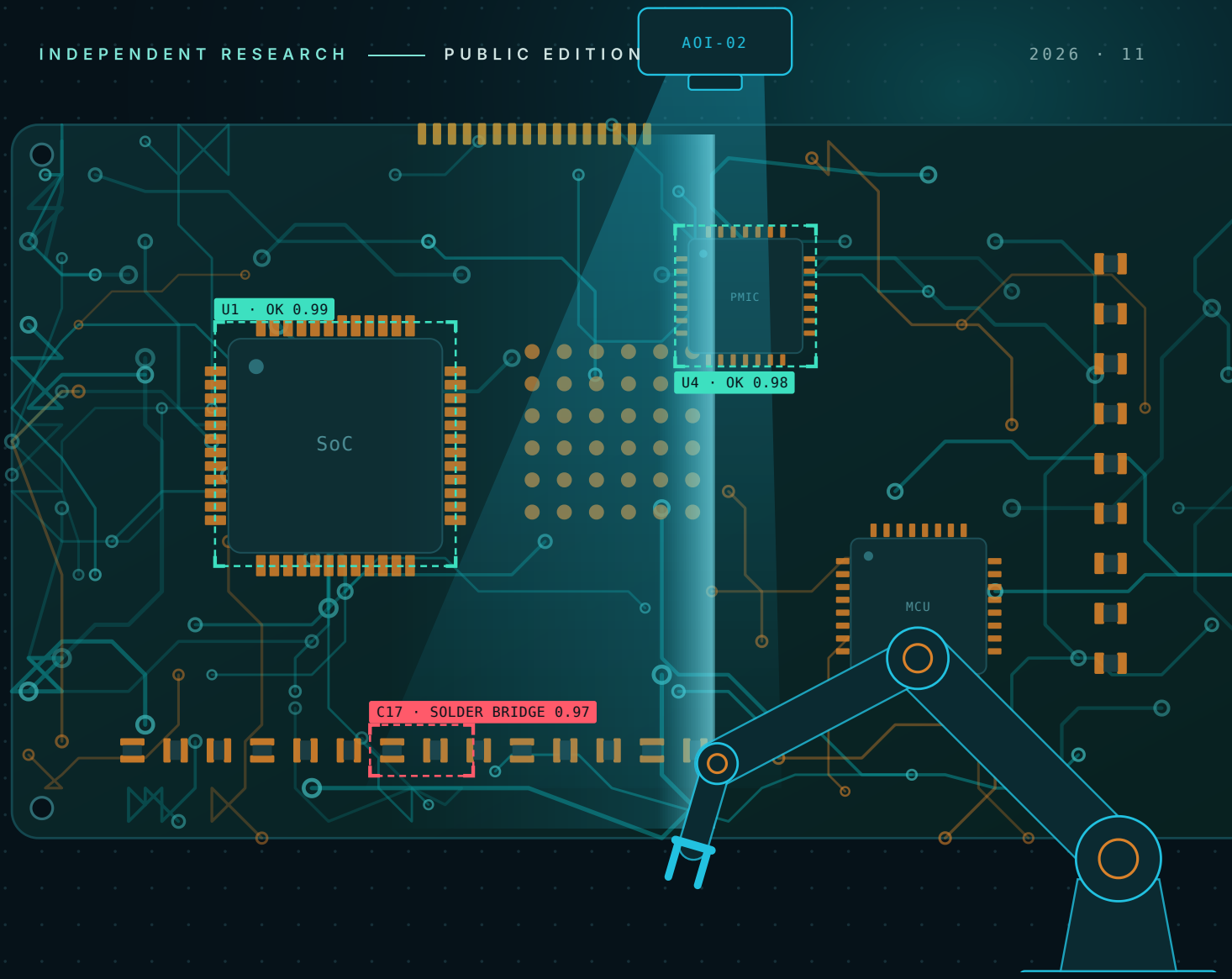


ELECTRONICS MANUFACTURING · AI INSPECTION × ROBOTICS

AI Visual Inspection and Robot Arms in Electronics Factories

2026 Assessment

Mainstream configurations · Regional feasibility · On-premises deployment · Teacher–student models
Return on investment · Competition · What only a pilot can answer

KEY FINDING **AI review is worth doing, narrowly. Embodied robot arms can wait.**

Tan Shuai
tanshuai.com • i@tanshuai.com

Data as of November 11, 2026
Mainland China · Taiwan · Southeast Asia · United States

CONTENTS

Contents

◆ **Executive summary** — Worth doing, but narrowly, now and with partners — 4

01 **The line today** — What electronics factories run today, and what is technically feasible — 7

02 **On-premises** — Running AI inside the plant: clean rooms, air gaps, hardware and quantization — 15

03 **Teacher and student** — Teacher and student: frontier models can pre-label inspection data; for robot arms the real teachers are still people — 24

04 **Costs and payback** — Costs and payback: AI review pays back in months, machine tending in years, pick-out arms never — 33

05 **Market and competition** — Market, competition and how buyers pay — 40

06 **Embodied AI** — Embodied AI: first it changes how robots are taught, later how they move — 49

07 **Open questions** — What the public record cannot answer, and how a pilot settles it — 55

◆ **About the author** — Tan Shuai — 57

Confidence tags in the text

verified	two or more independent sources, or two independent primary documents, agree
single source	one source only, including primary documents
vendor claim	published by the seller of the product or service
estimate	the author's arithmetic or inference from cited data

How to read this report

Key numbers carry a confidence tag. verified means two or more independent sources agree, or two independent primary documents agree; single source means only one source (including primary documents); vendor claim means published by the seller of the product or service, even if widely repeated; estimate means my own arithmetic or inference from cited data, with the inputs sourced. "Primary" describes the type of source (annual report, prospectus, regulation, standard, official statistics); it does not change the confidence class. Where sources conflict, the figures are given side by side and not averaged. Currency conversions use the People's Bank of China central parity of November 9, 2026, $1 = RMB 6.7188 (Xinhua), and for Taiwan SMBC's fixing on the same day, $1 = NT$32.12 (SMBC). Mainland labour and subsidy data use the Pearl River Delta, where electronics manufacturing is most concentrated (official wage data from Shenzhen), with the Yangtze River Delta for comparison. Claims from industry interviews are treated as claims to be checked, never as facts. Questions the public record cannot answer are collected in Chapter 7.

FIGURES

Figures



01	Where AI fits on the factory floor today	6
02	A volume SMT line: the people sit behind the AOI	8
03	About 97–99% of AOI alarms are false calls	9
04	A general-purpose camera is not an AOI	10
05	An air-gapped deployment	15
06	Clean and ESD zones decide where compute can sit	16
07	Speed limits by use	18
08	On-prem compute: memory, price and what can legally be bought in mainland China	19
09	One big workstation does not make a shared assistant	20
10	Quantization: 8-bit is safe, 4-bit depends on the recipe, 3-bit and below is out	21
11	Teacher candidates: generic scores doubled in a generation	25
12	"Is there a defect?" is still hard for general models	26
13	Two teacher–student pipelines	27
14	How big the golden set must be	28
15	Edge policy latency	30
16	Robot students: near production grade only when the station is structured	31
17	What one inspection or handling worker costs a year	34
18	The price ladder	35
19	Payback by region	37
20	Why picking out bad boards does not pay	38
21	US duty on robots, AOI and edge computers	39
22	Market sizes differ by more than ten times	41
23	Bottom-up: 40,000 lines → 16,000 serviceable → ≈RMB 320M one-off	42
24	The retrofit window is about five years	43
25	The review wedge is encircled on three sides	43
26	The margin sits in productized hardware	45
27	How buyers pay	46
28	Days sales outstanding	47
29	The embodied-AI stack	49
30	What in-plant robot deployments actually disclose	50
31	Net labour replaced per robot	51
32	Standards and regulation timeline	52

EXECUTIVE SUMMARY

Worth doing, but narrowly, now and with partners

KEY FINDING

For electronics factories, what is mature in 2026 and what plants will pay for is AI review and niche inspection on AOI images, plus programmed robot machine tending, not "a large model looking at boards plus embodied robots picking them".

0.46–0.65 yr
AI review payback in the US, Taiwan and Shenzhen, at RMB 50k per line

≥95% · <20 PPM
Coverage and escapes of a top EMS's in-house AI review: the bar for selling to tier-1 EMS

100–700 RMB M
Serviceable third-party retrofit market in mainland China, one-off, over a window of about five years

0 scores
Published industrial-defect benchmark scores for the November 2026 flagships: teachers, not inspectors

1.1–1.5 yr
Payback of a cobot machine-tending cell over two shifts in the US, Taiwan and Shenzhen

0.003–0.14 FTE
Labour a pick-out robot arm saves per shift on an SMT line: it never pays back

Inspection
ROI

Supervised deep learning applied to the images an existing AOI already produces cuts false calls or review hours by 60–90% (vendor claims). DBG, a large EMS company, reports in its annual report that its in-house system covers more than 95% of flagged components with escapes below 20 PPM and more than 85% less human review; that is the bar a third-party product must clear to sell to tier-1 EMS. At RMB 50,000 per line, AI review pays back within a year in mainland China, Taiwan and the US (Shenzhen 0.65 years, Taiwan 0.53, US 0.46), cobot machine-tending cells in 1.1–1.5 years and Southeast Asia in 2–6 years, while a robot arm added to an SMT line only to pick out bad boards never pays back anywhere, because above 90% first-pass yield picking takes only 0.003–0.14 of a person per shift estimate .

Models
ROBOTS

Multimodal large models still cannot be the primary inspector. As of November 11, 2026, none of the flagships (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 and others) has a published score on any industrial-defect or PCB benchmark, and in the ACM MM 2026 PCBA challenge the top five systems, trained on 6,600 real line images, scored only 77–82 on "is there a defect". Frontier models can pre-label and triage. For robot arms, the real teachers are still human demonstrations, real-robot reinforcement learning and simulation: students reach 90–99.4% at fixtured, structured stations, but only 30–80% on moving conveyors, unseen parts and new products; a frontier agent completed 104 of 108 quasi-static real-robot tasks, at 280 seconds each.

On-premises
DEPLOYMENT

Air-gapped deployment in clean, controlled plants is fully practical. An ordinary SMT hall is usually an ESD protected area rather than a graded clean room: a fanless Jetson or domestic NPU box goes at the line and GPU servers go in the server room. Under NVIDIA's own export classifications, Jetson Thor and DGX Spark can be sold to an ordinary mainland factory without a US li-

cence; DGX Station cannot. Precision defaults to 8-bit; vendor quantization-aware 4-bit is usable with conditions; 3-bit and below is out.

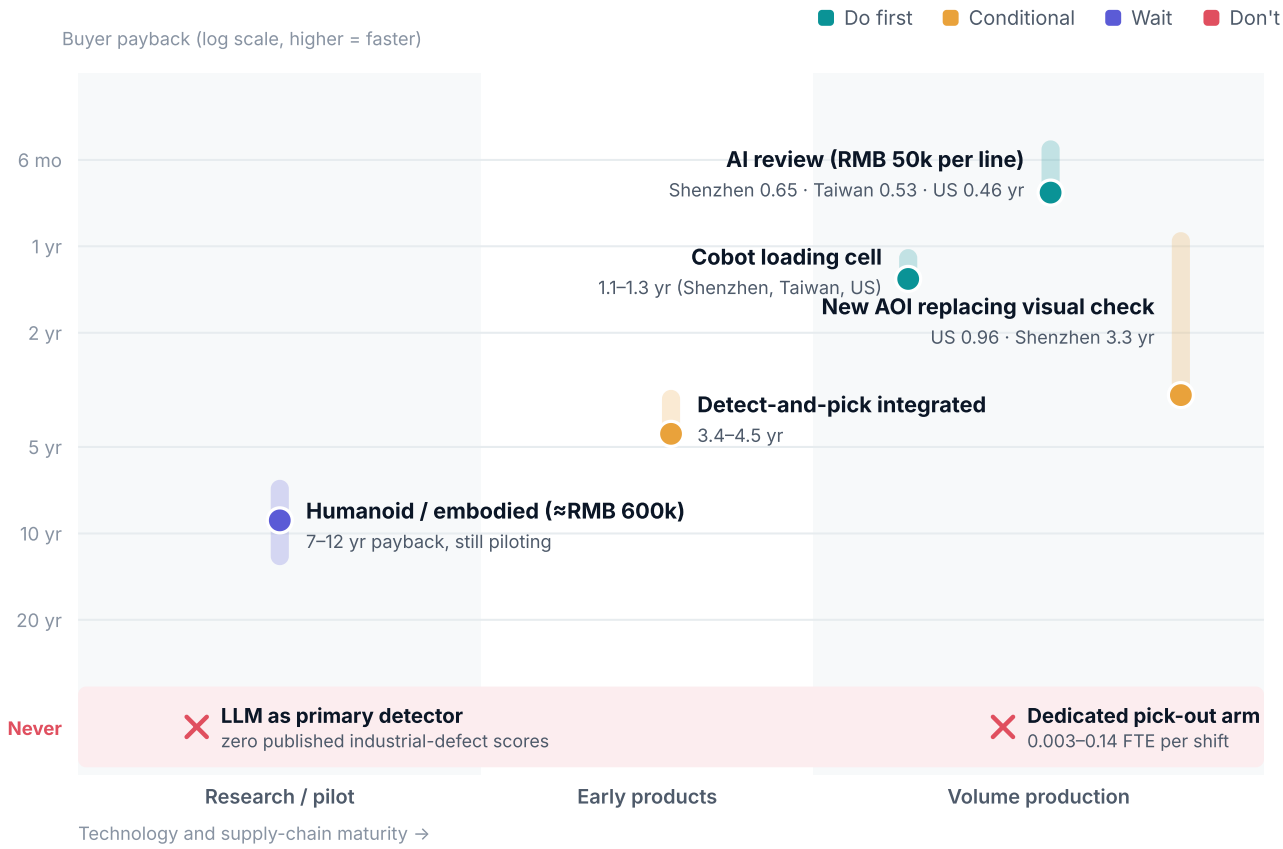
Market
COMPETITION

The market is real but limited. Mainland China has about 35,000–46,000 SMT lines and about 26,000–66,000 inline AOIs not running AI; the serviceable market for third-party retrofits is about RMB 100–700 million one-off, over a window of about five years, and AOI makers have started building review into new machines estimate. The competition is already strong: the review wedge is encircled by AOI makers, in-house EMS teams and platform vendors, and the margin sits with productized hardware.

