn mực, vậy, là khoảng 30% khi ký (20–30%, bằng không ở một số nhà máy như BYD), 60–65% sau nghiệm thu (trả 30–60 ngày sau, thường bằng hối phiếu ngân hàng hoặc thương mại sáu tháng, trì hoãn thêm nửa năm tiền mặt dùng được), và 5–10% giữ lại 3–12 tháng (12 tháng phổ biến nhất) verified trên nhiều hợp đồng.

FIG 27 **Người mua trả thế nào *30% khi ký, 60% sau nghiệm thu, 10% giữ một năm: nhận đủ mất 2–3 năm.***



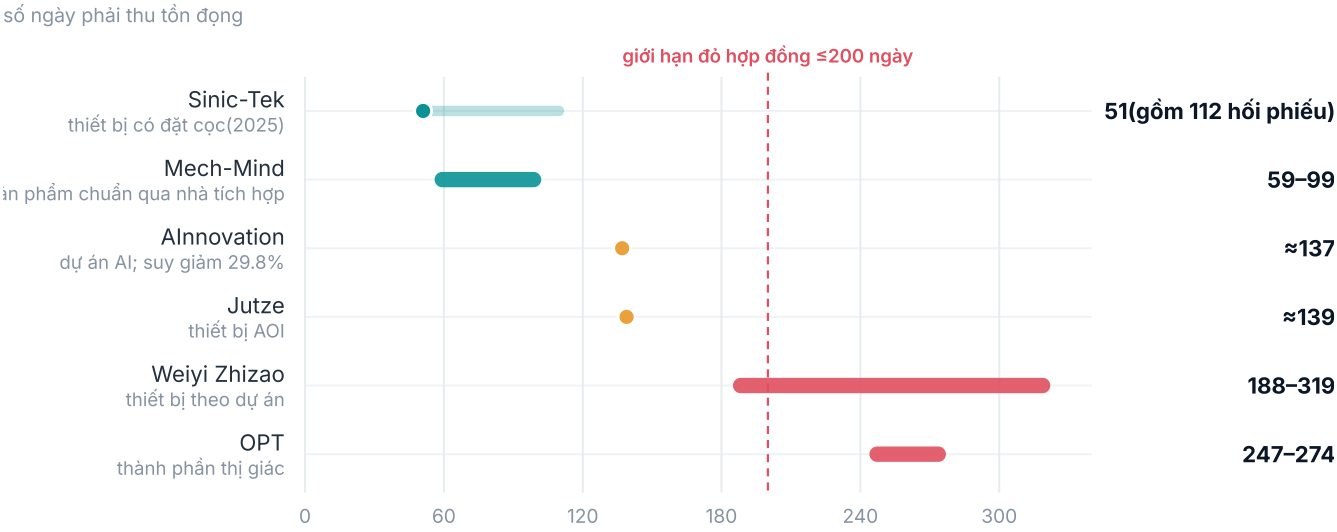
Nguồn: Trả lời hỏi đáp Sinic-Tek, Mech-Mind, GKG, JT và khác; sơ đồ điều kiện thường, thời gian ước tính

Nghiệm thu và chạy thử chậm và do người bán trả phí. Hàng đã giao nhưng chưa nghiệm thu của Sinic-Tek mất khoảng 143–158 ngày để quay vòng, vì khách hàng EMS và thương hiệu lớn cần kiểm tra tích hợp toàn dây chuyền và một đợt chạy thử kéo dài (Sinic-Tek inquiry reply 8-1-2); các dự án AI của Weiyi mất 4–8 tháng từ đánh giá đến triển khai (Weiyi Zhizao prospectus) single sources, primary. Không có dữ liệu chuyển đổi cho thí điểm rà soát AI; Sinic-Tek chỉ chuyển đổi khoảng 18–19% máy demo trực tiếp thành doanh số. Tôi lập kế hoạch ở mức một trên ba đến một trên năm cho thí điểm miễn phí và khoảng 80% cho thí điểm trả phí sau sàng lọc estimate.

Công ty	Mô hình kinh doanh	Ngày doanh thu tồn đọng
Sinic-Tek (2025)	Thiết bị có đặt cọc	≈51 (≈112 gồm hối phiếu)
Mech-Mind	Sản phẩm tiêu chuẩn qua nhà tích hợp	59–99
Jutze	Thiết bị AOI	≈139
AInnovation (2025)	Dự án AI	≈137 cuối năm, với 29.8% trích lập dự phòng
Weiyi Zhizao	Thiết bị theo dự án	188–319
OPT	Linh kiện thị giác	247–274

Nguồn: các báo cáo thường niên và bản cáo bạch ở trên; khoản phải thu (gồm tài sản hợp đồng) ÷ doanh thu × 365 primary + estimate .

FIG 28 **Số ngày phải thu tồn đọng** ***Sản phẩm chuẩn 51–99 ngày; thiết bị theo dự án 188–319 ngày.***



Nguồn: Hồ sơ công ty; phải thu(gồm tài sản hợp đồng) ÷ doanh thu × 365

Giấy phép rà soát AI bán qua hãng AOI hoặc nhà tích hợp nên thu hồi như mức 60–100 ngày của Mech-Mind; bán dự án trực tiếp như mức 190–320 ngày của Weiyi. Gộp thành dòng thời gian tiền mặt: thí điểm miễn phí một đến ba tháng, phê duyệt tại một ODM lớn tới khoảng mười hai tháng, xây dựng, giao hàng và nghiệm thu ba đến bảy tháng; 30% về cùng đơn hàng, 60% khoảng bốn đến chín tháng sau đó và 10% cuối cùng khoảng 16–21 tháng sau đó. **Từ thí điểm đầu tiên tại một EMS lớn đến thanh toán đầy đủ có thể mất hai đến ba năm**; EMS quy mô trung bình không đấu thầu chính thức có thể rút ngắn phê duyệt xuống vài tuần estimate . Tiêu chí nghiệm thu trong các hợp đồng được tiết lộ chỉ bao phủ tích hợp toàn dây chuyền, thông lượng (một dự án Đài Loan yêu cầu 100 chiếc mỗi giờ) và vài tháng chạy không lỗi; không hợp đồng nào quy định tỷ lệ lọt lưới hay cảnh báo giả (Anda inquiry reply) single source, primary . Tiêu chí nghiệm thu bằng văn bản cho rà soát AI phải được thỏa thuận trước khi thí điểm.