Question	Judgement	Key numbers (sources and confidence in the chapters)
① **What do mainstream electronics plants run today?**	A volume SMT line is "3D SPI + pre- and post-reflow AOIs + sampled X-ray + ICT/FCT + repair and visual inspection", with human review behind every AOI. AI enters review three ways: built in by AOI makers, built in-house by top EMS, or added by third parties. Assembly and handling use arms, SCARAs and AMRs with learned perception and programmed motion; embodied AI is still piloting	Takt ≈13–22 s per board; one ODM budgeted RMB 30.6M of AOI for 18 lines, ≈3–5 AOIs per line; ≈97–99% of AOI alarms are false calls
② **How far does the technology go, can money buy zero people, and what can each region use?**	Inspection is industrial-grade as "dedicated detector + AI review + a few people"; even unlimited money does not reach zero people. The mainland uses only domestic models or local open weights; private plants in Taiwan can use nearly anything, but fabs and critical infrastructure exclude mainland brands in practice; Southeast Asia has the fewest restrictions; the US has the full stack plus tariffs and certification	AI review cuts false calls or review hours 60–90% vendor claims ; a 4K camera over a whole panel ≈78 µm per pixel against 8–20 µm for an AOI; mainland-made arms pay ≈40% combined US duty, AOI-type instruments ≈37.5%
③ **Can it run on-premises in clean rooms and air-gapped plants, on which hardware, fast and accurately enough?**	Yes; the constraints are procedural, not compute. A ≤100 W fanless box at the line, GPU servers in the server room; Thor for stations, a Mac Studio or domestic appliance for plant Q&A no DGX Station on the mainland. Speed is fine when model size matches the task; keep 8-bit, use vendor quantization-aware 4-bit with conditions, never 3-bit	≈0.5 s per crop for an 8B vision model on Thor estimate ; Thor dev kit RMB 59,104 in China; DGX Station $123–140k, 139 tokens/s in total for 32 users on a 743B model
④ **Can the newest flagship act as a teacher over dozens of rounds and be distilled into local students, including for flexible manufacturing and robot arms?**	Inspection: pre-labelling and triage; auto-accept teacher labels only for classes with "≥299 defects, zero misses" on a golden set. Robots: the teachers are still human demonstrations, real-robot RL and simulation, and students approach production grade only at structured stations	No November flagship has an industrial-defect score; students 97.8% on trained connectors, 29–67% on unseen ones, 37–59% on moving conveyors; frontier models driving joints directly 0–5.5%
⑤ **What does it cost, and what is the ROI by region?**	AI review pays back within a year in mainland China, Taiwan and the US; cobot tending 1.1–1.5 years; Southeast Asia 2–6 years; a pick-out arm on an SMT line never pays back	RMB 50k review box per line: Shenzhen 0.65 years, Taiwan 0.53, US 0.46, Vietnam 2.2 estimate ; acceptable third-party add-on price ≈RMB 10–30k per line one-off or RMB 5–10k a year
⑥ **Are competitors already strong enough, and is it better to start now or wait?**	Demand is not saturated, but supply is crowded; OEM AI limits the retrofit window to about five years. Enter now with a narrow product through channel partners, without fighting platform vendors head-on; wait for 2027–2028 production data before betting on embodied AI	≈26–66k inline AOIs without AI on the mainland, shrinking 8–12% a year estimate ; the leading AI vision vendor had 2025 revenue of RMB 1.086B and a net loss of RMB 991M; the closest listed comparable's 3C revenue fell 79.9% year on year in January–May 2026

FIG 01 **Where AI fits on the factory floor today**

AI review of AOI false calls and cobot machine tending are mature and pay back fast; a dedicated pick-out arm and an LLM as primary inspector are not worth doing.



Source: Payback from the author's model (Chapter 4); maturity axis is the author's judgement.

What to build and what not to build

	Recommendation
Build	AI review and defect classification on AOI images, for mid-size EMS with older or mixed-brand AOIs and no AI upgrade from the AOI maker; inspection the SMT AOI covers poorly (through-hole and hand soldering, conformal coating, cables, connectors and FPCs, final appearance and screws); cobot loading and unloading cells where every part is handled (test fixtures, trays, packing), with inspection as an add-on; a golden set and shadow mode on a real line as the basis of every claim
Don't build	A general "AI plus robot arm" platform; a multimodal LLM as the primary inspector; a robot arm on an SMT line only to pick out NG boards (an NG diverter or buffer does it); humanoids and robots-as-a-service; frontier models inside a control loop or a safety function; piles of GPU memory at stations to run "full-size" models; direct sales to fabs, Taiwanese critical infrastructure or US defense electronics plants
Who to work with	AOI makers and distributors (channel and image interfaces; distributors brought in 46.68% of Sinic-Tek's 2025 revenue); cobot makers (certified arms and safety functions); EMS companies with 10–40 lines (rolling expansion and deposits); local integrators in Thailand and elsewhere (to meet BOI local-linkage rules)

THE LINE TODAY

01 What electronics factories run today, and what is technically feasible

The standard volume line: several AOIs, sampled X-ray, central review, 13–22 seconds per board

Primary filings from listed ODMs show what a 2025–2026 volume line actually looks like. Longcheer's Hong Kong prospectus (data to September 30, 2025) lists its SMT quality checks as "X-ray inspection, solder paste inspection (SPI), automated optical inspection (AOI), PCBA functional test and visual inspection"; its process flow runs pre-reflow AOI, post-reflow AOI, sampled X-ray, NG boards to an SMT repair station, then a final visual check. Its Huizhou, Nanchang and Vietnam plants have 34, 24 and 8 SMT lines and 67, 60 and 13 final-assembly lines (Longcheer H-share prospectus) single source, primary filing. Huaqin's 2023 laptop plant planned 18 SMT lines with an AOI budget of RMB 30.6 million, about RMB 1.7 million per line, which at domestic 3D AOI prices is three to five AOIs per line for pre- and post-reflow inspection of double-sided boards (Huaqin A-share prospectus) primary filing + estimate. A volume line therefore usually has two to four review points, not one.

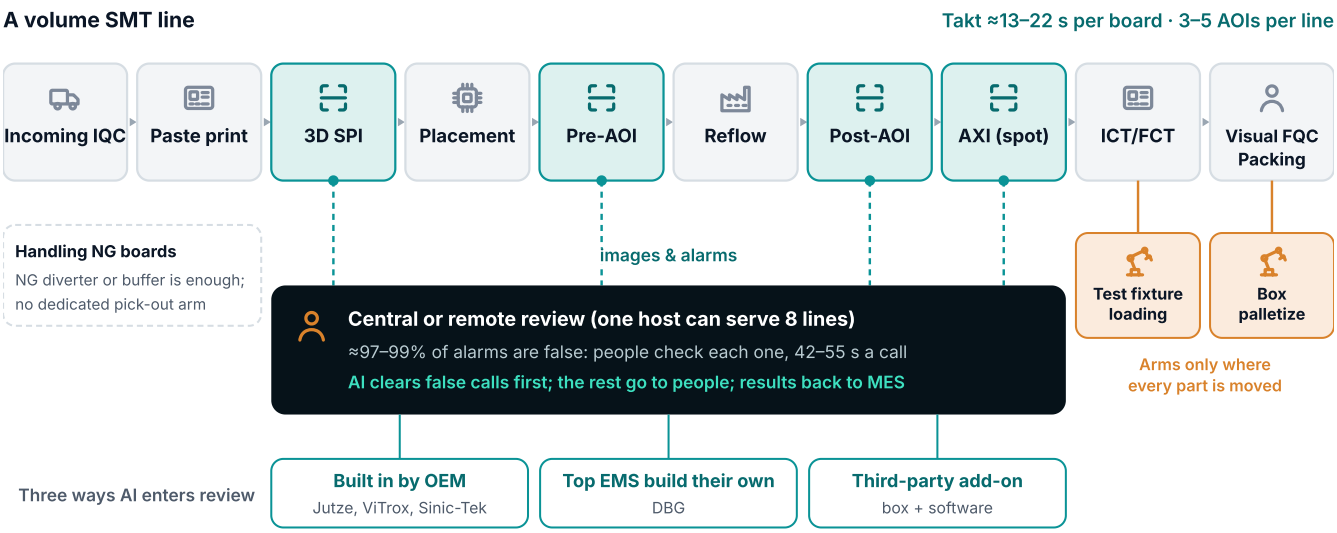
The machines are already highly automated: Huaqin reports 96% automation on its SMT lines (Huaqin H-share prospectus) single source. The remaining people are in review, repair, visual inspection, test and final assembly, and no company publishes how many reviewers it has per line (Chapter 7).

Three independent estimates agree on a takt of **13–22 seconds per board** (a 2026 environmental filing for five Hebei lines (Hebei SMT project EIA), Huaqin's laptop lines and an iFactory case), or 7–11 seconds per side for double-sided boards estimate. The hard part of a high-mix line is not volume but new product introduction and changeover, and every changeover is another cold start for an AI model.

On the software side, the AOI makers' own review stations are already centralized: Sinic-Tek's repair-station software lets "one host control the equipment of eight SMT lines" (Sinic-Tek prospectus) single source, primary, 2022, and after Jabil deployed ViTrox V-ONE remote AOI control in 2026, one engineer could manage 5–8 AOIs remotely against 1–2 locally (AtlasPCB) vendor claim. A third-party AI review product plugs into exactly this central station, through the AOI's image and result export (IPC-CFX, IPC-HERMES-9852 or a vendor's own format). The AOIs themselves are adding AI fast: Yamaha says its AI judgement takes about one second per call, helps inspectors after 30 NG images and becomes "very high" quality after 1,000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **A volume SMT line: the people sit behind the AOI**

The labour is concentrated in the review seats after automated optical inspection. That is where AI enters.



Source: Public filings and vendor material (Huaqin, DBG, Sinic-Tek and others); compiled by the author.

Stage	Mainstream products and technology	Representative suppliers	How far AI has got
After printing	3D SPI (structured light, moiré, some laser triangulation)	Koh Young, Sinic-Tek, TRI	OEM AI paste judgement in production; Jutze says some models are “fully AI”
After placement and reflow	2D/3D AOI with multi-angle, multi-colour LEDs and structured light, 8–20 µm per pixel, 8–12 MP cameras	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek and others	OEM AI review (Jutze in production, ViTrox V510Ai, Sinic-Tek with Tencent); top EMS build their own (DBG); third-party add-ons
Bottom-terminated and hidden joints	Sampled or inline X-ray (AXI)	—	Hikvision claims its Guanlan model misjudges X-ray voids less than 1% of the time vendor claim
Review and repair	Central review station (one host, eight lines), remote review, repair station	AOI makers’ software, ViTrox V-ONE	AI closes false calls, people get the rest
Visual inspection, IQC, FQC	Mostly manual; some AI appearance machines with robot loading	SmartMore and others	Appearance machines in production; no public headcount data for IQC/FQC
Handling and machine tending	Six-axis arms and cobots, SCARA, AMRs, 2D/3D vision guidance	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Learned perception plus programmed motion in production; VLAs and humanoids still piloting
Data and systems	MES, IPC-CFX, IPC-HERMES-9852	—	Results written back to MES; support in older equipment is uneven

Sources for the table: the filings above and (Koh Young, Jutze 2025 annual report, Sinic-Tek 2025 annual report, ViTrox, Sohu on Hengtianyi, Xinhua on Hikvision).

The people sit behind the AOI, because most alarms are false

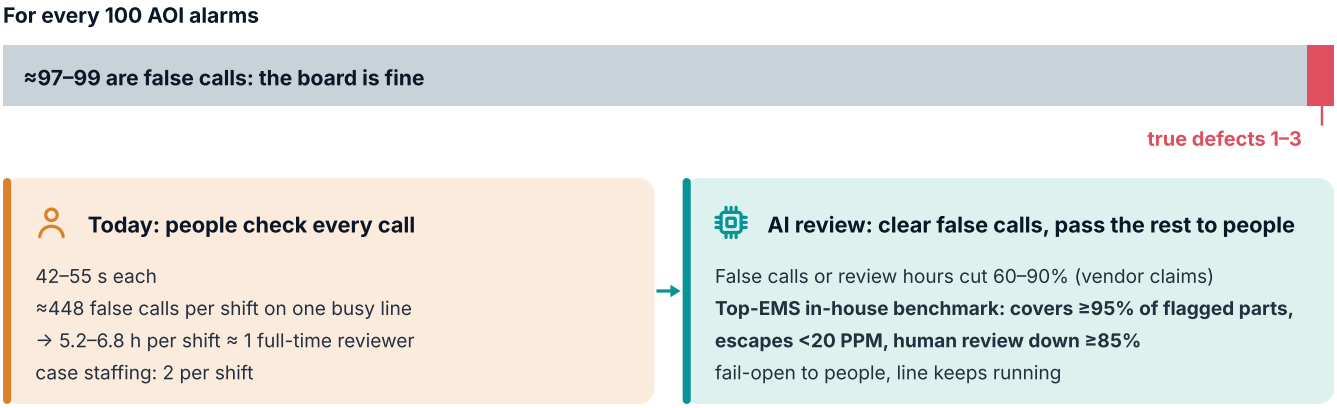
The inspection chain on a mainstream SMT/PCBA line runs: 3D SPI after printing for missing, insufficient or excess paste, bridging and offset; AOI after placement (optional) and after reflow for missing, shifted, rotated, reversed or tombstoned parts and solder volume; X-ray when there are BGAs or other bottom-terminated parts; then in-circuit or flying-probe test, functional test and manual visual inspection. Every AOI has a review seat and a repair station

behind it (Sinic-Tek prospectus, JLCPCB) verified. Post-reflow AOI is the most common and best value, but it sees only shape and area; joints, voids and wetting under a BGA need X-ray or electrical test (JLCPCB).

The labour sits in review because most alarms are false. One UK EMS recorded **54,019 false calls against 544 real defects across 917 part numbers, about 99% of alarms** (arXiv 2505.03576) single source, preprint; at Siemens' Rastatt plant **97.5% of the boards the AOI rejected were good** (Siemens case study) vendor claim. A busy line sees about 448 false calls per shift at 42–55 seconds each, 5.2–6.8 hours per shift, roughly one full-time reviewer (iFactory) vendor claim; hours are an estimate.

FIG 03 **About 97–99% of AOI alarms are false calls**

Today a person confirms every call; AI review clears the false ones first and passes the rest to people.



Source: Siemens case study, iFactory (vendor claim), DBG 2025 annual report; review hours estimated by the author.

From "review hours per shift = boards × false calls per board × seconds per call", one reviewer can cover about 0.7–1 volume line with a rule-based AOI. Once AI removes 60–90% of false calls, judgement time alone would allow 2–14 lines per person, but walking, NG buffer handling and repair hand-offs limit local coverage to two to four adjacent lines; only central or remote review gets close to the theoretical ceiling estimate; seconds per call come from one source. Since mainstream AOIs scan 60–80 cm² per second (Yunheng 2026 AOI guide), a 20×15 cm side takes 4–5 seconds; the AI only has to process the few flagged regions within one takt, not rescan the board.

What already works in production: supervised classification on the AOI's own images

The proven pattern is supervised deep-learning classification on images from the AOI a plant already owns. Fujitsu's add-on on ADLINK's AOI in Taiwan cut false calls by **more than 90%** (Fujitsu case); OKI has used AI since July 2026 to cut manual confirmation after AOI by about **80%** in its own EMS business (Circuits Assembly); Siemens reports 60% less manual analysis, €160,287 saved in the first year and an eight-month payback (Siemens case study) all vendor claims. Two things are common to these cases: almost none publishes an escape rate, and all of them keep people on the remaining alarms. The one case that gives an escape figure (0.8% down to 0.3%, against a contractual 0.5%) is internally inconsistent (iFactory). AI cuts review hours, not the review job itself.

AI review has moved from third-party add-ons to OEM features and in-house systems, which changes what a newcomer is measured against. Jutze's 2025 annual report lists its "3D AOI upgrade" as in volume production and says it "uses AI algorithms to remove manual review at the inspection end" (Jutze 2025 annual report); Sinic-Tek has developed a "multimodal AI-assisted manual review system" since 2019 and is co-developing a top-and-bottom 3D AOI with Tencent, targeting 99.5%+ defect coverage and 98% recognition accuracy (Sinic-Tek 2025 annual report) vendor claims in primary filings. Large EMS companies build their own. DBG's 2025 annual report lists "AI judgement for visual inspection" as a completed project that "covers more than 95% of AOI-flagged components", with an "escape rate below 20 PPM" and "more than 85% less manual review", fully traceable (DBG 2025 annual report) single source, primary. That is the only electronics-plant figure I found that gives coverage, escapes and reduction together. **A**

third-party product sold to tier-1 EMS has to match DBG's 95% coverage, 20 PPM escapes and 85% reduction; below tier 1, the benchmark is the 50–85% reduction vendors advertise.

Anomaly detection trained only on good parts is fast (EfficientAD-S takes 2.2 ms per image (EfficientAD, WACV 2024)) but weak on "logical" anomalies such as missing or wrong parts, and the best methods average under 60% AU-PRO on the harder MVTec AD 2 (MVTec AD 2) single source. Missing and wrong parts are better checked against CAD and BOM.

The ceiling at any cost: layers of machines, very few people, but not zero

I found no electronics plant that removed manual review entirely and reached zero escapes. Practice for IPC-A-610 Class 3 (high-reliability) products still has trained inspectors decide the borderline calls automatic thresholds cannot classify, and one PCB assembler calls "automated inspection can replace human sign-off" a "common misconception" (PCBCart) single source. The human baseline needs correcting too: citing a 2012 Sandia study, a single visual inspector at best catches about **80%** of defects, two in series about 96% (Instrumental) single source, cited by a vendor. The defensible claim is "AI plus a small human review loop beats all-manual inspection", not "AI replaces inspection".

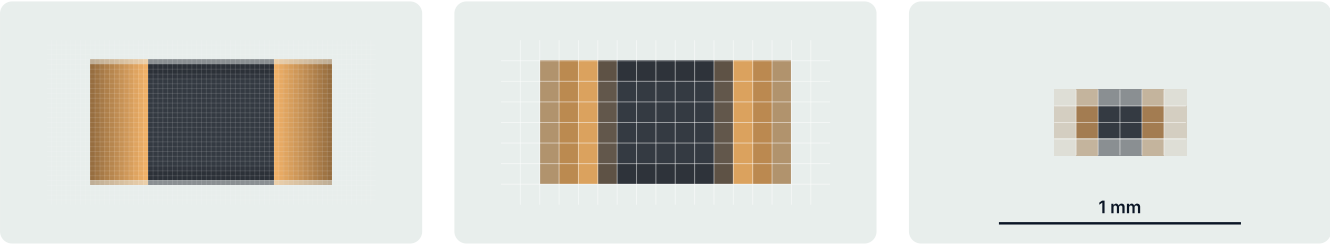
With no budget limit, the ceiling is 3D SPI, 3D AOI, X-ray, AI review and AI appearance stations stacked together, with people left only for borderline calls, new products and changeovers, new defect types and customer sign-off. Flagship plants report 85–92% less manual inspection (I-Connect007, 2021, WiAdvance) vendor claims; my estimate of the no-budget ceiling is review labour at 8–15% of today's estimate. Final assembly is still labour-intensive: automating the iPhone 16 buttons was dropped because of high defect rates (9to5Mac) single source.

A general high-resolution camera is not an AOI

Production AOIs image tile by tile at 8–20 µm per pixel, with 8–12 MP cameras, programmable multi-angle multi-colour lighting and structured-light 3D (Koh Young, Saki) verified. The industry noted long ago that an 01005 part covers only about 8×16 pixels at 25 µm per pixel, "possibly not enough" for reliable judgement (Circuits Assembly, 2007). By the same arithmetic estimate, a 4K camera across a 300 mm panel gets about **78 µm per pixel**: an imperial 0402 part is about 13×6 pixels, enough for presence and gross offset but not joint shape, and an 01005 is about 5×3 pixels, effectively invisible. Reaching 15 µm per pixel over a whole panel would take about 330 megapixels, which is why AOIs move a camera over the board. When someone says "0402", check which one they mean: metric 0402 is imperial 01005, far beyond any wide-angle camera.

FIG 04 **A general-purpose camera is not an AOI**

A 4K camera over a 300 mm panel gives about 78 µm per pixel: an 0402 part shrinks to ≈13×6 pixels and an 01005 disappears.



AOI: 8–20 µm per pixel
0402 ≈ 100×50 px

4K camera over a 300 mm panel: ≈78 µm/px
0402 shrinks to ≈13×6 px: enough for presence and gross offset only

Same camera on an 01005
≈5×3 px: effectively invisible

Source: Circuits Assembly; author's calculation for a 300 mm panel and a 4K sensor. Pixel grids are illustrative.

On lines that already have AOIs, the best place for AI is on top of the AOI's images. Only lines without AOI, or end-of-line stations, justify an enclosed camera station with controlled lighting; an open high-resolution camera over the line is the most exposed to ambient light and shadows (CSDN).

Assembly and handling: learned perception plus programmed motion is in production

The IFR's 2026 report says "machine vision for inspection and robot guidance is mature and industrial-grade", while "general-purpose physical AI and learning from simulation and experience are not yet industrial-grade" (IFR World Robotics 2026 executive summary) single source, primary . Electronics is the largest robot buyer (142,235 units installed worldwide in 2025, 59% of all installations in China (IFR, IFR China release)), yet China has only 166 robots per 10,000 manufacturing workers against 307 in the US (IFR, April 2026). The pattern in production is learned perception plus deterministic motion; end-to-end learned motion is near production grade only at structured stations (Chapters 3 and 6).

What each region can use

Mainland China relies on domestic models and local weights; in Taiwan the customer's identity decides; in the US, tariffs and certification do.

Element	Mainland China	Taiwan (private plants)	Southeast Asia	United States
Claude (any channel)	Not available	Conditional: the supplier must not be majority-owned or controlled by persons or entities in mainland China, Hong Kong or Macau	Same	Same
GPT and Gemini cloud	Not available	Available	Available (lists not checked country by country)	Available
Chinese cloud models (Qwen, DeepSeek, Doubao, GLM, Kimi)	Available (filed models; offering a service requires registration)	Legal, but mainland hosting rarely passes customer security review; banned in the public sector, state enterprises and critical infrastructure	Legal (international nodes)	Legal, strict review
Local open weights of Chinese origin (Qwen, GLM, DeepSeek)	First choice; private in-plant deployment needs no model or algorithm filing (law-firm reading)	Available (public sector restricted)	Available	Conditional: origin review
Local open weights of non-Chinese origin (e.g. Gemma 4)	Available	Available	Available	Available
Classic non-generative vision models	Available (outside the generative-AI rules)	Available	Available	Available
Jetson Thor/T5000/T4000, DGX Spark	Available: NVIDIA classifies them 5A992.c, no US licence needed for an ordinary civilian plant; Spark has a China SKU; channel prices about 40–60% above US list	Available	Available	Available

Element	Mainland China	Taiwan (private plants)	Southeast Asia	United States
RTX 5090, RTX PRO 6000, DGX Station	🔴 Not available: the first two are 4A090.a, DGX Station 5A992.z; legal alternatives are 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	🟢 Available	🟢 Available, but buyers headquartered in or ultimately owned from China count as Chinese; Malaysia requires permits for export and transshipment	🟢 Available
Mac Studio M5 Ultra 512 GB	🟢 Available (Apple China store)	🟢 Available	🟡 Presumed available	🟢 Available
Huawei Ascend and other Chinese accelerators	🟢 Available	🔴 High risk	🔴 High risk	🔴 Not usable in practice
Mainland-made fixed arms, AOIs, industrial cameras	🟢 Available	🟡 Importable; since December 1, 2025 the authorities can act directly on mainland brands, and agencies, state enterprises, certain foundations and critical-infrastructure providers can be restricted or barred from using them	🟢 Available	🟡 Legal, but combined duty is about 40% on arms, 37.5% on AOI-type instruments and cameras, 25% on edge computers
Foreign-made AMRs and mobile manipulators	🟢 Available	🟢 Available	🟢 Available	🟡 New models need conditional approval from the US Department of War
Hikvision and Dahua cameras	🟢 Available	🟡 Private use allowed, public sector banned	🟢 Available	🔴 New FCC authorizations blocked; banned for federal contractors
Robot safety and certification	🟡 GB 11291.1-2011 and GB 11291.2-2013 mandatory, under revision since April 2026; CR certification voluntary; GB/T 39404-2026 covers controller cybersecurity	🟡 Fencing by default; fenceless collaboration must meet CNS 14490 or ISO 10218 with a written assessment	🟡 Per customer and export destination, confirmed in writing before shipment	🟡 Electrical equipment must be “approved” (in practice an NRTL mark or field evaluation); UL 1740 requires robot cells to meet ISO 10218-2 since August 11, 2025
Fabs and controlled-product plants	—	🟡 TSMC’s purchase contracts require SEMI E187; mainland equipment being removed from 2 nm fabs	—	🔴 ITAR technical data includes “photographs”; denial policy for China; mainland-controlled suppliers cannot in practice serve defense electronics plants

Sources for the table: (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).

Mainland China: a complete stack, but neither closed frontier models nor the strongest single machines can come in. Filed domestic cloud models, local open weights (Qwen under Apache-2.0, GLM and DeepSeek under MIT), classic vision models and domestic AOIs, cameras and cobots are all usable. NVIDIA’s own export classifications list the Jetson AGX Thor kit, the T5000/T4000 modules and DGX Spark as 5A992.c, so an ordinary EMS or PCBA plant needs no US licence, while the RTX 5090 and RTX PRO 6000 (4A090.a) and DGX Station (5A992.z,

TPP 58,484) cannot be bought on the mainland (NVIDIA export classifications, eCFR) primary; the legal reading is mine and not legal advice. RS China lists the Thor kit at RMB 59,104 including tax, about 1.4 times the US list price (RS China).

Four law firms agree that a private in-plant deployment for the plant's own use needs no model or algorithm filing, while a hosted multi-tenant service sits in a grey zone (Covington, FPF, JunHe) verified: consistent readings, no official ruling. GB 11291.1-2011 and GB 11291.2-2013 remain the mandatory robot-safety baseline and have been under revision since April 28, 2026; GB/T 39404-2026 on controller cybersecurity took effect on November 1, 2026 (revision plan 20262542, GB/T 39404-2026) single source, primary. Until the new GB 11291 arrives (2028–2029 estimate), design to ISO 10218:2025 and keep GB 11291 compliance files.

Taiwan: private plants can use almost anything; the customer's identity decides. Frontier APIs, open weights and mainland equipment are all legal (data.gov.tw). But since the amended Cyber Security Management Act took effect on **December 1, 2025**, the authorities can act directly on products that are "mainland brands" and restrict them at government bodies, state enterprises and critical-infrastructure providers (Laws & Regulations Database, review rules) verified; the list of critical-infrastructure providers is confidential; and TSMC has written SEMI E187 into its purchase contracts (CNA) and is reported to be removing mainland equipment from its 2 nm fabs (TrendForce) single source. Private EMS, assembly and small PCBA plants are open; fabs, park utilities and government-related customers are closed in practice. A mainland-controlled supplier should design to E187, ship a software bill of materials, offer offline updates and never connect remotely from the mainland.

Southeast Asia: the fewest restrictions; the problems are slow payback and Chinese parents. Everything is usable, but US guidance of May 31, 2026 requires a licence to export advanced computing items to entities headquartered in, or ultimately owned from, China even when the consignee is in a third country (Greenberg Traurig) single source, and Malaysia has its own permits for advanced chips (Fulcrum). Demand is moving here: TrendForce expected Thailand, Vietnam and Malaysia to approach 50% of server-ODM SMT capacity in 2026 (TrendForce) single source, a 2023 forecast.

United States: the full stack, at the price of tariffs, certification and controlled data. Under chapter 99 of the US tariff schedule, a mainland-made robot arm pays about 40% in combined duty, AOI-type instruments and cameras about 37.5% and edge computers 25%; made in Malaysia or Indonesia about 10–12.5%, in Vietnam, Thailand or the Philippines 12.5–15%, in Taiwan 10%, with computers at 0% outside the mainland (USITC HTS chapter 99, Great Lakes Customs Law) primary schedule + estimate. USTR's November 10 notice extended 178 old Section 301 exclusions to January 9, 2027, none of them covering robots, inspection instruments, cameras or computers (Federal Register 2026-22925) verified.

OSHA requires workplace electrical equipment to be "approved" (in practice an NRTL mark, or a field evaluation of about $3,000 to $20,000+ per cell), and since August 11, 2025 UL 1740 brings whole robot cells under ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). Defense electronics is closed to mainland-controlled suppliers in practice: ITAR treats production photographs as technical data and applies a denial policy to China (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). A US-bound cell should use a fixed arm, unrestricted cameras and local or US-cloud models, with final assembly in Southeast Asia where that genuinely changes origin.

At any budget, the ceiling differs by region but nowhere reaches zero people estimate. The mainland can stack domestic 3D SPI, 3D AOI, X-ray, AI review and an in-house AI team to bring review labour down to 8–15%, but without closed frontier models or DGX Station, plant-wide LLM Q&A has to run on local open weights and domestic compute (a full-size model for many users needs at least two eight-card Ascend servers). Taiwan and Southeast Asia can technically use the best of everything, including frontier APIs and DGX Station, limited by customer identity, mainland brands and Chinese-parent licensing. The US has no technical limits, at the price of tariffs, NRTL certification and personnel separation for controlled data. Everywhere, people remain for borderline calls, new products and changeovers, new defect types and customer sign-off.

A product sold across regions therefore needs two model stacks: a mainland stack on filed domestic models or local open weights, and an overseas stack on frontier APIs plus a local model of non-Chinese origin. Anthropic's pol-

icy effective November 12, 2026 judges by ownership and control, so a mainland-controlled supplier cannot build on Claude even through an overseas subsidiary, a cloud platform or a reseller (Anthropic) single source, primary .

ON-PREMISES

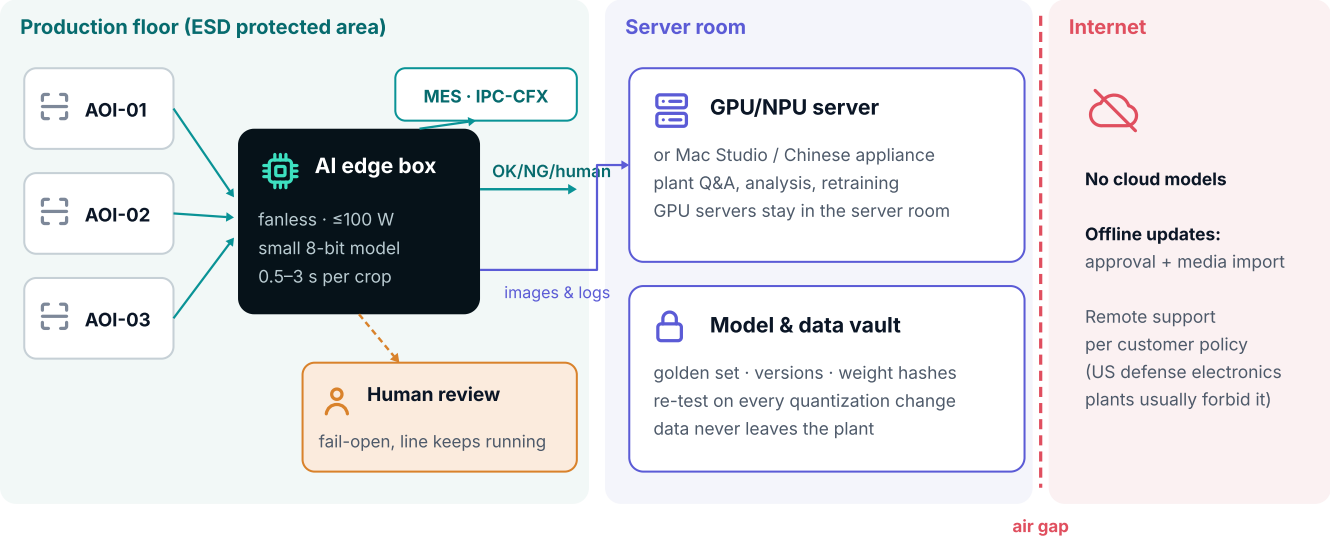
02 Running AI inside the plant: clean rooms, air gaps, hardware and quantization

In an air-gapped plant the constraint is process, not compute

Fully offline vision inspection is routine; the constraints are procedural. China's 2024 MIIT guide on industrial control system security asks for zoning, firewalls or data diodes, audited time-limited remote maintenance, scanned media and tested patches (guide text) single source ; semiconductor customers add SEMI E187 and E188, which TSMC wrote into its purchase contracts in 2023 (SEMI, TXOne) verified . Results go back to MES over IPC-CFX and board identity travels over IPC-HERMES-9852 (Hermes standard). The common trap is licensing that assumes a network: Roboflow's offline mode caches weights on a 30-day lease (Roboflow). An offline product needs offline activation, signed updates on scanned media, a software bill of materials, USB control, local audit logs and a CFX adapter.