Bốn kiểu mẫu thất bại

Các thất bại được ghi nhận rơi vào bốn kiểu mẫu: giao hàng trước khi sản phẩm đáng tin cậy (Rethink (The Robot Report)); một lớp phần mềm trên phần cứng của hãng dẫn đầu mà hệ sinh thái của hãng đó sau này thay thế (READY Robotics (The Robot Report)); hết tiền ở đáy chu kỳ (Dextrous Robotics (Robots & Startups)); và các công ty mô hình kết thúc bằng acqui-hire (Covariant (GeekWire)). Các bài học: sản phẩm hóa trong phạm vi hẹp thay vì tinh chỉnh từng dự án; giữ vốn lưu động cho doanh thu theo nghiệm thu và thu hồi công nợ dài; cung cấp dịch vụ địa phương cho nhà máy chạy 24 giờ; và đừng mở rộng trước khi sản phẩm được chứng minh.

Thời điểm: kiểm tra không được lợi gì từ chờ đợi; phần cứng cánh tay rẻ hơn

Các lập luận cho việc chờ đợi thì yếu. Phần cứng không rẻ hơn trong 2026: các hãng robot tăng giá 5–15%, một số sản phẩm 25–50% (Sina Finance) single source, và các bộ NVIDIA tăng lần thứ ba (36Kr). VLA cấp sản xuất sẽ không đến trong cửa sổ này (IFR, Instrumental). Trong khi đó mỗi lần thay AOI lại loại bỏ một mục tiêu nâng cấp và mỗi lần triển khai AI OEM lại khóa dữ liệu và quy trình của một dây chuyền; năng lực mô hình đang bị hàng hóa hóa, và thứ khan hiếm là quyền tiếp cận nhà máy và dữ liệu nhà máy đã gán nhãn. Hỗ trợ địa phương cũng có cửa sổ: quy tắc "thành phố tiên phong AI" 2026 của Thâm Quyến, với tới 2 triệu RMB cho dự án trình diễn ứng dụng AI và phiếu mua mô hình chi trả 30% giao dịch mua, chạy đến **ngày 30 tháng 6 năm 2027** (Yicai, May 2026) điều khoản single source.

Thứ vẫn đang rẻ đi nhanh là robot hình người và bàn tay khéo léo (bàn tay giảm từ khoảng 50,000 RMB xuống dưới 10,000 trong hai năm (Sina Finance)), và cobot nhập môn khi cơ cấu sản phẩm dịch xuống; điều đó ủng hộ việc hoãn phần hiện thân, không phải phần kiểm tra. **Tổng thể: kiểm tra ngay, như một sản phẩm hẹp; cụm robot ngay, qua đối tác và tích hợp; AI hiện thân sau, chỉ với một trạm R&D hôm nay.**

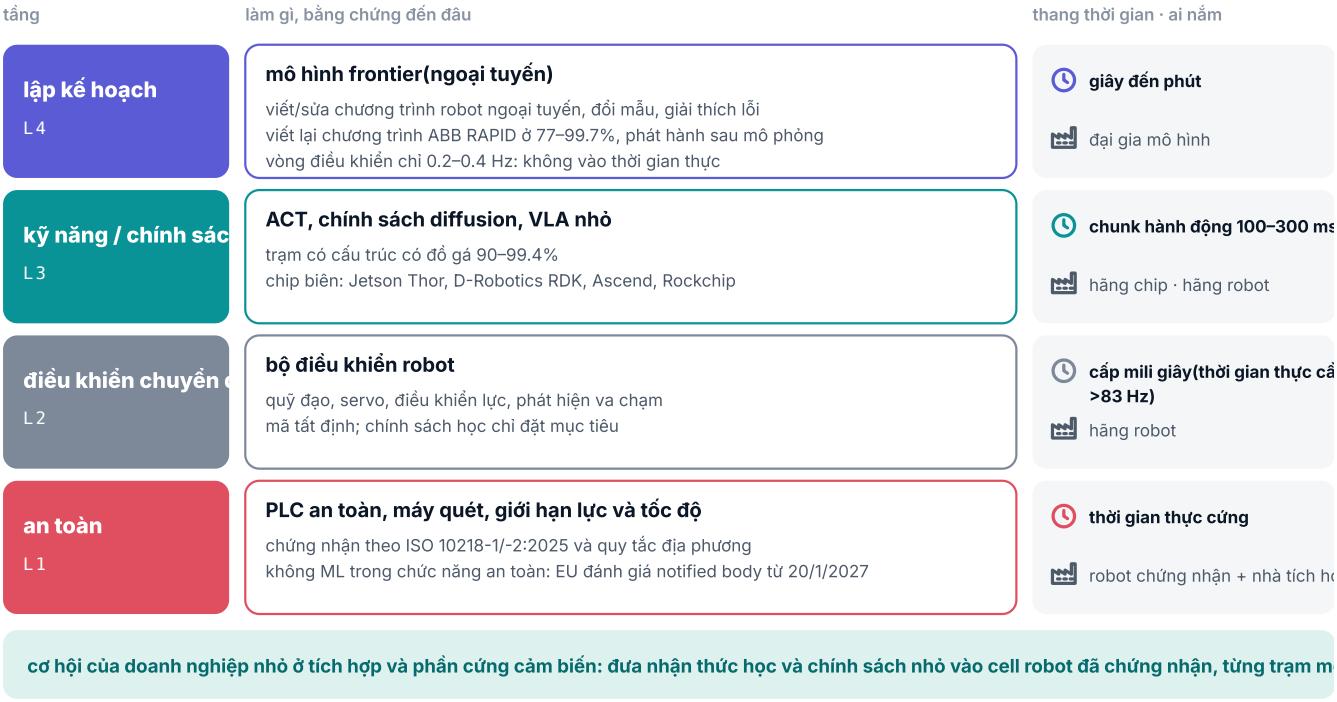
AI hiện thân

06 AI hiện thân: trước tiên nó thay đổi cách robot được dạy, sau đó là cách chúng di chuyển

Điều đang thay đổi là nhận thức và lập trình; chuyển động vẫn mang tính xác định

Mọi hãng dẫn đầu cánh tay robot giờ đều tiếp thị "AI vật lý", nhưng thứ đã xuất xưởng rơi vào ba lớp: nhận thức học được không cần dạy (AI Masterless Vision của FANUC, xuất xưởng tháng 1 năm 2027 (FANUC); lựa chọn cầm nắm bằng nơ-ron của KUKA (KUKA)); các mô hình lớn viết chương trình và lập kế hoạch nhiệm vụ (iiQWorks.Copilot của KUKA, cảnh báo rằng mã do AI sinh nên được xác thực trong bản sao số (KUKA); Yaskawa với Gemini Robotics-ER, nơi "AI sinh quyết định làm gì" và robot quyết định "di chuyển thế nào" (Yaskawa)); và các nền tảng tính toán và dữ liệu (AI Accelerator của UR với Jetson Orin (Universal Robots)) vendor claims . Không hãng dẫn đầu nào công bố tỷ lệ thành công sản xuất cho các tính năng AI của mình, và các hãng dẫn đầu Nhật Bản hợp tác với Google, SoftBank, Microsoft và các bên khác thay vì tự xây VLA frontier (FANUC news index): lớp chính sách học được sẽ đến từ các công ty mô hình, cánh tay và bộ điều khiển đã chứng nhận thì từ các hãng cánh tay.