FIG 05 **An air-gapped deployment**

A small box at the line decides; the server room answers questions; failures fall back to people and data never leaves the plant.



Source: Author's synthesis (Chapter 2).

Closed frontier models do not, in practice, enter a plant's server room. Gemini is offered air-gapped only through Google Distributed Cloud for governments and large enterprises (Google Cloud); Microsoft's Foundry Local opens large models only to "qualified customers" and names no GPT model (Microsoft); Claude is offered only in cloud government regions, with no on-premises deployment (Claude for government); none of the three is offered in mainland China. Plant equipment has to be designed around open weights; frontier models can at most label or explain images at headquarters, offline, when the customer allows anonymized images out.

Server-room hardware should be treated as office-grade: the Mac Studio is rated for 10–35 °C and DGX Spark for 5–30 °C, with reports of thermal shutdowns under sustained inference (Apple, NVIDIA forum) verified . The line gets fanless industrial boxes, such as a Jetson T4000 industrial PC rated for −25 to 60 °C (AAEON) vendor claim . Because an hour of downtime costs a large plant $36,000 to $2.3 million (Siemens True Cost of Downtime 2024) vendor survey , every AI step on the line must **fail open**: on any fault it falls back to the existing AOI-plus-people process.

Clean and ESD zones: an ordinary SMT hall is an ESD protected area, not a graded clean room

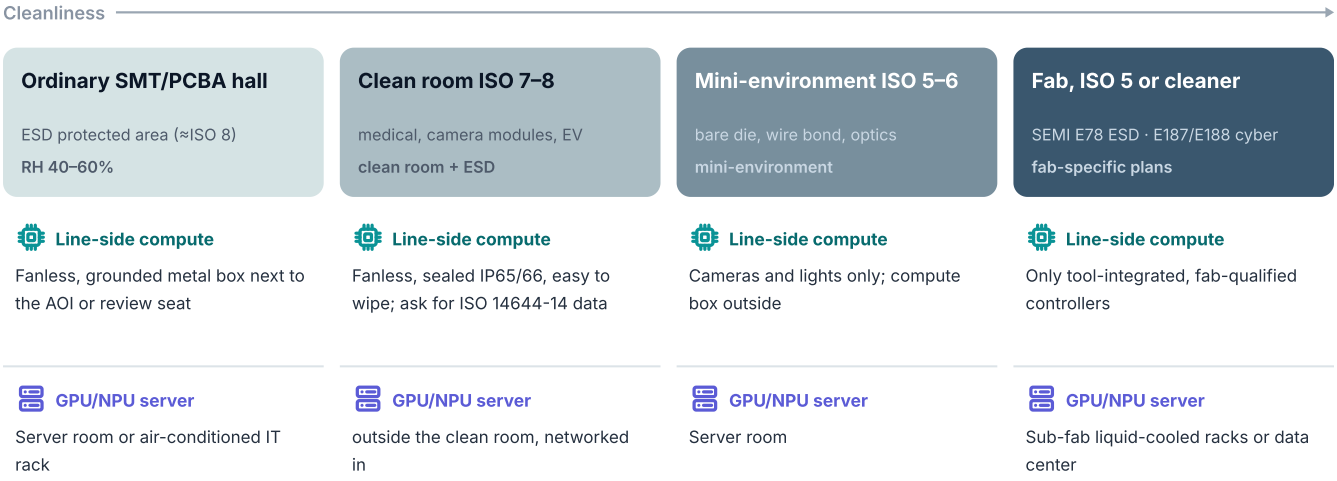
What limits where compute can sit is static and airflow, not cleanliness class. Most SMT assembly areas are “not formally particle-controlled”, and world-class plants “may reach or approach ISO 8” (Circuitnet expert panel) single source. Medical PCBA and camera modules typically sit in ISO 7, bare die and optical alignment in local ISO 5–6 (PCBSync) single source; clean-room builders quote ISO 7–8 for SMT generally (Angstrom Technology) vendor claim.

The universal requirement is ESD control: ANSI/ESD S20.20-2021 requires isolated conductors below 35 V and keeps strong-field insulators at least 30 cm from sensitive devices (ANSI, ESDA, Desco) verified. Fabs put compute in the sub-fab or the data centre, in liquid-cooled racks that meet SEMI S2/S8/F47 (Prodrive) vendor claim.

Whether equipment can enter a clean area is tested for particle emission under ISO 14644-14; robots hold such certificates (Fraunhofer IPA), but I found no industrial PC, edge AI box or GPU server with one. CE, FCC and IEC 60068 show that a box tolerates its environment, not that it may enter an ESD protected area; the customer’s ESD coordinator checks that. GPU servers never go on the line: a DGX Station drew 1,292 W at the GPU in testing (StorageReview).

FIG 06 **Clean and ESD zones decide where compute can sit**

An ordinary SMT hall is an ESD protected area, not a clean room. GPU and NPU servers belong in the server room.



Source: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, SEMI standards and vendor material; compiled by the author.

Zone	Typical environment	Compute at the line	GPU/NPU servers
Ordinary SMT/PCBA hall (consumer, industrial)	ESD protected area under S20.20 or IEC 61340-5-1, usually not a graded clean room (≈ISO 8 equivalent); 40–60% RH	Fanless box in a grounded metal case (Jetson T4000/Thor, Chinese NPU box) next to the AOI or review seat; plastic monitors and housings out of the area or ≥30 cm from boards	Server room or air-conditioned IT rack
Medical, camera modules, EV electronics (ISO 7–8 clean room)	Clean room plus ESD protection	Fanless, IP65/66 sealed, smooth and easy to wipe; ask for ISO 14644-14 data or an IPA certificate	Outside the clean room, networked in
Bare die, wire bond, optical stations (local ISO 5–6)	Mini-environment	Only cameras and lights inside; compute box outside	Server room
Fab clean room (ISO 5 and cleaner)	Fab-specific ESD and electrostatic-attraction plans, equipment to SEMI E78; cybersecurity to SEMI E187/E188	Only tool-integrated, fab-qualified controllers; no add-on PCs with fans	Sub-fab liquid-cooled racks or the fab data centre

The table is my synthesis of the sources above estimate. For a product, that means a fanless line-side box under about 100 W, a metal case bonded to protective earth, 12–24 V DC input, wide-temperature operation at real cabinet airflow, and an ISO 14644-14 particle-emission report ready for customers at ISO 7 and cleaner. Any design where "one GPU server serves four lines" puts the server in the server room and leaves only a small box or a camera at the line.

Speed limits by use: accuracy is harder than speed

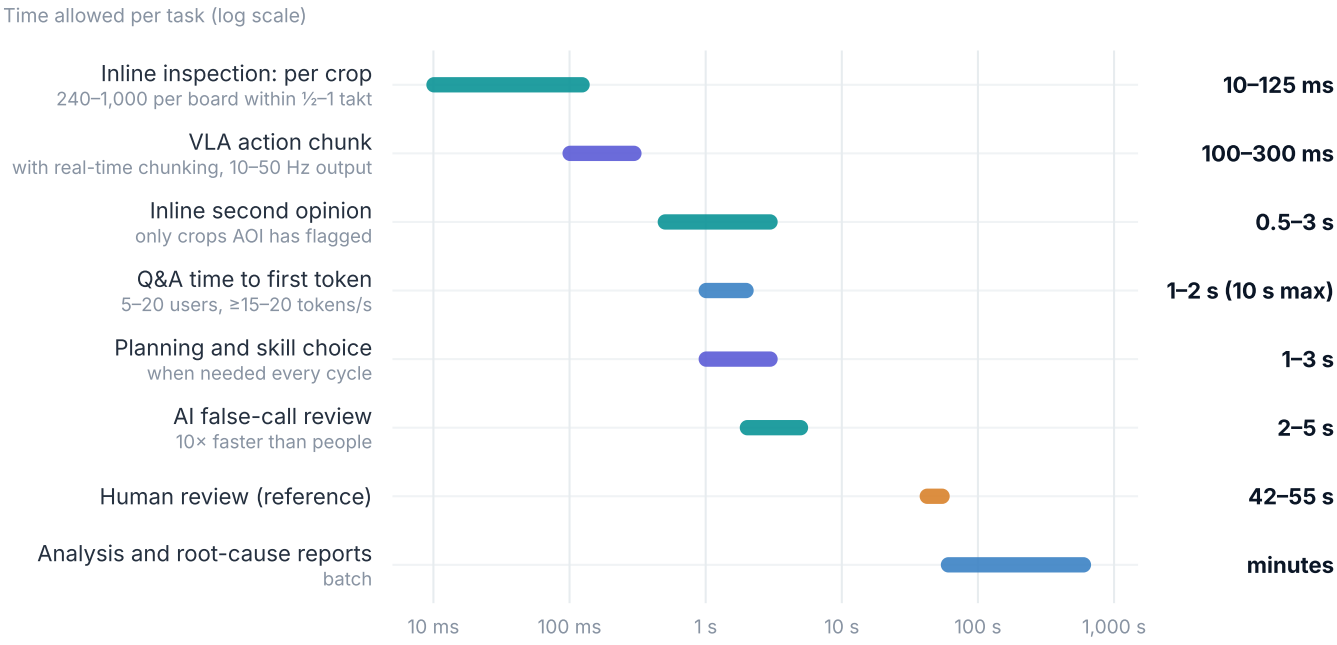
Plants and vendors almost never publish AI latency or availability targets. The thresholds below are derived from takt, human review time, robot control rates and interaction research estimate :

Use	Speed limit	Accuracy limit	Availability
Inline inspection of every part	Within ½–1 takt: 240–1,000 crops in a 10–30 s takt, 8–100 crops per second	Escapes no worse than today's "AOI plus people" (contract example <0.5%)	Fail open, never stop the line; target ≥99.9% (no public SLA)
Inline second opinion on AOI-flagged crops only	Flags (usually ≤5–20 per board) within ½ takt: 0.5–3 s per crop	Same; trained on the line's own images	Falls back to human review
AI review of false calls	2–5 s per call, more than ten times faster than a person's 42–55 s	Escapes no worse than a human reviewer	Degradation adds people, never stops the line
Robot control (VLA)	100–300 ms per action chunk with real-time chunking, 10–50 Hz output	≥99% task success to beat programmed motion	Safety system independent of the model
Robot planning and skill choice	1–3 s when needed every cycle; ≤1 min if only on exceptions or changeovers	Output must be verifiable code or a skill ID	Falls back to the fixed program
Q&A (SOPs, MES queries)	First token ideally ≤1–2 s, at most 10 s; ≥15–20 tokens/s per user; 5–20 concurrent users	Accurate citations; faithful over long context	Office-IT grade (≈99–99.5%)
Analysis and root-cause reports	Minutes are fine (batch)	Highest; faithful over long context	Office-IT grade

Physical Intelligence measured 97 ms of inference for a 3B-class VLA (about 108–139 ms with network and preprocessing), found that real-time chunking keeps throughput unchanged up to 200 ms of extra delay, and completed fine tasks such as striking a match or plugging a network cable at more than 300 ms (Physical Intelligence) vendor research; interaction research puts the limit for keeping a train of thought at one second and for attention at te (NN/g). Accuracy is the hard part: open Qwen3-VL-8B averages only 56.8 zero-shot across five industrial anomaly datasets (IndusAgent) single source. No hardware rescues a vision-language model that has not been tuned on the line's own images.

FIG 07 **Speed limits by use**

Inline inspection is measured in milliseconds, Q&A in seconds, analysis in minutes. People take 42–55 s per call.



Source: Author's synthesis (Chapters 1–2).

Hardware: Thor for the station, a Mac Studio or domestic appliance for plant Q&A, and no DGX Station on the mainland

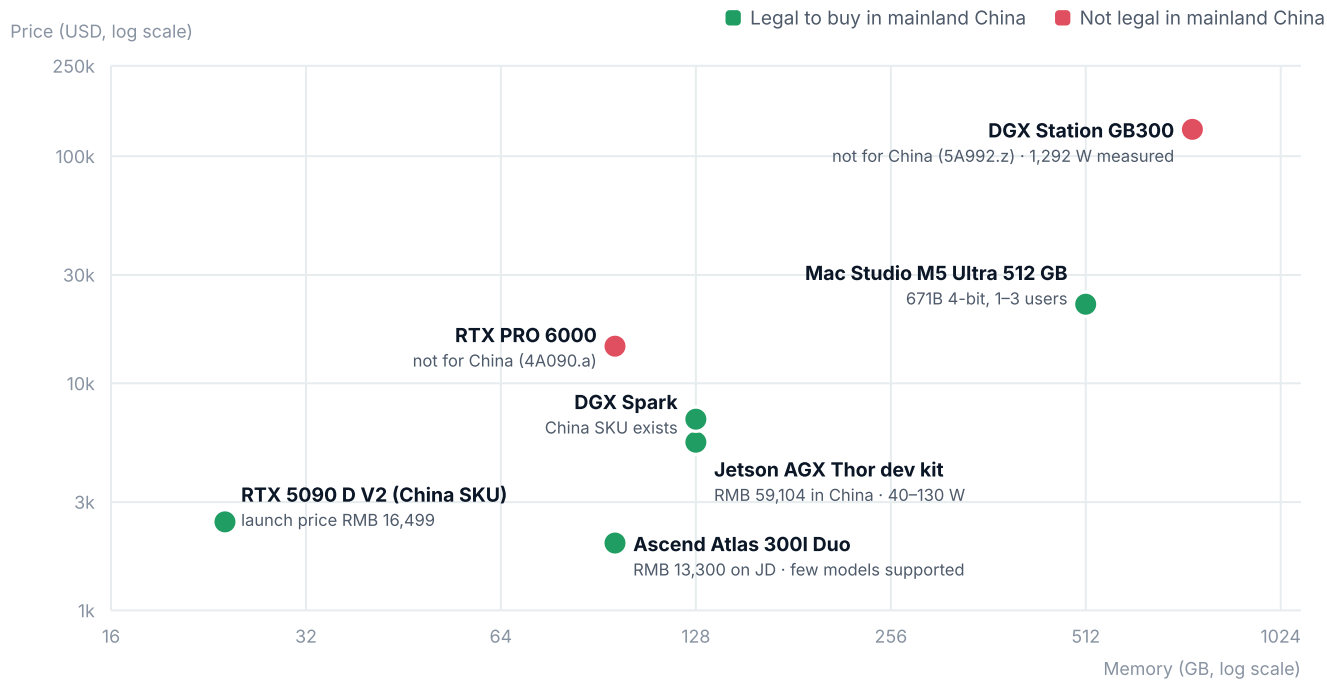
Device	Memory and bandwidth	Price (November 2026)	What it is good for	Mainland China
Jetson AGX Thor dev kit / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 at launch → $5,499 (July 2026); T5000 $4,999, T4000 $2,999; RS China dev kit RMB 59,104 incl. tax	2–8B vision models re-checking crops; real-time 3B-class VLAs	● Yes (5A992.c, no licence for civilian plants)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; the same-chip ASUS GX10 China model RMB 32,999 on JD (Oct 2025)	Multi-user 8–35B MoE	● Yes (5A992.c, China model, authorized channels)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (June 2026) → ≈$16,005 average eBay listing in November	Multi-user 27–35B FP8; 120B-class MoE at 4-bit	● No (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB unified, 1.2 TB/s; 480 W max sustained	From $22,299 in the US; ≈RMB 173–186k estimated in China	DeepSeek 671B at 4-bit for 1–3 users	● Yes
DGX Station GB300	748 GB (252 GB HBM3e at 7.1 TB/s + 496 GB LPDDR5X at 396 GB/s); 1,292 W measured at the GPU	$123–140k street price in November (≈RMB 830–940k before duty); $95–117k in August	High concurrency on 120–235B MoE that fits in HBM; only a few users on 433–743B models	● No (5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	RMB 13,300 from a JD merchant	Low-volume batch review; few supported models	● Yes (domestic)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Not obtained; full DeepSeek needs two at W8A8, four at BF16	Plant-wide multi-user Q&A and analysis	● Yes (domestic)

Device	Memory and bandwidth	Price (November 2026)	What it is good for	Mainland China
RTX 5090 D V2 (China model)	24 GB	RMB 16,499 at launch (Aug 2025), up since with memory prices	8B vision models, small VLAs	Yes

Sources: (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). Mainland prices for the 512 GB Mac Studio are extrapolated from US configuration premiums and are low-confidence estimate.