FIG 29 **Ngăn xếp AI embodied** ***Mô hình lớn lập kế hoạch ngoại tuyến, chính sách nhỏ chạy trong cell, chuyển động và an toàn giữ tất định.***



Nguồn: Tổng hợp của tác giả(Chương 3 và 6); ISO 10218:2025; Quy định máy móc EU

Cách công việc phi cấu trúc, linh hoạt, lô nhỏ bao xa: trạm có cấu trúc trong 1–3 năm, lắp ráp linh hoạt 5+

Các con số nhà máy thực đã bắt đầu xuất hiện, nhưng chúng do nhà cung cấp báo cáo, hẹp và từ các trạm có cấu trúc. Tám robot hình người của AgiBot trên dây chuyền kiểm tra cuối máy tính bảng của Longcheer ghi nhận 64,828 nhiệm vụ ở 99.99% trong hơn 64 giờ (AgiBot, Yicai Global) vendor và khách hàng , nhưng đây là công việc phục vụ mà một cánh tay truyền thống có thị giác có thể làm. UBTech nói Walker S2 của họ làm việc ở 30–50% hiệu suất con

người (Irish Times/FT) single source. 11 tháng của Figure 02 tại BMW Spartanburg tạo ra các con số khối lượng (1,250+ giờ, 90,000+ linh kiện), trong khi các con số chu kỳ, thành công và can thiệp là mục tiêu, không phải kết quả (Figure, BMW Group) verified, chỉ từ số. Nucleus nói khoảng 40% video nhà máy của họ là điều khiển từ xa (Interesting Engineering) vendor claim, và một điều tra của Reuters trích khoảng 20% sản lượng của một công nhân cho một số robot hình người Trung Quốc (Notebookcheck) single source. Không triển khai nào công bố đồng thời tỷ lệ thành công, thời gian chu kỳ và tỷ lệ can thiệp, và không triển khai nào được kiểm toán, nhưng chính ba thứ đó là những gì một phép tính hoàn vốn cần.

Triển khai	Tỷ lệ thành công	Takt	Can thiệp hoặc điều khiển từ xa	Thời gian hoạt động hoặc khoảng lỗi	Lượt chạy khối lượng
Figure 02 @ BMW Spartanburg	Chỉ mục tiêu (>99% mỗi ca)	Chỉ mục tiêu (84 s)	Chỉ mục tiêu (0 mỗi ca)	Không tiết lộ (điểm lỗi là đầu cắng tay)	1,250 h
Agility Digit @ GXO	Không tiết lộ	Không tiết lộ	Không tiết lộ	Không tiết lộ	100,000 thùng
Apptronik Apollo @ Mercedes-Benz	Không tiết lộ	Không tiết lộ	Không tiết lộ	Không tiết lộ	Không tiết lộ
Agile ONE @ Audi	Không tiết lộ	Không tiết lộ	Không tiết lộ	Không tiết lộ	Không tiết lộ
Nucleus (nhà máy không nêu tên)	Không tiết lộ	Không tiết lộ	≈40% điều khiển từ xa (phương pháp không tiết lộ)	Không tiết lộ	≈2 h video
AgiBot G2 @ Longcheer	99.99% trong 64,828 nhiệm vụ	18–20 s mỗi bước	Không tiết lộ	Một lỗi truyền thông, ≈2 phút khôi phục	64+ h
Cụm lai Neuromeka	322/324	Chậm hơn người 12.8%	Không tiết lộ	Không tiết lộ	324 chu kỳ
Physical Intelligence π0.6 tại địa điểm khách hàng	—	—	96.4% tự động trong một ca	Không tiết lộ	Một ca

Nguồn: chương này và Chương 3 chủ yếu vendor claims.

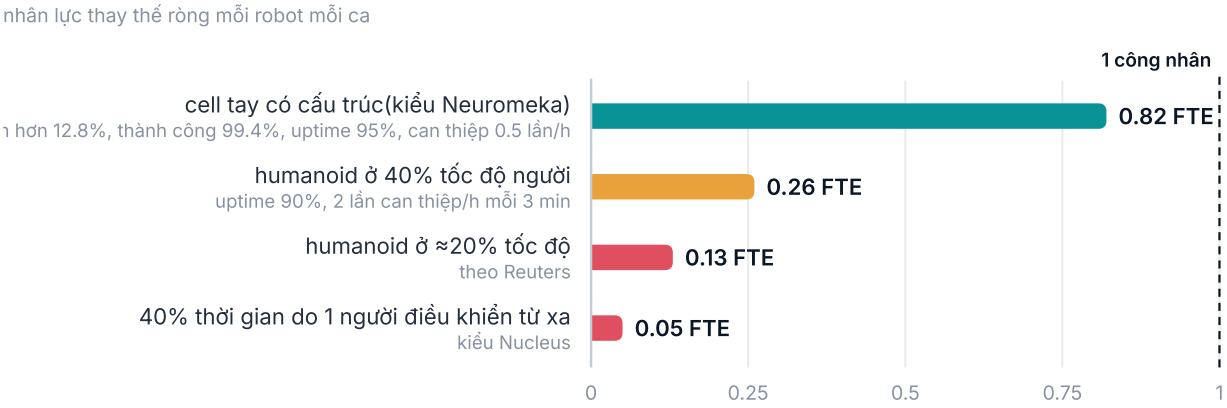
FIG 30 **Điều triển khai robot trong nhà máy thực sự công bố** ***Đa số chỉ công bố số lượng; không ai công bố đồng thời tỷ lệ thành công, takt và tỷ lệ can thiệp.***



Nguồn: Tài liệu công khai từ BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence và khác

Điều khiển từ xa quan trọng với kinh tế hơn tỷ lệ thành công. Số người ròng được thay thế mỗi robot mỗi ca ≈ (chu kỳ người ÷ chu kỳ robot) × thời gian hoạt động × thành công, trừ thời gian can thiệp và điều khiển từ xa estimate, công thức của tôi. Một trạm có cấu trúc kiểu Neurometka (chậm hơn 12.8%, thành công 99.4%, thời gian hoạt động 95%, 0.5 lần can thiệp 2 phút mỗi giờ) thay thế khoảng 0.82 người; một robot hình người ở 40% tốc độ người, 90% thời gian hoạt động và hai lần can thiệp 3 phút mỗi giờ khoảng 0.26; ở mức 20% của Reuters, khoảng 0.13; và nếu, như Nucleus, một người vận hành điều khiển từ xa 40% thời gian, khoảng 0.05, gần như không có gì. Ở khoảng 600k RMB mỗi robot hình người và 85k RMB mỗi người mỗi năm ở đại lục, ba cái sau hoàn vốn trong nhiều thập kỷ; chỉ trạm cánh tay có cấu trúc có hoàn vốn thực tế (Chương 4) estimate.