FIG 08 **On-prem compute: memory, price and what can legally be bought in mainland China**

Export classifications decide the short list before benchmarks do.



Source: NVIDIA export classifications, Apple, RS China, JD.com and review prices (November 2026).

Jetson Thor makes station-level deployment work. NVIDIA's TensorRT Edge-LLM tests show Qwen3-VL-8B (NVFP4) prefilling 2,048 tokens in 104.5 ms and decoding 44.4 tokens/s on Thor (TensorRT Edge-LLM) vendor claim; for an AOI crop and a 20-token answer that is about 0.5 s with the 8B model and 0.2 s with the 2B, or 8.5–21 crops per second batched estimate. Output length, not parameter count, sets the latency, so constrained decoding to a class token or short JSON is part of the design. GR00T N1.7 (3B) runs at 25.1 Hz on Thor (Jetson AI Lab) vendor claim. An Orin NX takes about 1.8 s per crop and an Atlas 300I Duo has a p95 above 23 s, fit only for batch wor (CSDN) single source. On the mainland, budget NVIDIA products at 40–60% above US list price estimate.

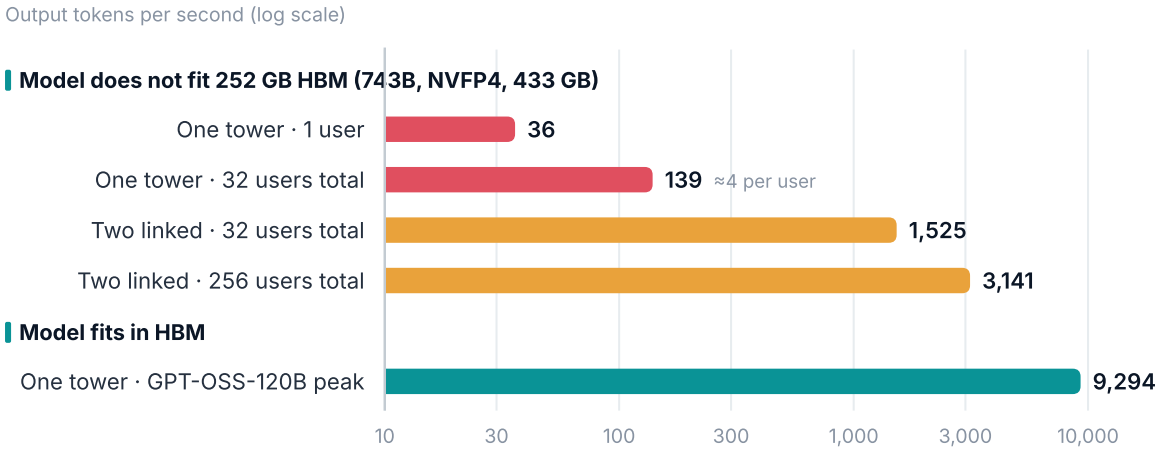
A 512 GB Mac Studio makes a full-size model on one machine possible, for a few users. The M5 Ultra (August 2026) offers up to 512 GB at 1.2 TB/s from about $22,299 (Apple, TechRadar); DeepSeek 671B at 4-bit (≈380–404 GB) decodes 17–20 tokens/s on an M3 Ultra (Pinggy) verified and an estimated 25–29 on the M5 Ultra. Long prompts are the weak point: a 10,000-token retrieval prompt takes roughly 17–70 s estimate, beyond the 10-second interaction limit.

DGX Station is the strongest single machine, but not available on the mainland and five to six times the price of a Mac. The shipping GB300 has **748 GB**, of which only 252 GB is fast HBM (Wccftech, Tom's Hardware, Aug 2026) verified. Independent tests show about 9,300 output tokens/s on GPT-OSS-120B, which fits in HBM, but a 743B model that spills into LPDDR5X gave 32 users only 139 tokens/s in total, about 4 each; two linked units reached 1,525 (StorageReview, Aug 2026, StorageReview, Oct 2026) single source. Street prices rose to $123–140k by

November (CyberPowerPC, monitors.com). On the mainland, the realistic options for the same job are the Mac Studio, KTransformers-style CPU+GPU servers (KTransformers) vendor test or domestic Ascend servers.

FIG 09 **One big workstation does not make a shared assistant**

When the model does not fit in high-bandwidth memory, 32 concurrent users get about 4 tokens per second each.



Source: StorageReview (Aug and Oct 2026); ServeTheHome. Single-source independent tests.

Memory shortages moved the same parts 20–100% in a year (DGX Spark went from $3,999 to $6,950 (TechRadar)), so equipment quotes should be indexed rather than fixed.

Domestic accelerators run full-size models; software maintenance is the bottleneck. Ascend 910B has 64 GB of HBM, but vLLM-Ascend marks DeepSeek V3/R1 as "no longer maintained" and supports only a few models on smaller cards (vLLM-Ascend) single source, primary, and each new model takes one to two months to adapt (Caijing/Anue); ship a frozen, validated model version. Serving full DeepSeek 671B to 20–100 users needs at least two eight-card Atlas 800I A2 servers (China Telecom Cloud) verified, yet public tenders for LLM appliances in China totalled only RMB 27.5 million in January–September 2025 (C114) single source: plants pay for integration and service, not for a box of compute.

Quantization: 8-bit is safe, 4-bit depends on the recipe, 3-bit and below is out

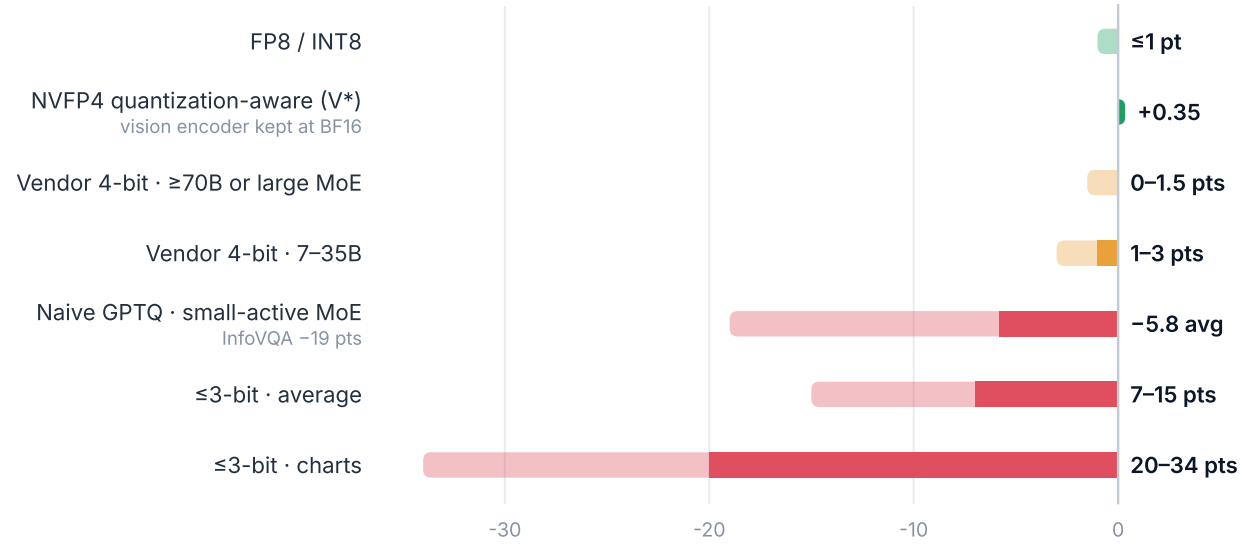
8-bit (FP8 or INT8) is essentially lossless for vision-language models (Red Hat, Hugging Face) verified. At 4-bit, the loss depends on size and recipe: Qwen's 72B AWQ build loses under a point on MMMU while the 7B loses 2.8 (Qwen AWQ) verified, and detail-heavy tasks fall further (naive 4-bit GPTQ costs Qwen3-VL-30B-A3B 19 points on InfoVQA (VEQ)). At 3-bit and below, the average loss is 7–15 points and charts lose 20–34 (VLMQ) verified; beyond 64k tokens, 4-bit loses up to 59%, worse on non-English input (arXiv 2505.20276). Keep the vision encoder at FP16/BF16.

FIG 10 **Quantization: 8-bit is safe, 4-bit depends on the recipe, 3-bit and below is out**

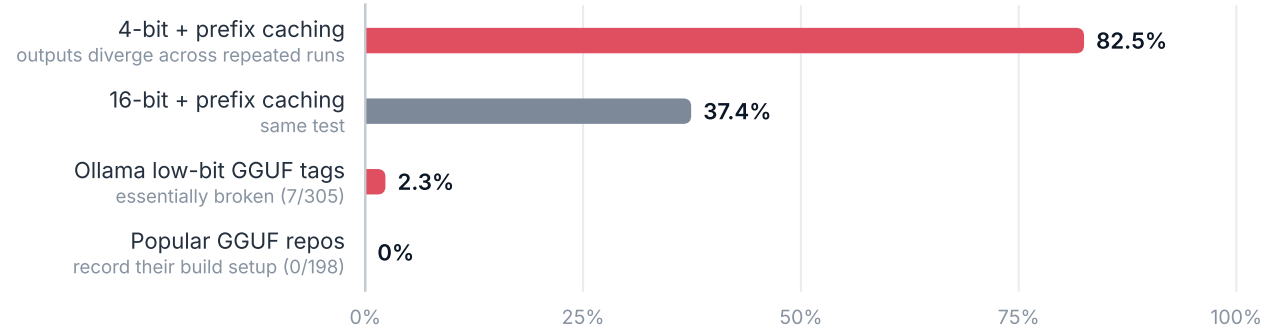
Engineering mistakes in the serving stack cost more accuracy than quantization itself.

Accuracy: 8-bit is safe, 4-bit depends on the recipe, ≤3-bit is out

Score change vs full precision (points)



Engineering failures outweigh quantization error



Source: NVIDIA Nemotron model card, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 and others.

Two warnings matter more. First, **as of November 11, 2026, no study or model card has measured quantization on PCB/SMT defect images**, so revalidate on the golden set. Second, serving-stack mistakes are often bigger than quantization error: the same model scored about 30 points lower under Ollama than under the reference implementation (SiliconBench); 7 of Ollama's 305 official code-model tags were essentially broken although the files were intact (arXiv 2609.05881); none of 198 popular GGUF repositories recorded its build (arXiv 2610.19498); and with prefix-cache reuse, 82.5% of episodes diverged at 4-bit against 37.4% at 16-bit (arXiv 2609.04748) single sources. Vendor quantization-aware 4-bit is close to lossless on detail tasks (NVIDIA Nemotron 3.5) vendor claim, but 4-bit VLAs lose 1.5–4.3 points under perturbation and run slower on Thor (LoopQ) single source. Manage a quantized model like a new gauge: pin weight hashes, record the build, reset prefix caching and revalidate after any change of precision, engine or template.

Precision	≥70B dense or ≥100B MoE	7–35B	≤4B or detail-heavy tasks	Verdict
FP8 / INT8	Safe (≤1 point)	Safe	Safe; the default for inspection	Use
4-bit AWQ/GPTQ (vendor recipe, vision encoder FP16)	0–1.5 points	1–3 points; naive GPTQ on small-active MoE −5.8 on average, −19 on InfoVQA	2–3 points; validate detail tasks first	Conditional

Precision	≥70B dense or ≥100B MoE	7–35B	≤4B or detail-heavy tasks	Verdict
NVFP4 (vendor recipe)	≈0.3–1 point below FP8 (vendor data)	Quantization-aware builds near lossless on V*, OCRBench	No public data	**Conditional (prefer quantization-aware builds)**
GGUF Q4_K_M	Untested	High variance on domain tasks	Not for small text or defects	**Not for detail judgement**
GGUF Q5_K_M and above	Presumed safe	Near lossless in two tests	The floor for detail on llama.cpp	**Use**
≤3-bit (including 1T models squeezed to 1.8–2 bits on one machine)	Not for decisions	No	No	**No**

Which configuration for which job

Configuration	Inline inspection	AI review of false calls	Robot planning and control	Q&A and analysis
Dedicated detectors or anomaly models (YOLO, RF-DETR, PatchCore class), INT8/FP16, any NPU or GPU edge box	● Yes: the main judge, 2–12 ms per crop	● Yes (classifier tuned on the line)	● Perception only	—
2–4B vision models, NVFP4/AWQ/FP8 with FP16 vision encoder, on Thor/T4000 or DGX Spark	● Conditional: re-check AOI-flagged crops at 0.2–0.3 s each	● Conditional, under 1 s after tuning	● VLA backbone (2–3B VLA at ≈40–130 ms per chunk)	● Weak
7–9B vision models (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 or vendor 4-bit, on Thor, Spark or a 24–32 GB card	● Conditional: ≈0.5 s per flagged crop	● Yes: 0.5–1 s per call, best value	● Planning and skill choice (1–3 s), not in the control loop	● Basic SOP Q&A for one or a few users
26–35B MoE (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● No (too heavy)	● Yes, as an explainer or for hard cases	● Planning	● Yes (8 users on a Spark, ≈0.1 s first token)
106–241B MoE at 4-bit on a Mac Studio, RTX PRO 6000 or Spark	● No	● No	● Offline planning	● Yes, for one or a few users
DeepSeek 671B or Kimi K2.x 1T at 4-bit on a 512 GB Mac, KTransformers server or DGX Station	● No	● No	● Offline program generation only	● Conditional: 1–3 users and short prompts; long prompts take 16–170 s
Anything at ≤3-bit; GGUF Q4_K_M for small-model detail work; Ollama for many users	● No	● No	● No	● No

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

So, to the two usual worries: **it does not have to be slow** if the model matches the job (small and mid-size vision models on Thor or Spark meet every inline and review budget; 671B–1T models stay off the line), and **quantization does not have to make it wrong** with FP8/INT8 by default, or vendor-recipe 4-bit above 7B validated on the golden

set. The bigger risks are the roughly 57-point zero-shot industrial accuracy of general models and serving-stack mistakes. Give inspection requests fixed batch sizes and their own queue (temperature 0 is not bit-reproducible under batching (vLLM documentation)), keep an N+1 hot standby, and make every step on the line fail open.