FIG 31 **Nhân lực thay thế ròng mỗi robot** ***Cell tay có cấu trúc thay ≈0.82 người; humanoid điều khiển từ xa 40% thời gian ≈0.05.***



Nguồn: Ước tính theo công thức của tác giả

Rời khỏi trạm có cấu trúc, các mô hình tổng quát vẫn đạt 40–90% trên nhiệm vụ mới: Gemini Robotics 2 đạt 74.2% trên pick-and-place phòng thí nghiệm, 78.9% trên kitting dụng cụ và 44% trên buộc túi (eesel/DeepMind); trên RoboChallenge Table30, 30 nhiệm vụ robot thật, π0.5 đạt 43.7% và mô hình mở tốt nhất 50.3% (RoboChallenge, QbitAI).

Các dự báo bên ngoài khác nhau, nhưng không dự báo nào ủng hộ sử dụng rộng rãi trong hai năm. IFR nói AI vật lý tổng quát "chưa đạt cấp công nghiệp" và hầu hết robot hình người "vẫn là nguyên mẫu hoặc thí điểm" (IFR); Interact Analysis ước tính chỉ khoảng 10% robot hình người sản xuất năm 2025 đi vào sử dụng thực, với điểm bùng nổ thương mại "sau 2032" (automation.com); UBS thấy "khoảnh khắc EV" của robot hình người khó xảy ra trước 2030 (Yicai Global); người trong ngành Trung Quốc ước tính còn hai đến ba năm nữa trước khi sử dụng quy mô nhà máy (Gasgoo) single sources. Phán đoán của tôi estimate: trong các nhà máy điện tử và ô tô có nguồn lực tốt với nhà cung cấp tại chỗ, **nhiệm vụ trạm có cấu trúc** (phục vụ, phân loại OK/NG, cắm cố định) sẽ đạt ≥99% ở tốc độ gần người trong khoảng 1–3 năm; **kitting linh kiện hỗn hợp** với VLA tổng quát trong khoảng 3–5; **bó dây định tuyến tự do và lắp ráp linh hoạt nhiều bước** trong hơn 5 năm. Hoàn vốn robot hình người không khả thi hôm nay: một Walker S2 ở khoảng 600k RMB (10jqka) single source, ở 30–50% hiệu suất người trong hai ca, bằng khoảng 0.6–1.0 công nhân, một hoàn vốn đơn giản 7–12 năm ở chi phí lao động đại lục trước tích hợp, bảo trì và thời gian ngừng estimate.

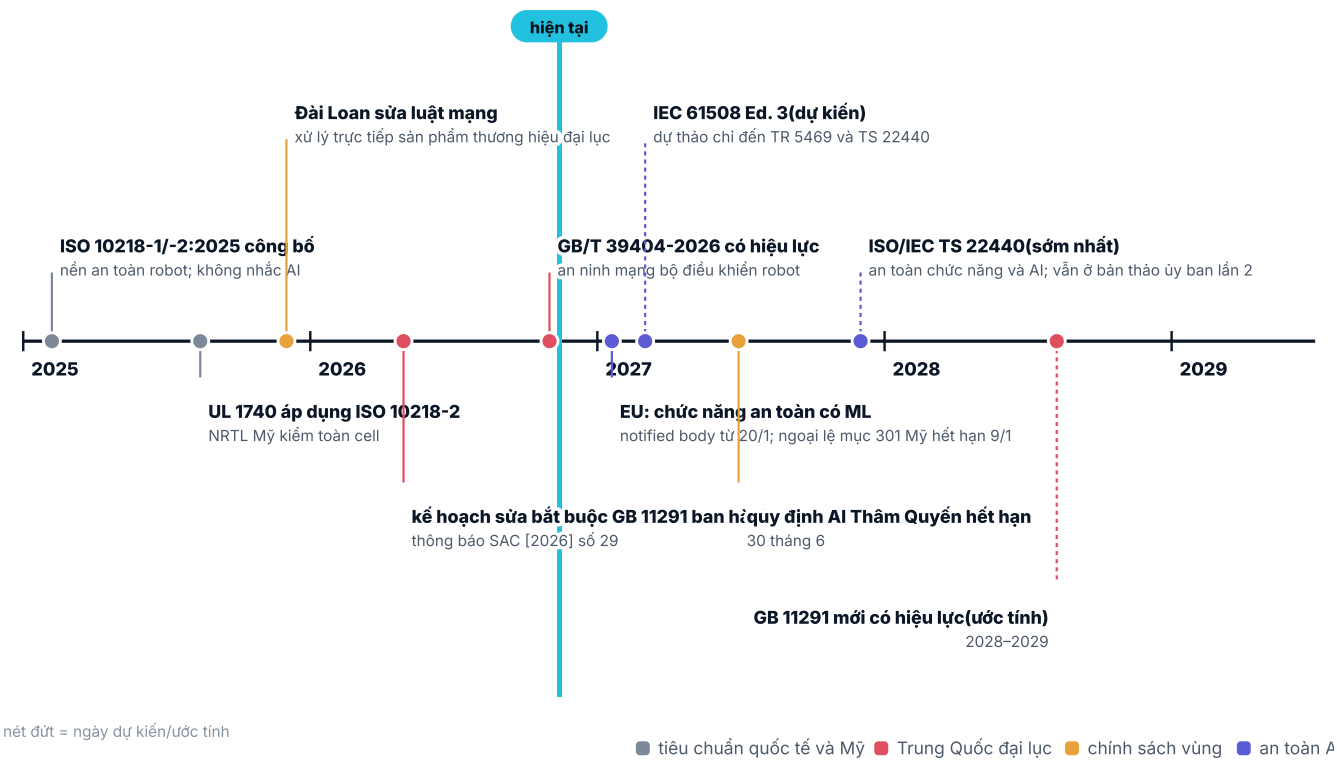
Nơi nó đang kẹt

Khâu	Bằng chứng năm 2026
Dữ liệu	Ngành ước tính cần ít nhất 10 triệu giờ dữ liệu thực chất lượng cao và chỉ khoảng 500,000 giờ đã được thu thập vào đầu 2026 (Gasgoo); Reuters báo cáo khoảng 500,000 giờ so với khoảng 100 triệu giờ cần thiết (Notebookcheck); một người vận hành từ xa mới cần khoảng 300 lần thử mỗi chuyển động dùng được; nhà máy sẽ không chia sẻ dữ liệu quy trình; chính sách huấn luyện bằng mô phỏng rơi mạnh trên robot thật
Độ tin cậy	Trạm có cấu trúc 98–99.99% (vendor claims, lỗi không được định nghĩa); mô hình tổng quát 40–90% trên nhiệm vụ mới; đầu nối chưa thấy 29–67%