TEACHER AND STUDENT

03 Teacher and student: frontier models can pre-label inspection data; for robot arms the real teachers are still people

November 2026 teacher candidates: generic vision doubled, industrial-defect scores are missing

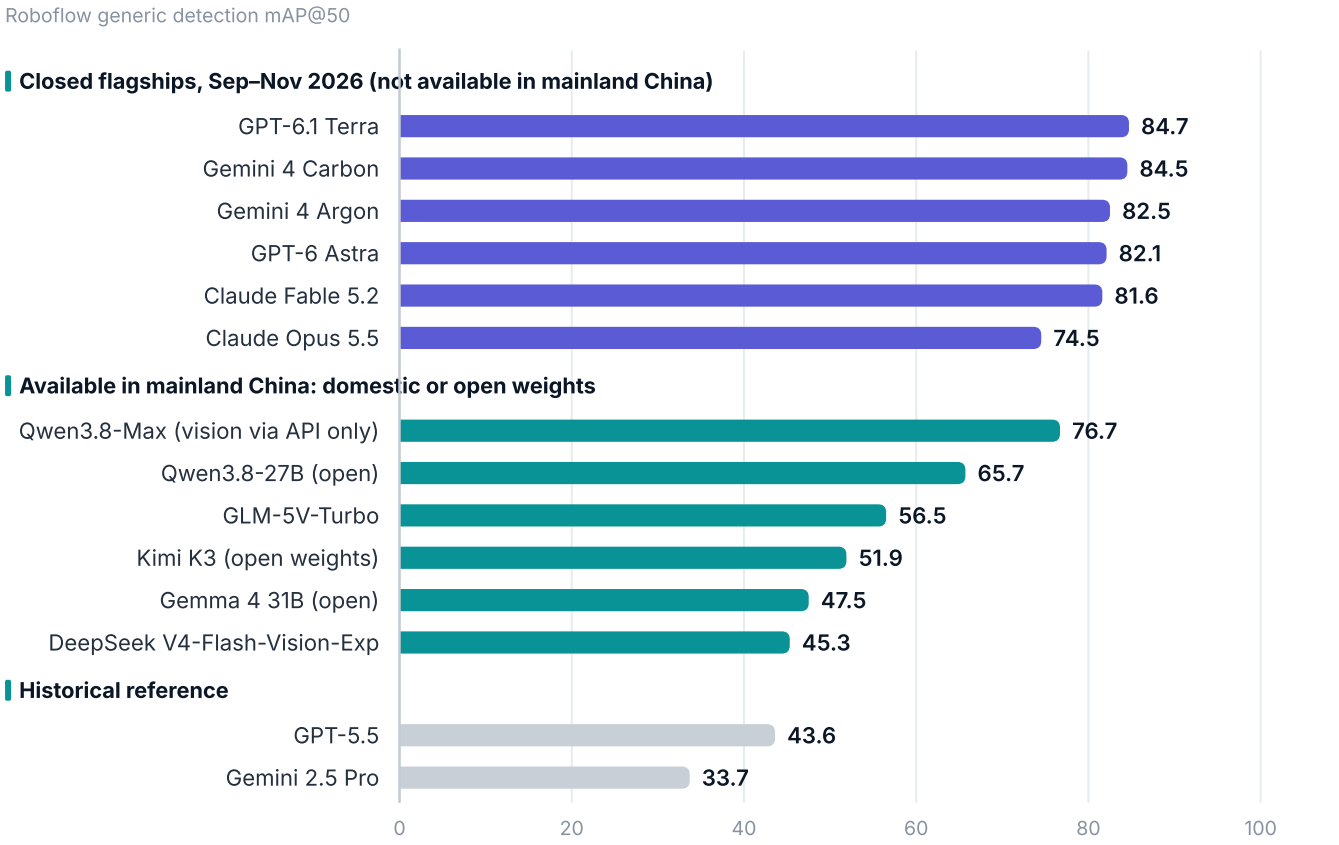
Every closed flagship that could act as a teacher in November 2026 accepts images: Claude Fable 5.2 (generally available November 9, $8/$40 per million tokens) and Opus 5.5 ($4/$20) (Anthropic, Claude models overview) verified; GPT-6 Astra ($10/$50) and GPT-6.1 Terra ($1/$5) (OpenAI pricing); and Gemini 4 Argon and Carbon, opened in stages (Google Carbon). Usable in mainland China are Qwen3.8-Max (vision via API) and open models such as Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp and Gemma 4 (Roboflow blog, Hugging Face).

The key gap has to be stated plainly: none of these models has a published score on MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB or any PCB/SMT dataset. MMAD still has no leaderboard (MMAD GitHub), and the two newest industrial anomaly papers (August–September 2026) still use Gemini 2.5 Pro and Claude Sonnet 4 as their newest general baselines (InspectorGPT, Anomaly-LR). An arXiv full-text search on November 11 found zero hits for “Fable 5.2” or “Gemini 4 Carbon” in this context (arXiv API) single source, absence of evidence. The closest substitutes:

- The newest frontier model with an MMAD score is GPT-5.5: 69.95% on “defect or not” and 76.20% averaged over seven tasks on a 500-image subset, against 95.24% and 86.65% for human experts (AnomalyClaw) single source; GPT-4o scored 68.63% and 74.92% in 2024 (MMAD, ICLR 2025), about 1.3 points of progress in two years.
- On LDU-Bench, real SEM review images from IC lithography and the only electronics benchmark with 2026 models, good-or-bad triage is decent (macro F1 0.932 for GPT-5.4, 0.884 for GLM-5V-Turbo, 0.876 for Claude Opus 4.6, 0.504 for Gemma 4 31B), but classifying 11 defect morphologies reaches only 0.20–0.41 (LDU-Bench) single source.
- On Roboflow’s independent vision evals (updated November 10, 2026), object-detection mAP@50 is 84.5 for Gemini 4 Carbon, 84.7 for GPT-6.1 Terra and 81.6 for Fable 5.2, against 33.7, 43.6 and 38.6 for Gemini 2.5 Pro, GPT-5.5 and Claude Opus 4.8 (Roboflow Vision Evals) single source, independent evaluator, private datasets.

FIG 11 **Teacher candidates: generic scores doubled in a generation**

But not one of the 2026 flagships has a published score on an industrial-defect benchmark.



Industrial defect benchmarks (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): zero published scores for these 2026 flagships

Generic detection scores doubled in a generation, but that is not defect judgement; test on your own golden set

Source: Roboflow Vision Evals; model release notes; MMAD, MVTec AD 2 and other benchmark pages (as of 11 Nov 2026).

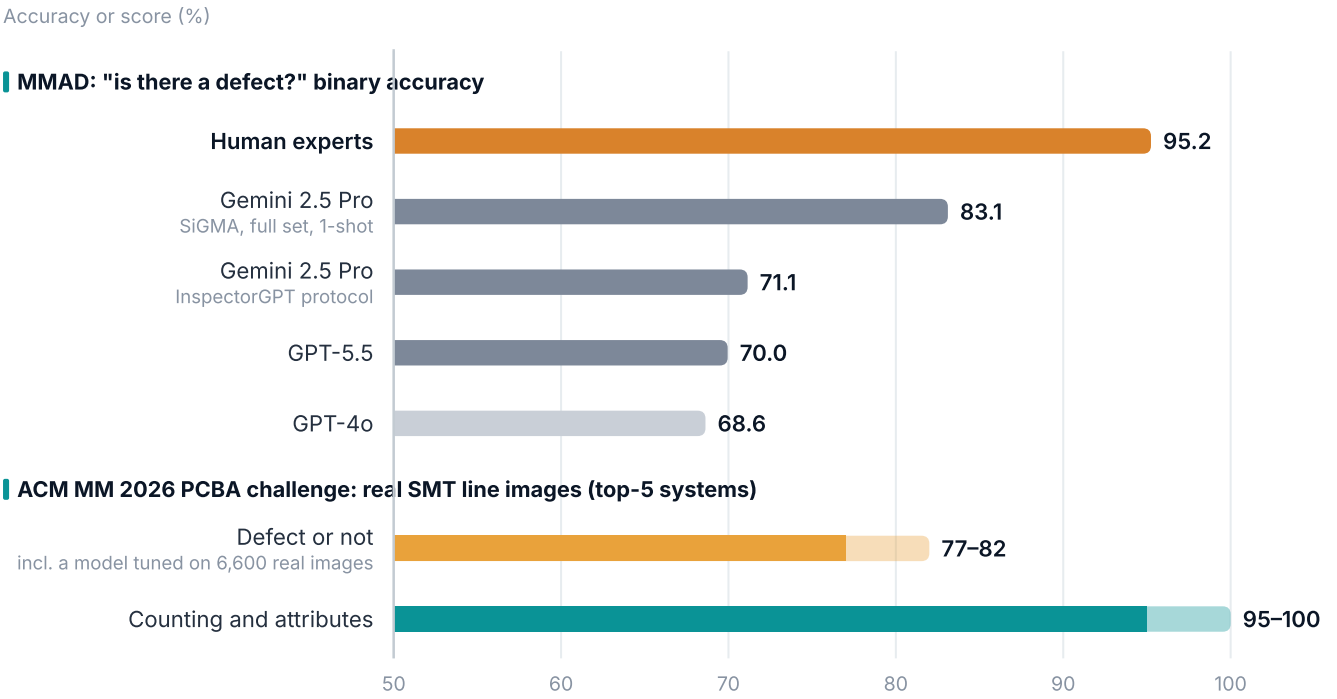
Teacher candidate (Nov 2026)	Price per M tokens (in/out)	Roboflow detection mAP@50 / counting	Industrial-defect benchmarks	Usable in mainland China
Claude Fable 5.2	$8 / $40	81.6 / 86.5	None	No
Claude Opus 5.5	$4 / $20	74.5 / 80.6	None	No
GPT-6 Astra	$10 / $50	82.1 / 80.2	None	No
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	None	No
Gemini 4 Carbon	$2 / $10 (promotional)	84.5 / 90.1	None	No
Qwen3.8-Max (vision via API only)	$2 / $6	76.7 / 81.1	None	Yes
Qwen3.8-27B (open)	Local	65.7 / 64.9	None	Yes
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench triage 0.884	Yes
Gemma 4 31B (open)	Local	47.5 / 51.4	LDU-Bench triage 0.504	Yes (local weights)
Reference: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD binary 69.95% / 68.63%	No

Teacher candidate (Nov 2026)	Price per M tokens (in/out)	Roboflow detection mAP@50 / counting	Industrial-defect benchmarks	Usable in mainland China
Human experts	—	—	MMAD binary 95.24%	—

Newer papers still test the previous generation, but point the same way. On the full MMAD set, human experts score 95.2% on defect-or-not and Gemini 2.5 Pro 83.1% (SiGMA), yet the same model scored only 71.14% in another paper (InspectorGPT): the test protocol moves the score as much as a model generation does. In the CVPR 2026 VAND 4.0 challenge, off-the-shelf vision-language models peaked at 61.88 AP against 90.84 for supervised methods, and the organizers concluded they "cannot currently replace fine-tuned classifiers" (VAND 4.0) single source. In the ACM MM 2026 PCBA challenge, the top five scored 76.92–82.47 on defect-or-not while component type and pin counting reached 95–99.5 (arXiv 2609.21276) single source: recognizing is much easier than judging.

FIG 12 **"Is there a defect?" is still hard for general models**

The best general model trails human experts by more than 12 points on MMAD; tuned systems do better on real line images.



Source: MMAD; SiGMA; InspectorGPT; ACM MM 2026 PCBA challenge.

Three conclusions follow. Seeing and counting doubled in a generation, but judging abnormality moves slowly, depends heavily on the protocol and comes with systematic biases; how much the new models gain can only be measured on a plant's own images. Inside their training distribution, fine-tuned 7B specialists still beat general models (Anomaly-LR, UniPCB). And real line data collapses scores: on assembled electronic panels, image-level AUROC fell from 97–98 on MVTec AD to 31.5–78.6 (arXiv 2606.01992) single source. For mainland companies the strongest available teacher, Qwen3.8-Max, sits about 8 points below the global frontier.

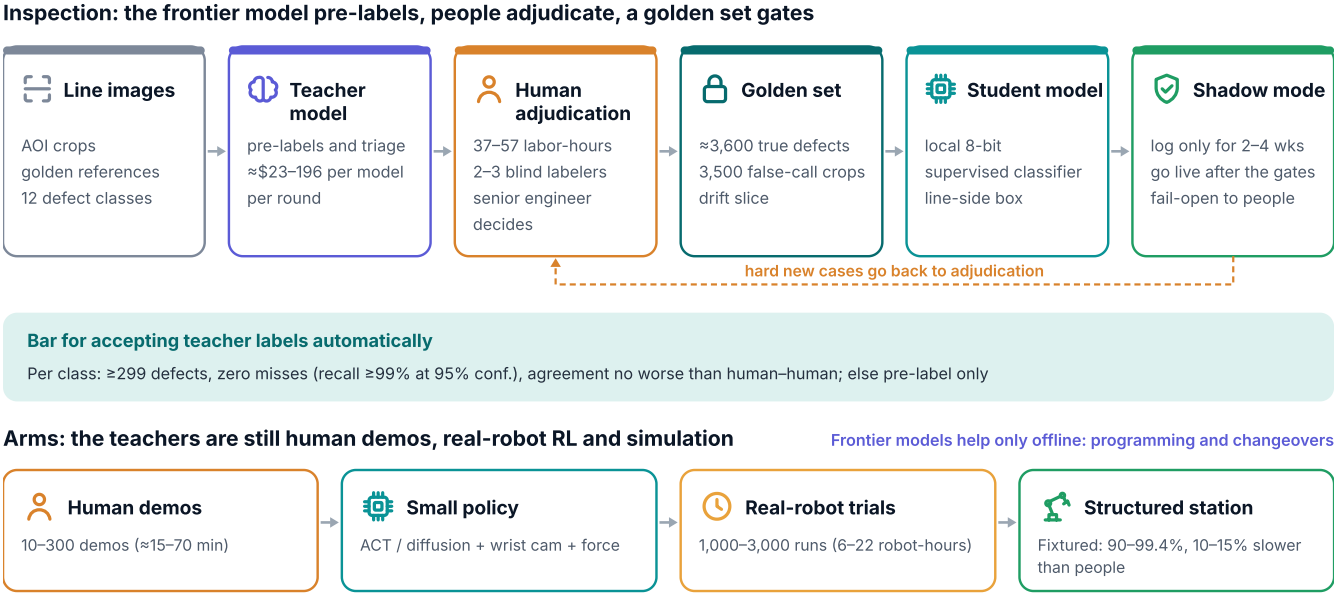
Running dozens or hundreds of rounds: money is not the problem, label quality is

Money is not the problem. At November 2026 list prices, 1,000 one-megapixel images cost about $3 (GPT-6.1 Terra) to $32 (GPT-6 Astra) per round, and a hundred rounds over 100,000 images about $32,000–135,000 before batch discounts estimate (Claude vision docs, OpenAI vision guide, Roboflow Vision Evals). For simple labels a person costs about the same as Terra; the teacher's advantage is speed and scale, not price. Customers' NDAs usually keep product images inside the plant, so the golden set is labelled on site.

The real obstacle is quality, and more rounds do not fix systematic errors. Roboflow runs every task three times and scores move only ±0.1–1.4 points (Roboflow), so a model mostly repeats its mistakes and majority voting cannot save it. Rounds help only when each adds information: one or two good reference images add about 3–4 points, more add nothing (AD-Copilot); having a model argue against a good reference over several turns adds 6.23 points of cross-domain AUROC for GPT-5.5 and 2.95 on MMAD, peaking after about three calls (AnomalyClaw) single sources. "A student trained on teacher labels approaches human labels" has only been shown on natural images: a ResNet-101 trained on Qwen3-VL labels of COCO and Objects365 reached 95.7–99.4% of the human-label result, and the authors say domain data is untested (arXiv 2602.20972) single source.

FIG 13 **Two teacher–student pipelines**

For inspection a frontier model pre-labels and people adjudicate; for robot arms the real teachers are still human demonstrations.



Source: Author's synthesis (Chapter 3); costs and sample sizes are estimates.

A workable protocol estimate:

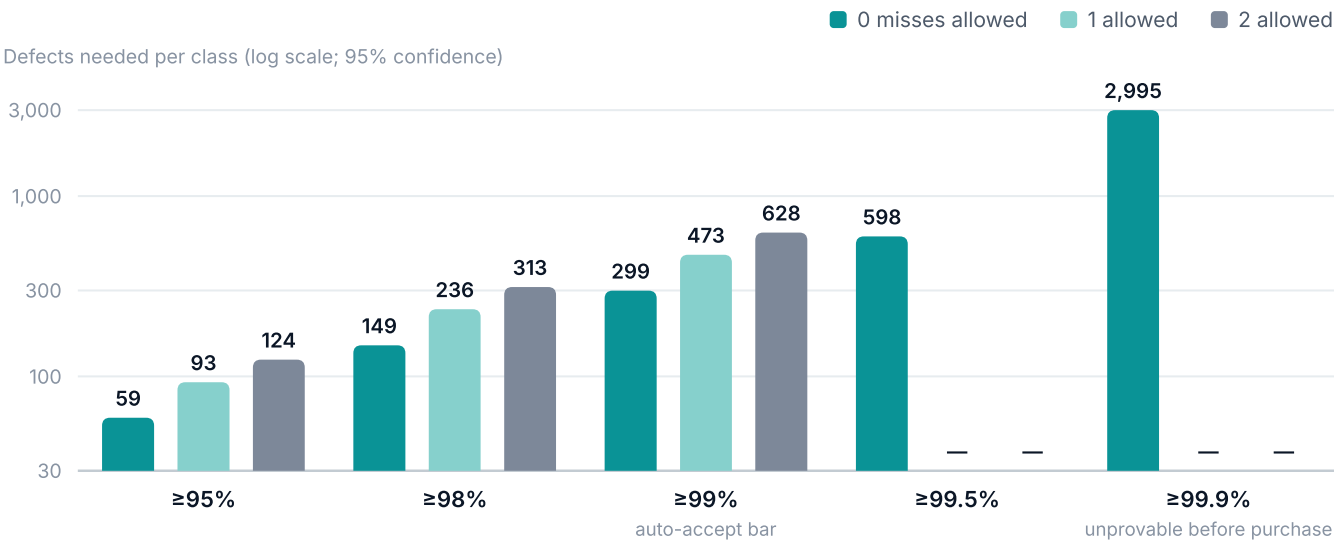
1. **Collect the golden set from the plant's own AOI archive and retained NG boards**, stratified into 12 true defect classes (missing, offset, tombstone, polarity, wrong part, bridge/short, insufficient solder, excess solder, solder ball, lifted lead/open, damaged part, foreign material) and four false-call causes (shadow, flux, silkscreen, tolerance), plus a "drift slice" of at least 300 defects from a second line or different lighting, board colour or part numbers.
2. **Label blind with two or three IPC-A-610-trained inspectors**, with a senior engineer deciding and X-ray, electrical test or cross-sections settling ambiguous samples. Accept a class definition only if inspector agreement reaches κ ≥0.75 (≥0.90 for critical classes); human agreement on 11 electronics defect morphologies in LDU-Bench was only 0.727, so some classes will need merging or redefining. AIAG-style attribute agreement (effectiveness ≥90%, miss ≤2%, false alarm ≤5%, κ >0.75; >0.90 for release) is a sensible minimum (Fabrico, iFactory).
3. **Size it with exact binomial (Clopper–Pearson) one-sided 95% bounds** (binomial confidence intervals, rule of three:

Recall to prove	0 misses allowed	1 allowed	2 allowed
≥95%	59 defects	93	124
≥98%	149	236	313

Recall to prove	0 misses allowed	1 allowed	2 allowed
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **How big the golden set must be**

Proving 99% recall with zero misses takes 299 defects per class; 99.9% cannot be proven before purchase.



Source: Exact binomial confidence intervals (95%).

Sixty defects per class are enough for a pass/fail screen; 300 are needed to decide on automatic acceptance. The full set, about 3,600 defects plus 3,500 false-call crops, takes about 37–57 labour-hours to label and adjudicate, and running a teacher candidate over it costs $23–196; the real cost is two to six weeks of collection, set by the rarest class estimate. Send single-component AOI crops, not whole boards, with the target at least 42–48 pixels across (arXiv 2610.09313).

1. **Accept teacher labels automatically only for a class with zero misses on ≥299 golden defects**, model–truth agreement no worse than human–human agreement, and at least 50% of false calls cleared. Critical classes such as shorts, missing parts, polarity and wrong parts need 99.5–99.9% (598–2,995 defects); otherwise anything the AI cannot clear with high confidence goes to a person. An escape rate below 0.1% of defects cannot be proven before purchase (2,995 defects with zero misses; 0.01% would take 29,956), so the line keeps human sampling of AI-cleared calls, downstream ICT/FCT as a backstop and SPC on escapes. Verified labels are then distilled into a conventional detector or classifier and revalidated on every new product.

Terms and data rules also decide the teacher. Anthropic forbids using its services to train competing models (Anthropic commercial terms); OpenAI allows outputs to train in-house classifiers that are not sold to third parties (OpenAI business terms); DeepSeek explicitly allows distillation (DeepSeek terms) primary terms. None defines a "competing model", so commercial use needs legal review, and customers' data rules usually keep images on site. The safest default is a locally run open-weight teacher, with cloud frontier models only when the customer allows it in writing.

Inspection students: small and fast; the bottlenecks are optics, data and validation

The student is not the bottleneck: YOLO26n takes 3.80 ms per image in INT8 on a $249–399 Orin Nano Super (Ultralytics) vendor test, NVIDIA reports 98.51% on three PCB defect classes after adapting a foundation model on

about 600 images (NVIDIA) vendor claim, and synthetic defects add only 0.9–2.5 mAP points (arXiv 2605.04635). The constraints are optics, line data, validation and licences: Ultralytics YOLO is AGPL-3.0 or enterprise-licensed (Ultralytics), so closed products need a licence or an Apache-2.0 detector such as RF-DETR.

Robot students: structured stations are near production grade; flexible and dynamic work is not

In November 2026 there are three kinds of robot teacher, and only one can become a local student: open, fine-tunable VLAs (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Closed robot models (π0.6/π0.7 through a beta API, Gemini Robotics 2 for trusted testers) cannot be deployed on site (The AI Insider, DeepMind). Frontier general models are slow and poor at fine manipulation: in Artificial Analysis's test they scored 46–69 with 28–45% grasp success and about 30 minutes per task, "far too slow and expensive for any production use" (AA-Robotics); GPT-6 Astra managed a precision slot only 2 times in 20 (RoboCurve); and driving joints directly gives 0–5.5% on LIBERO at 0.2–0.4 Hz against about 83 Hz needed (Anthropic) verified across three sources. Fable 5.2's 64.9% on the simulated RoboDojo benchmark is an internal run (Anthropic) vendor claim.

Third-party studies since October agree: about 33–44% on LIBERO-10 for the three leading agents (arXiv 2610.30051); agents falsely claiming success on 24–39% of subtasks (arXiv 2611.08209); and 104 of 108 quasi-static real-robot tasks for Opus 5.5, but at 280 seconds each and with no insertion or assembly (AMP1) single sources. Frontier models belong offline: writing programs, orchestrating tasks and drafting success criteria.

Student quality has a clear edge: good inside the training distribution, collapsing outside it. π0.7 "often exceeds 90%" on seen tasks and reaches 60–80% on unseen ones (π0.7 paper) vendor research; ContactDP reaches 97.8% on trained connectors but 28.6–66.7% on unseen ones (ContactDP); slight camera or pose changes drop 95% below 30% in LIBERO-Plus (LIBERO-Plus, CVPR 2026); learned grasping from moving conveyors peaks at 58.89% (ForeTime-VLA) single sources, a job classic vision tracking has done in production for years.

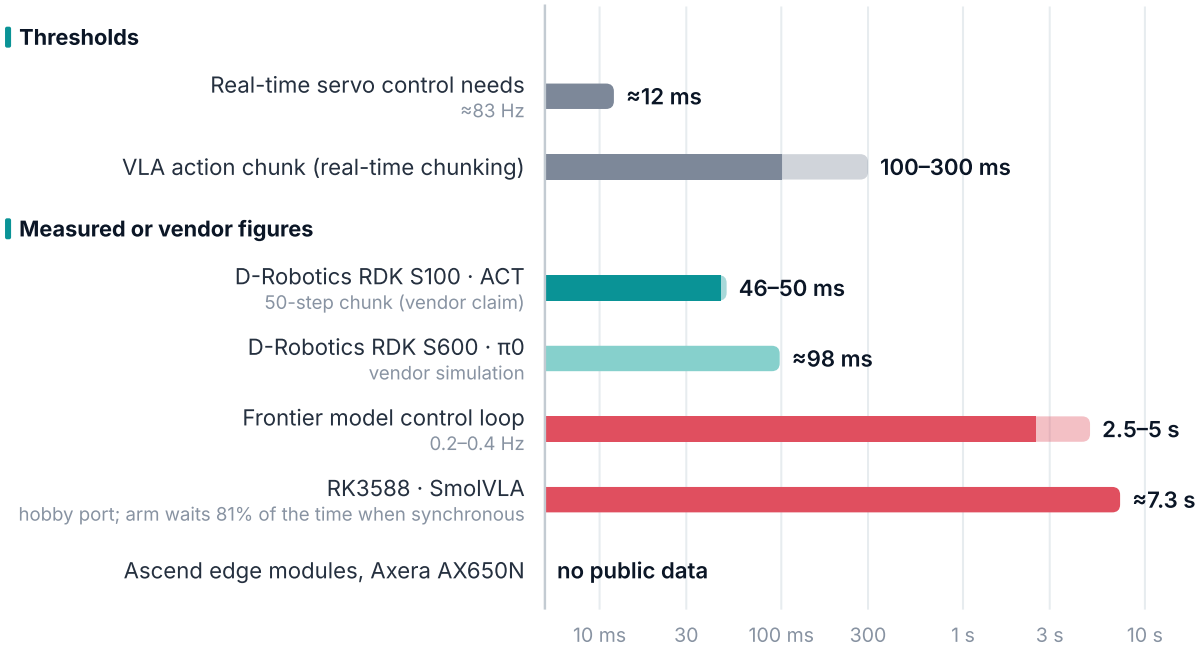
The near-production evidence all comes from structured stations, human demonstrations and real-robot correction. Neuromeka kept a state machine and pre-taught motion and learned only grasping, cable insertion and soldering, each from under 20 minutes of data: 322 of 324 in production, 12.8% slower than a person (Neuromeka) single source, vendor. Fraunhofer reached 92.8% across four connectors from 68 minutes of demonstrations (Fraunhofer); Berkeley's HIL-SERL succeeded 100 times out of 100 on RAM, SSD and USB insertion after 1–2.5 hours of real-robot RL (HIL-SERL); AgiBot ran 18–20 seconds per step within a 20-second takt on a tablet test line (IThome) vendor claim.

Distillation buys speed, not generalization: π0.5 runs at 0.70 Hz on Jetson Orin raw and 10.30 Hz after asynchronous scheduling (EagleVLA); layer distillation takes it from about 3 Hz to about 10 Hz (Shallow-π) verified. A student inherits only what the teacher covered; each new product still needs task data.

FIG 15 **Edge policy latency**

A small ACT policy runs in about 50 ms on an edge board; a VLA ported to a hobby chip takes 7.3 s.

Inference latency per action chunk or control cycle (log scale)

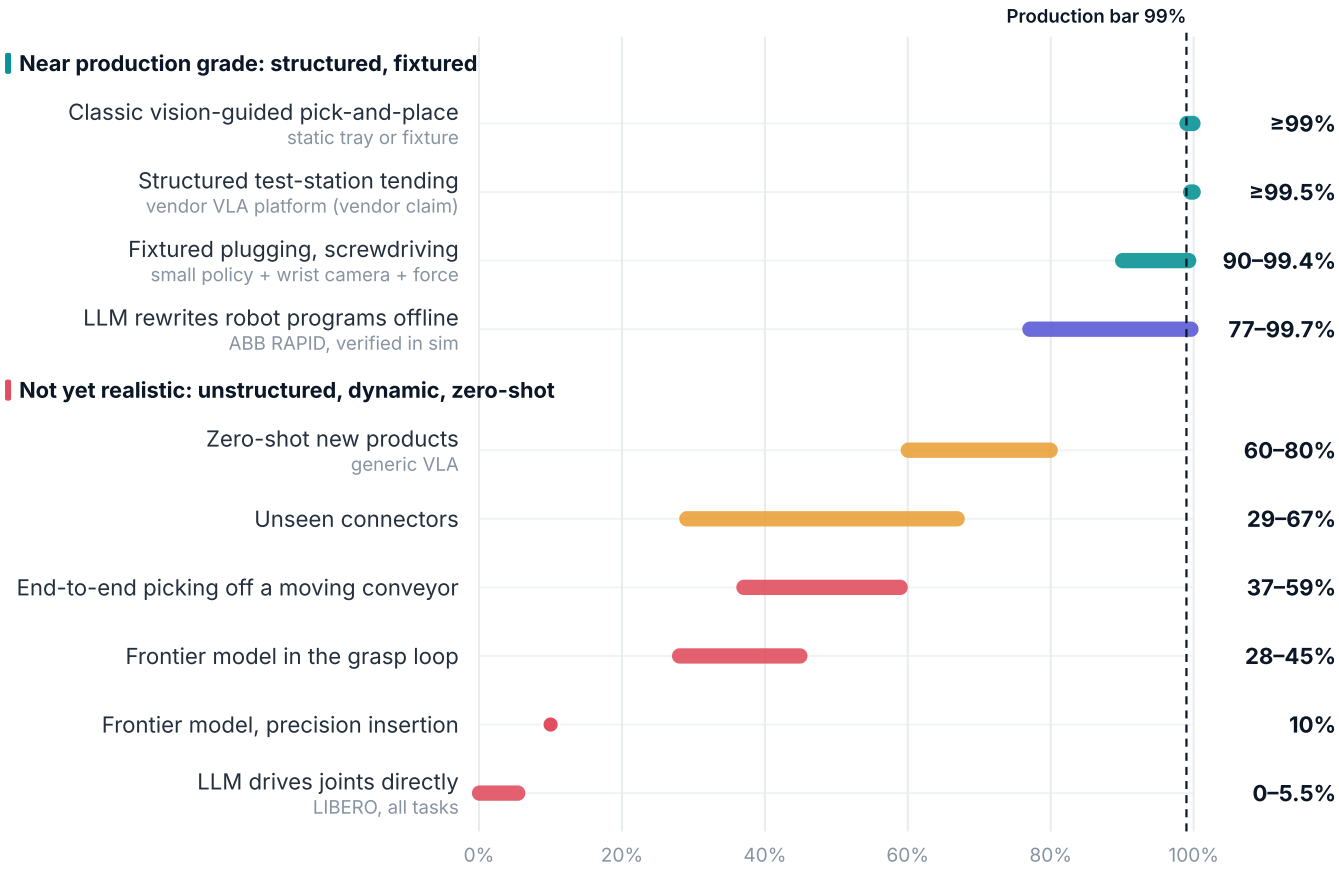


Source: D-Robotics documentation and forum (vendor figures); GitHub community ports; Anthropic research.