Khâu	Bằng chứng năm 2026
Tốc độ	Robot hình người ở 30–50% tốc độ người (UBTech), một số khoảng 20% (Reuters); tác từ frontier khoảng 280 s mỗi nhiệm vụ robot bán tĩnh; vòng điều khiển LLM ở 0.2–0.4 Hz so với khoảng 83 Hz cần thiết (Anthropic)
Chi phí	Walker S2 trung bình khoảng 600k RMB; hoàn vốn 7–12 năm estimate
Chứng nhận an toàn	ISO 10218-1/-2:2025 làm rõ hơn an toàn chức năng và hấp thụ các yêu cầu hợp tác, nhưng không nói gì về AI (The Robot Report); ISO/IEC TS 22440 (an toàn chức năng và AI) vẫn là bản dự thảo ủy ban thứ hai vào tháng 11 năm 2026 (ISS, ISO Open Data); hướng dẫn IEC 61508-3 khuyên không dùng AI trên SIL 1 (Analog Devices); theo Quy định Máy móc EU, chức năng an toàn có học máy cần đánh giá bởi tổ chức được công nhận từ ngày 20 tháng 1 năm 2027 (EUR-Lex)
Bàn tay và dụng cụ	Bàn tay khéo léo đạt khoảng 30 N (≈3 kg) mỗi ngón so với khoảng 10 kg của tay người (Gasgoo); thay dụng cụ vẫn thủ công
Tích hợp	Đọc nhầm cảm biến, lệch giao thức và liên kết MES tại hiện trường; tỷ lệ can thiệp hầu như không bao giờ tiết lộ

Hàm ý an toàn là trực tiếp: các tiêu chuẩn không cấm cũng không công nhận chính sách học được. ISO 10218-1/-2:2025 không nói gì về AI; ISO/IEC TS 22440 (an toàn chức năng và AI) vẫn ở bản dự thảo ủy ban và khó ra trước cuối 2027 hoặc 2028 estimate ; ISO/IEC TR 5469:2024 chỉ là hướng dẫn, khuyến nghị dự phòng phi AI, ràng buộc giám sát và bỏ phiếu dư thừa; IEC 61508 bản 3 dự kiến đầu 2027 (ISO Open Data, BSI, 61508 Association) verified . Chưa tiêu chuẩn công bố nào đưa ra lộ trình được chấp nhận để chứng nhận nhận thức học máy là một chức năng an toàn. Các quy tắc khu vực cũng đang dịch chuyển: Trung Quốc đang sửa GB 11291 (có hiệu lực khoảng 2028–2029) (revision plan 20262543); UL 1740 đã đưa ISO 10218-2 vào đánh giá NRTL của Mỹ từ tháng 8 năm 2025; EU yêu cầu đánh giá bởi tổ chức được công nhận đối với chức năng an toàn có học máy từ ngày 20 tháng 1 năm 2027; và chưa robot hình người nào được chứng nhận sử dụng nhà máy không rào trong khi ISO 25785-1 vẫn ở bản dự thảo (i-SCOOP). Kiến trúc có thể bảo vệ: học máy chỉ đưa ra quyết định quy trình (“linh kiện nào”, “tấm mạch này tốt không”), trong khi một lớp an toàn xác định đã chứng nhận (thường là PL d / Category 3) và một PLC an toàn quyết định tốc độ, lực và dừng. Hãy coi bất kỳ tuyên bố “AI camera cho an toàn không rào” nào là chưa được chứng minh cho đến khi một chứng chỉ PL d hoặc SIL 2 nêu tên thành phần học máy.

FIG 32 **Dòng thời gian tiêu chuẩn và quy định*Lộ trình chứng nhận AI trong chức năng an toàn chưa ổn định trước 2027–2029.***



Nguồn: ISO, IEC, UL, EUR-Lex, SAC và cơ sở dữ liệu quy định quốc gia

Lớp nào là cơ hội: mô hình thuộc về các ông lớn, hộp biên thuộc về hãng chip, doanh nghiệp nhỏ thuộc về tích hợp và phần cứng cảm biến

Lớp mô hình cần vốn và dữ liệu và thuộc về các ông lớn (AgiBot huấn luyện GO-2 trên "hàng chục nghìn giờ" (iFeng Tech); các mô hình mạnh nhất là đóng hoặc chỉ trên đám mây). Lớp hộp biên thuộc về hãng chip: Jetson Thor là tiêu chuẩn thực tế, và ở đại lục, các board RDK của D-Robotics (D-Robotics) và bộ đồng xử lý Rockchip (CSDN) cạnh tranh, chủ yếu không có kết quả VLA công bố.

Doanh nghiệp nhỏ có hai lối mở. Một là phần cứng cảm biến trực tiếp mua độ tin cậy: cảm ứng nâng thành công cắm của ContactDP từ 58.7% lên 97.8%, và cảm biến lực/mô-men và camera cổ tay làm công việc tương tự; bộ thu thập dữ liệu cũng bán được (10jqka), dù bàn tay, cảm biến và cánh tay đều đang giảm giá. Lối còn lại là lớp tích hợp giữa hãng cánh tay và công ty mô hình, sở hữu dữ liệu nhiệm vụ, đường bao an toàn và liên kết MES của từng trạm, thứ mà các hãng cánh tay không bán như một sản phẩm hôm nay. Đối thủ rơi vào bốn nhóm: hãng dẫn đầu cánh tay (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot và các hãng khác), doanh nghiệp thị-giác-cộng-cánh-tay (Mech-Mind, Hikrobot, SmartMore), công ty AI hiện thân (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) và các bên tham gia chip ô tô.

Sự chấp nhận của người mua: không con số nào mà không có mẫu số

Vì không nhà cung cấp nào công bố tỷ lệ can thiệp, khoảng lỗi hay số lần thành công dài hạn, những thứ này thuộc về điều khoản nghiệm thu của hợp đồng. Về thống kê, 3,000 chu kỳ sản xuất liên tiếp với không lỗi chứng minh tỷ lệ lỗi tối đa 0.1% ở độ tin cậy 95%; 10 lỗi trong 3,000 chỉ chứng minh ≤0.57%; chứng minh ≤1%, ≤0.5%, ≤0.1% và ≤0.01% với không lỗi cần 299, 598, 2,995 và 29,956 chu kỳ estimate, nhị thức chính xác. Theo tiêu chuẩn đó, 322/324 của Neurometka có khoảng 95% hai phía là 97.8–99.9% và không chứng minh ≥99%; khoảng 6 lỗi trong 64,828 nhiệm vụ của AgiBot ngụ ý 99.98–100%, nhưng chỉ khi "lỗi" được định nghĩa theo cách người mua định nghĩa estimate. Đây là các yêu cầu nghiệm thu mà tôi sẽ đặt ra cho bất kỳ cụm cánh tay robot hay hiện thân nào; các ngưỡng dựa trên chính mục tiêu của Figure tại BMW và trên các ca Neurometka và AgiBot, và chúng là phán đoán, không phải tiêu chuẩn estimate:

Hạng mục	Yêu cầu	Ngưỡng vượt gợi ý
Định nghĩa	Định nghĩa có ký về chu kỳ, thành công, lỗi (thành công một phần, cầm lại, rơi, lọt lỗi chất lượng), can thiệp (chạm, điều khiển từ xa, đặt lại, dừng khẩn cấp) và cách tính mức tự động	Định nghĩa có trước mọi chỉ số; không phần trăm nào mà không có công thức
Can thiệp và điều khiển từ xa	Theo giờ robot, theo loại và thời lượng; thời gian trung bình giữa các lần can thiệp; phút điều khiển từ xa theo tỷ trọng thời gian chạy; số người vận hành mỗi robot	Trạm có cấu trúc ≤0.5 lần can thiệp mỗi giờ robot (cách nhau ≥2 h); không điều khiển từ xa trong nghiệm thu
Thành công dài hạn	≥3,000 chu kỳ sản xuất liên tiếp trong ≥3 ca, ≥2 mã linh kiện và biến thiên ánh sáng và cung ứng bình thường; phân loại lỗi; cận dưới chính xác	Cận dưới ≥99.5% cho trạm nội tuyến dừng dây chuyền, ≥99.0% ngoại tuyến; lọt lỗi chất lượng báo cáo riêng
Thời gian chu kỳ	P50/P90/P99 mỗi linh kiện gồm khôi phục, so với ≥30 chu kỳ người bấm giờ tại cùng trạm	≤1.15 × P50 người, hoặc trong takt dây chuyền
Thời gian hoạt động và lỗi	Giờ chạy, nhật ký ngừng, lỗi theo linh kiện, thời gian thay linh kiện	≥95% thời gian hoạt động trong ≥2 tuần; MTTR ≤20 phút với phụ tùng tại chỗ
Khôi phục	Tiêm lỗi kịch bản: rơi, che khuất, đổi ánh sáng, dừng băng chuyền, mất MES hoặc mạng, mất điện, dừng khẩn cấp	≥90% khôi phục tự động, trở lại sản xuất trong 2 phút
Chuyển đổi	Giờ, số lần trình diễn, kỹ năng người vận hành và chi phí thu thập dữ liệu mỗi mã linh kiện hoặc đồ gá mới	Trạm có cấu trúc: ≤1 ca và ≤100 lần trình diễn mỗi mã linh kiện mới

Hạng mục	Yêu cầu	Ngưỡng vượt gợi ý
An toàn	Cánh tay và cụm theo ISO 10218-1/-2:2025, đánh giá rủi ro cụm, quy tắc địa phương (GB 11291 ở đại lục; CNS 14490 cộng đánh giá bằng văn bản ở Đài Loan; ANSI/A3 R15.06-2025 và UL 1740 ở Mỹ)	Không chức năng an toàn nào do học máy thực hiện; một lớp an toàn đã chứng nhận (thường PL d / Cat 3) giới hạn tốc độ và lực; không robot có chân không rào cho đến khi ISO 25785-1 được công bố và đánh giá
Tính toán và hành vi ngoại tuyến	Phần cứng suy luận, độ trễ P99, hành vi khi không mạng, phiên bản mô hình, tái xác thực sau mỗi cập nhật	Suy luận trọng yếu takt không phụ thuộc đám mây; nghiệm thu lại đầy đủ sau mọi thay đổi mô hình, lượng tử hóa hoặc engine
Bằng chứng	Video chưa chỉnh sửa và nhật ký thô cho toàn bộ cửa sổ nghiệm thu; được đội chất lượng của khách hàng hoặc bên thứ ba chứng kiến	Không highlight reel hay tóm tắt chỉ của nhà cung cấp
Điều khoản thương mại	Mốc thanh toán gắn với chỉ số nghiệm thu; phạt tính sẵn sàng khi tính phí theo giờ robot; quy tắc sở hữu dữ liệu và truy cập từ xa	Thanh toán cuối và khoản giữ lại chỉ giải phóng sau nghiệm thu 3,000 chu kỳ và ≥4 tuần chạy ổn định
Kinh tế	Số người ròng thay thế = tỷ số tốc độ × thời gian hoạt động × thành công − thời gian can thiệp − tỷ trọng điều khiển từ xa	Từ chối mọi đề xuất mà mức thay thế ròng không đỡ nổi hoàn vốn đã tuyên bố

Nguồn: (Figure, Neuromeka, Yicai Global, i-SCOOP); các ngưỡng là gợi ý của tôi.

Các câu hỏi bỏ ngỏ

07 Những gì hồ sơ công khai không thể trả lời, và cách một thí điểm giải quyết nó

Các câu hỏi mà nguồn công khai không thể trả lời rơi vào năm nhóm: rà soát AI thực sự bán giá bao nhiêu và người mua sẽ trả bao nhiêu; cách các hãng AOI đóng gói AI và nâng cấp máy cũ; các AOI đã lắp đặt bao nhiêu tuổi và bao nhiêu đã chạy AI; thực sự có bao nhiêu người và giờ rà soát mỗi dây chuyền; và cách thí điểm chuyển đổi và người mua thanh toán. Trên các nhóm đó là cách các mô hình mới nhất, bản lượng tử hóa và cánh tay robot thực sự thể hiện trong điều kiện của một nhà máy cụ thể. Mỗi câu có thể biến thành một phép kiểm với ngưỡng rõ ràng. Các ngưỡng dưới đây là quy tắc quyết định tôi suy ra từ bằng chứng trong báo cáo này estimate ; hãy tính lại chúng khi dữ liệu nhà máy đầu tiên về.

Câu hỏi mở	Cách giải quyết	Kết quả tốt trông như thế nào
Giá và mức sẵn sàng trả cho rà soát AI	Một yêu cầu báo giá chuẩn (ví dụ bốn dây chuyền với AOI 2D cũ hỗn hợp và 3D mới hơn, hai ca, khoảng 2,000 tấm mỗi ca, 12 lớp khuyết tật, chỉ tại chỗ, tích hợp MES hoặc IPC-CFX) gửi cho vài nhà cung cấp bên thứ ba; một khảo sát độ nhạy giá bốn câu cho 20–30 nhà máy EMS	Hầu hết báo giá trong 40–120k RMB một lần hoặc 20–40k một năm cho bốn dây chuyền; ít nhất một nửa nhà máy chấp nhận 10–30k RMB mỗi dây chuyền với hoàn vốn trong 12 tháng
Cách các hãng AOI đóng gói AI và nâng cấp máy cũ	Phỏng vấn có cấu trúc ngắn với khoảng một tá hãng AOI và nhà phân phối	Liệu rà soát AI là chuẩn, tùy chọn hay vắng mặt trên mẫu 2025–26; liệu máy trước 2020 có nâng cấp được không và giá bao nhiêu
Tuổi của AOI đã lắp đặt và tỷ trọng chạy AI	Khảo sát tại chỗ khoảng 30 nhà máy EMS, phân tầng theo quy mô và khu vực, với danh sách thiết bị xem xét dưới NDA	Tuổi AOI trung vị ≥5 năm và ≥60% không AI nghĩa là nhóm nâng cấp là thật
Lao động và giờ rà soát	Phỏng vấn và 2–4 tuần nhật ký cảnh báo AOI, rà soát, sửa chữa và ca tại hai nhà máy trở lên, với lấy mẫu công việc	≥1 người rà soát chuyên trách mỗi dây chuyền mỗi ca và ≥3 giờ rà soát mỗi dây chuyền mỗi ca, với ít nhất 0.5 FTE mỗi dây chuyền nhà máy sẵn sàng điều chuyển
Hiệu suất của chính rà soát AI	Bộ dữ liệu vàng cộng chế độ chạy song song trên dây chuyền thực (Chương 3): thỏa thuận dữ liệu tuần 0, đường cơ sở và gán nhãn tuần 1–2, chạy song song tuần 2–4, chạy có trợ giúp tùy chọn tuần 4–8	Cảnh báo giả hoặc giờ rà soát giảm ≥50–60%; cận lọt lưới theo lớp không tệ hơn đường cơ sở người; không bỏ sót trên lớp trọng yếu; độ trễ p95 ≤½ takt; tự động mở-khi-lỗi sang người
Chất lượng nhãn thầy của mô hình frontier	Chạy hai hoặc ba ứng viên trên cùng bộ dữ liệu vàng, ba cách diễn đạt prompt mỗi cái, một ảnh tham chiếu tốt, ảnh cắt đơn linh kiện ≥42–48 px	Chỉ chấp nhận tự động cho lớp có không bỏ sót trên ≥299 khuyết tật và đồng thuận cấp người; nếu không chỉ gán nhãn trước và phân loại sơ bộ
Bản lượng tử hóa	Chỉ thay độ chính xác; ghim hash trọng số, template, tiền xử lý, giải mã, engine và kích thước batch; tắt hoặc đặt lại prefix caching; kiểm tra bản tham chiếu tái tạo kết quả nhà cung cấp; chấm điểm cặp trên bộ dữ liệu vàng và lát trôi	Không bỏ sót mới trên lớp trọng yếu; ≤1 bỏ sót mới trong 1,000 khuyết tật; tỷ lệ khép lại giảm ≤2 điểm; ≥99.5% quyết định giống hệt trong ba lần chạy; nếu không rơi về FP8/INT8
Phần cứng cạnh dây chuyền, độ sạch và ESD	≥2 chiếc mỗi hộp ứng viên stress-test ở gấp đôi tỷ lệ cảnh báo dự kiến; ngâm 72 giờ ở luồng khí tủ thực; kiểm ESD bởi điều phối viên khách hàng; kiểm ISO 14644-14 cho khách ISO 7 trở lên	p95 ≤½ takt; không treo hoặc giảm tốc trong 72 giờ; ký ESD; báo cáo hạt đúng cấp

Câu hỏi mở	Cách giải quyết	Kết quả tốt trông như thế nào
Điều khoản thanh toán, quy tắc thí điểm và chuyển đổi	Điều khoản thanh toán từ hợp đồng thiết bị kiểm tra gần đây tại các nhà máy khảo sát; phỏng vấn quy tắc thí điểm, tiêu chí nghiệm thu và thời gian phê duyệt	Khoảng 30/60/10 với tiền mặt trong 60 ngày sau nghiệm thu; thí điểm ≤8–12 tuần với một chủ thể ký duyệt có tên; ≥1 trong 3 lần thử chuyển đổi; khung bao phủ ≥5–10 dây chuyền trong năm đầu
Dữ liệu, IT và trách nhiệm	Phỏng vấn và rà soát hợp đồng với tư vấn pháp lý	Một hộp nội bộ có thể kết nối và truy cập dữ liệu; trách nhiệm giới hạn ở lỗi phần mềm
Tuân thủ khu vực trước khi ra nước ngoài	Đài Loan: trạng thái bằng văn bản của khách hàng theo Đạo luật Quản lý An ninh mạng, tự đánh giá E187; Mỹ: phán quyết hải quan về cách phân loại một cụm, báo giá NRTL, kiểm ITAR; Đông Nam Á: liên kết nội địa BOI và điều kiện phụ cấp Malaysia, công ty mẹ của khách hàng; đại lục: phạm vi CCC cho PC và nguồn đi kèm	Khách hàng không phải thực thể bị kiểm soát, và hoàn vốn giữ ≤2 năm khi đã gồm thuế và chứng nhận
Nghiệm thu một cụm robot hoặc hiện thân	Bảng nghiệm thu 12 hạng mục ở Chương 6: định nghĩa có ký, ≥3,000 chu kỳ sản xuất liên tiếp trong ≥3 ca và ≥2 mã linh kiện, tiêm lỗi kịch bản, ≥30 chu kỳ người bấm giờ, có chứng kiến	Cận dưới ≥99.5% nội tuyến (≥99.0% ngoại tuyến); chu kỳ ≤1.15 × người; ≤0.5 can thiệp mỗi giờ và không điều khiển từ xa; ≥95% thời gian hoạt động; MTTR ≤20 phút; ≥90% khôi phục
Chính sách ACT và VLA trên chip biên Trung Quốc	Bốn tuần: port và micro-benchmark (≥1,000 lần suy luận, phân bố độ trễ, công suất, ngâm 60 phút); mô phỏng vòng kín với độ trễ đo được tiêm vào; A/B robot thật với GPU; chạy bền 8–24 giờ (≥3,000 chu kỳ)	Đoạn ACT p95 ≤100 ms, VLA ≤300 ms; thành công robot thật trong 3 điểm so với GPU và chu kỳ ≤1.2 × người; nếu không giữ chính sách trên Jetson hoặc GPU
Giá cobot, AMR và tích hợp	Báo giá từng mẫu từ vài nhà phân phối cobot, và báo giá cụm đầy đủ từ hai hoặc ba nhà tích hợp cho nạp đồ gá kiểm tra (một trạm, hai ca)	Cụm đầy đủ ≤200k RMB và hoàn vốn ≤1.5 năm trong hai ca ở Thâm Quyến
Số dây chuyền SMT ở Đài Loan, Đông Nam Á và Mỹ	Dữ liệu IPC (trước đây New Venture Research) hoặc cơ sở lắp đặt của nhà phân phối thiết bị địa phương	Đủ dây chuyền có thể phục vụ mỗi khu vực để biện minh kỹ thuật ứng dụng địa phương

Trong số này, bốn câu quyết định các con số nhạy cảm nhất trong báo cáo: giá người mua sẽ trả, lao động rà soát AI thực sự giải phóng, liệu rà soát AI có thực sự không bỏ sót nhiều hơn người không, và tiền về nhanh đến đâu. Chúng đáng được giải quyết trước bất kỳ cam kết đáng kể nào, bởi một nhà máy đang cân nhắc mua hoặc bởi một nhà cung cấp đang cân nhắc sản phẩm.

Một trình tự thí điểm hiệu quả cho cả hai bên

Với một nhà máy, và với một nhà cung cấp muốn có tham chiếu đáng tin cậy, thứ tự là như nhau:

1. **Tuần 0–2: thống nhất dữ liệu và định nghĩa.** Ký thỏa thuận dữ liệu, định nghĩa các lớp khuyết tật và thế nào là một lần lọt lưới, và ghi lại đường cơ sở: cảnh báo giả mỗi tấm, giây rà soát mỗi lần gọi, lọt lưới phát hiện ở hạ nguồn.
2. **Tuần 1–4: xây bộ dữ liệu vàng** từ kho ảnh AOI và các tấm mạch NG được giữ lại, gán nhãn mù và phân xử (Chương 3).
3. **Tuần 2–4 (hoặc lâu hơn): chế độ chạy song song.** Mô hình ghi lại quyết định của nó trong khi con người quyết định như trước; so sánh theo từng lớp.
4. **Tuần 4–8: chạy có trợ giúp trên các ca ngẫu nhiên,** với mô hình chỉ khép lại các lớp đã qua cổng và mọi thứ khác chuyển cho người.

5. **Nghiệm thu, bằng văn bản, trước khi mở rộng:** kết quả theo lớp, độ trễ, hành vi mở-khi-lỗi, và ai sở hữu mô hình và nhãn. Mở rộng từng dây chuyền, và tái xác thực trên mọi sản phẩm, phiên bản mô hình, lượng tử hóa hoặc engine mới.

Giới hạn của báo cáo này

Đây là nghiên cứu bàn giấy. Tôi không tiến hành phỏng vấn gốc nào với các nhà máy cho ấn bản này; phần lớn dữ liệu dây chuyền đến từ các nhà cung cấp và chưa được kiểm toán độc lập. Một số trang web chính phủ, cơ sở dữ liệu gói thầu và nền tảng tuyển dụng không truy cập được, nên các mức trợ cấp nâng cấp 2026 của Thâm Quyến, hướng dẫn lương Đồng bằng Dương Tử, giá giao dịch rà soát AI, số người rà soát mỗi dây chuyền và tỷ lệ chuyển đổi thí điểm là ước tính hoặc suy diễn; mỗi cái xuất hiện ở trên như một câu hỏi mở. Tỷ giá là tỷ giá trung tâm của Ngân hàng Nhân dân Trung Quốc ngày 9 tháng 11 năm 2026 ($1 = 6.7188 RMB) và, cho Đài Loan, tỷ giá cố định của SMBC cùng ngày ($1 = 32.12 NT$); đồng Việt Nam dùng dữ liệu tháng 5 năm 2026. Mọi ngưỡng đều là quy tắc quyết định suy ra từ bằng chứng được trích dẫn, không phải tiêu chuẩn.

Về tác giả

Tan Shuai

Tan Shuai là một kỹ thuật gia và nhà sáng lập với hơn hai mươi năm làm việc trải dài qua chip, phần cứng và phần mềm hệ thống, cùng 20 bằng sáng chế đã được cấp. Công việc của ông đã đưa ông qua Thung lũng Silicon, New York, Chicago, Đài Loan, Khu vực Vịnh Lớn (Thâm Quyến và Hồng Kông), Việt Nam và Malaysia, với trọng tâm lâu dài vào sản xuất tiên tiến và chuỗi cung ứng. Ngày nay ông làm việc như một nhà nghiên cứu và phát triển AI, tập trung vào cách AI thực sự được sử dụng trên sàn nhà máy.

Ông hoan nghênh trao đổi với những người làm AI trong sản xuất, dù là xây đội ngũ, tìm đối tác hay đối chiếu ghi chú: i@tanshuai.com · tanshuai.com

Ghi chú và tuyên bố miễn trừ trách nhiệm

Báo cáo này là nghiên cứu bàn giấy độc lập dựa trên các tài liệu công khai có đến ngày 11 tháng 11 năm 2026. Nó được cung cấp chỉ để tham khảo và không phải tư vấn đầu tư, pháp lý, thuế hay kỹ thuật. Tên sản phẩm, giá, phiên bản mô hình, phân loại xuất khẩu, thuế quan và quy định thay đổi nhanh chóng; hãy xác minh bất cứ điều gì bạn dựa vào với nguồn gốc. Các con số đánh dấu vendor claim hoặc estimate nên được đọc như vậy. Tên công ty và sản phẩm thuộc về chủ sở hữu của chúng và chỉ được dùng để nhận diện; không công ty nào rà soát hay tài trợ cho báo cáo này.

© 2026 Tan Shuai. Bạn có thể chia sẻ báo cáo này nguyên văn kèm ghi công.

SẢN XUẤT ĐIỆN TỬ · KIỂM TRA AI × ROBOT

Kiểm tra thị giác bằng AI và cánh tay robot trong các nhà máy điện tử: Đánh giá 2026

Mỗi con số then chốt đều kèm nguồn và nhãn mức độ tin cậy: verified, single source, vendor claim hoặc estimate. Những gì hồ sơ công khai không thể trả lời được tập hợp ở Chương 7, cùng với cách một thí điểm giải quyết từng câu hỏi.

Tan Shuai
tanshuai.comi@tanshuai.com

ẤN BẢN CÔNG KHAI · DỮ LIỆU TÍNH ĐẾN NGÀY 11 THÁNG 11 NĂM 2026