On Chinese edge chips, the only measured learned policy is an ACT policy on D-Robotics' RDK S100 at 46–50 ms per 50-step chunk, with no success rate published (D-Robotics RDK S docs) vendor claim ; a hobby port of SmolVLA on an RK3588 takes about 7.3 s per chunk (GitHub). Quantization does not buy speed for free: no 4-bit π0.5 build was faster than 16-bit at batch size 1 (LoopQ). Demonstrations are expensive (Reuters reported about 300 attempts per usable motion for a novice teleoperator (Notebookcheck)), and validation costs more: proving a failure rate ≤0.1% at 95% confidence takes about 3,000 consecutive successes, 17–22 robot-hours estimate . No learned manipulation policy has safety certification.

FIG 16 **Robot students: near production grade only when the station is structured**

Fixtured stations reach 90–99%+; unstructured, dynamic and zero-shot tasks are not realistic yet.



Source: Papers and vendor material cited in Chapter 3; some figures are vendor claims.

Electronics task	Workable student	Expected success	Speed vs a person	Cost per changeover	2026–2028 verdict
High-mix pick-and-place, kitting, machine tending (static tray or fixture)	Learned perception plus programmed or planned motion	≥99% with classic vision guidance	Same as conventional stations	Hours	Feasible now
Fixtured connector, FFC and cable insertion, screwdriving, soldering	State machine + small BC/diffusion/ACT policy + wrist camera + force/torque (ideally touch) + human correction	90–99.4%	10–15% slower	10–300 demos (≈15–70 min) + 1,000–3,000 validation runs (6–22 robot-hours)	Feasible, R&D-heavy months for the first station
Structured test-station tending and OK/NG sorting	Vendor VLA platform + digital twin + real-robot RL	≥99.5% vendor claim, conflicting	18–20 s per step, within a 20 s takt	≤4 h retraining + 15 min calibration	Feasible on a vendor platform
Changeover programming	LLM writes or edits robot programs offline, verified in simulation	77–99.7% on ABB RAPID rewrites	No effect on takt	Minutes to hours per program	Feasible as an engineering tool
End-to-end learned picking from moving conveyors	VLA student	37–59%; ≤37% on fast belts	—	≥458 demos per setup	Not before 2028; use classic tracking plus learned perception

Electronics task	Workable student	Expected success	Speed vs a person	Cost per changeover	2026–2028 verdict
New products with no new data (zero-shot)	Generic VLA	60–80%; unseen connectors 29–67%	—	—	Not realistic
Frontier LLM in the loop driving the arm	GPT-6, Claude, Gemini agents	Grasp 28–45%; precision insertion 10%	≈30 min or 280 s per task vs ≈20 s	—	Not realistic
Free-routed wire harnesses and multi-step flexible assembly	—	No 2026 production evidence	—	—	Not before 2028

Sources: the studies above and (University of Passau on RAPID rewriting, GraspVLA, Komax) **mostly single source and vendor data** . In 2026–2028, a large model's value to robots is as a pre-trained prior and an offline programmer, not as something that "teaches" a student. Every near-production case I found (Neurometka, AgiBot at Longcheer, Fraunhofer, ContactDP) relies on human demonstrations, force or touch sensing, simulation or digital twins and supervised RL. None relies on labels from a frontier model.

COSTS AND PAYBACK

04 Costs and payback: AI review pays back in months, machine tending in years, pick-out arms never

Labour cost: the US is about 5.5 times Shenzhen, Taiwan about 1.5 times, most of Southeast Asia under half

The return on replacing labour starts with what the labour costs. Shenzhen's official wage guidance puts the 2024 median for a manufacturing quality inspector at **RMB 74,074** a year; wages are 86.92% of total labour cost in electronics manufacturing, so the fully loaded cost is about RMB 85,000. JETRO's 2025 survey puts the annual cost of a Chinese factory worker at $12,667, about RMB 85,100, within 0.2% (Shenzhen wage guidance, JETRO 2025) verified.

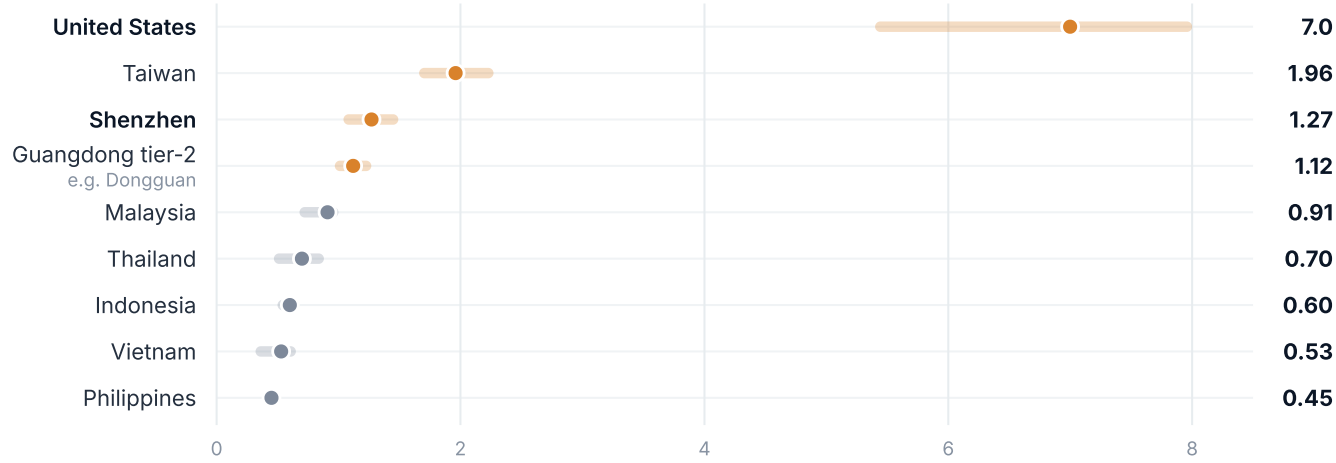
Region	Fully loaded annual cost of one inspector or handler (low / base / high)	Basis
Mainland China, Shenzhen	RMB 70k / **85k** / 100k ($10.4k / **12.7k** / 14.9k)	Official median × 1.15; matches JETRO verified
Guangdong tier-2 wage areas (Dongguan etc.)	RMB 65k / **75k** / 85k	Guangdong 2023 median × 1.15 plus two years of raises estimate
Yangtze River Delta (Suzhou, Kunshan, Wuxi)	As Shenzhen	No basis for cheaper inspection labour; Jiangsu's top minimum wage is about 5.6% above Shenzhen's estimate
Taiwan	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	Ministry of Labor occupational wage survey plus ≈1.18 statutory on-costs; JETRO as the high end estimate
Malaysia	$6.8k / **9.1k** / 10.0k	JETRO and minimum wages estimate
Thailand	$4.7k / **7.0k** / 8.8k	Same
Indonesia	$5.0k / **6.0k** / 6.3k	Same
Vietnam	$3.2k / **5.3k** / 6.5k	Same
Philippines	$4.0k / **4.5k** / 4.9k	Same
United States	$54k / **70k** / 80k	BLS May 2025 inspector median $48,570 ÷ wages at 65.5% of compensation estimate

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

FIG 17 **What one inspection or handling worker costs a year**

Fully loaded, the United States is about 5.5× Shenzhen and Taiwan about 1.5×.

Annual loaded cost per QC or handling worker ($10k; dot = base, bar = low–high)



Source: Shenzhen official wage guidance, Taiwan Ministry of Labor, JETRO, BLS; author's estimates.

Three corrections set the basis. First, Shenzhen raised its pension contribution floor from RMB 4,775 to 5,521 a month in July 2026 (Shenzhen Bendibao) single source, so a Shenzhen inspector in 2026 more likely costs RMB 80–90k. Second, a post is not a person: two 12-hour shifts need about 2.0 people (six days a week) to 2.4 (all year) in the Pearl River Delta, about 2.7–3.1 in Southeast Asia under overtime caps, and 4.2–4.5 for 24×7 in the US and Taiwan estimate. AI review usually lets one reviewer cover two to four lines, so the realistic saving is **0.5–1.5 full-time people per line, not one person per line per shift**. Third, buyers in Taiwan, Vietnam and Malaysia are not only chasing cost: in JETRO's survey, 61.1% of firms in Taiwan, 48.2% in Vietnam and 44.9% in Malaysia said hiring had become harder in the past two years, against 18.6% in mainland China (JETRO 2025) single source.

Equipment prices: four bases that should not be averaged

AOI prices disagree mainly because they measure different things. Online listings (November 2026, new, used and rental mixed) run about RMB 17–80k for offline or desktop 2D, RMB 50–170k for inline 2D and RMB 80–475k for in-line 3D, up to RMB 700k for Korean 3D (1688 inline AOI, 1688 offline AOI, 1688 3D AOI) single source, list prices. A November 2025 tender capped one 3D AOI plus one 3D SPI at RMB 800k (tender notice). Sinic-Tek's prospectus implies RMB 428–512k per 3D AOI in 2019–2022, about RMB 340–470k by its 2025 report (Sinic-Tek prospectus, Sinic-Tek 2025 annual report) primary + estimate. An overseas vendor estimates $3,000–20,000 for desktop 2D, $30–70k inline 2D, $50–90k entry 3D and $100–200k+ high-end 3D, with the same inline machine at $30–50k in China (Averroes) vendor estimate from a seller of competing AI software.

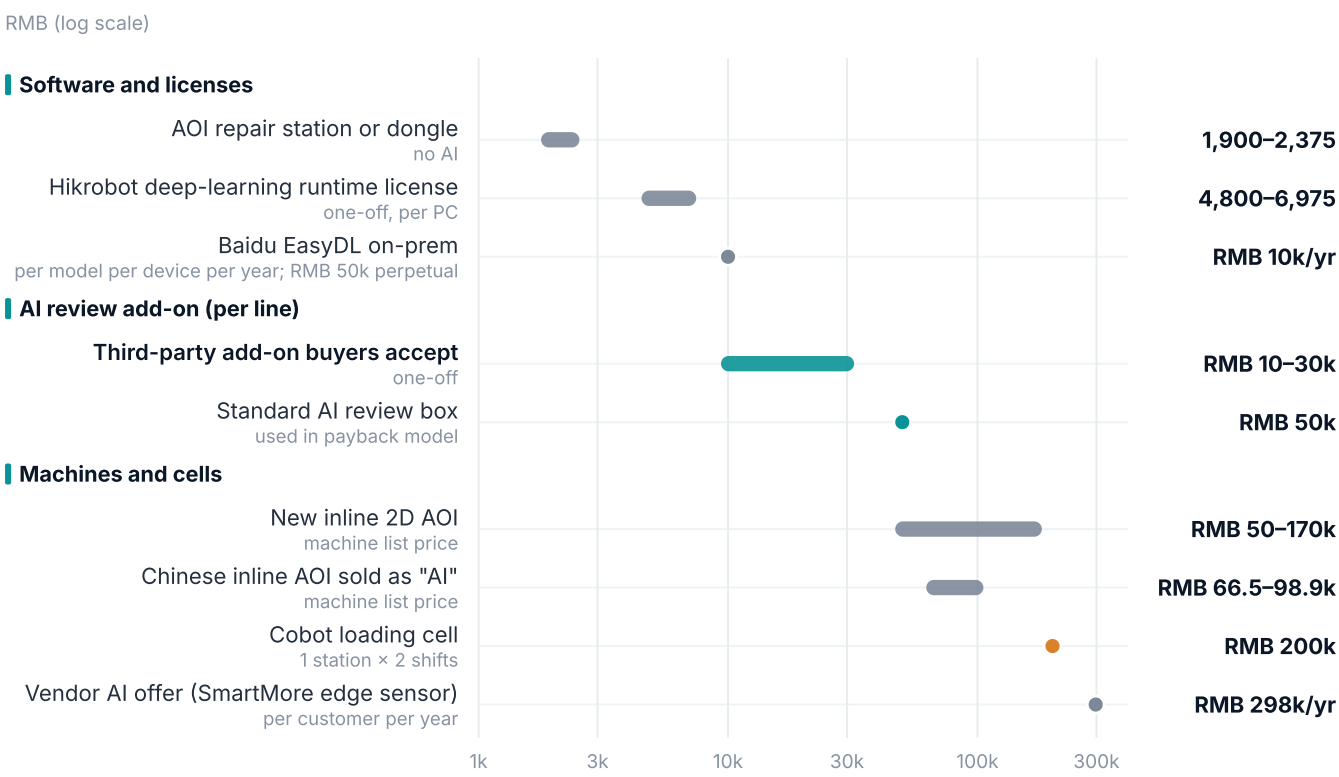
AI review itself has no public RMB price; vendors quote case by case after a free on-site evaluation (Sohu on Hengtianyi), and searches of e-commerce listings, tenders, vendor articles and Taiwanese media turned up no public quote (1688, 51SMT). The anchors: about €27k per line implied by the Siemens case estimate; $21,924 list for a US distributor's remote-review network module (Manncorp) vendor list price; about RMB 298k per customer per year for SmartMore's edge AI sensors in 2025 (36Kr) single source. A domestic inline 2D AOI sold as "AI" lists at RMB 66.5k (1688 inline AOI), and an NG buffer for manual re-inspection about RMB 50k (1688 AOI review). So the ceiling for a third-party review add-on is about RMB 10–30k per line, or a similar annual fee; above that, the buyer would rather replace the machine with a domestic AOI that has AI built in estimate.

Anchor	Price (RMB)	What it buys	Source
AOI repair-station or software dongle (older TRI, Zhenhuaxing)	≈1,900–2,375	Licence key, no AI	1688 list price

Anchor	Price (RMB)	What it buys	Source
General deep-learning vision runtime (Hikrobot VisionMaster dongle)	≈4,800–6,975 one-off per PC	The platform; the customer builds and maintains models	1688 list price
Baidu EasyDL on-premises deployment	10,000 per model per device per year, or 50,000 perpetual	Packaged inference service	Baidu AI docs 2022 price page
Domestic inline AOI sold as "AI"	66,500–98,900 per machine	Machine replacement	1688 list price
SmartMore edge AI sensors	≈298,000 per customer per year	Vendor solution across several machines	36Kr single source

FIG 18 **The price ladder**

From a few thousand RMB of software licence to RMB 200–300k for a robot cell or a vendor AI offer.



Source: 1688 listings, Baidu AI documentation (2022 price list), 36Kr, an industry interview (unverified).

A Baidu-style annual fee, with one inference server for one to four lines and one to four models each, implies a subscription of RMB 2.5–40k per line per year, realistically about RMB 5–10k; at 20% of a perpetual licence, subscription and purchase break even in year five, a structure worth copying estimate. Buyers are used to near-zero software fees: Sinic-Tek's entire "other" revenue (spares, service, rental, accessories) was 7.65% of 2025 sales, about RMB 2,300–3,100 per installed machine per year (Sinic-Tek 2025 annual report) primary + estimate. An AI subscription of RMB 5–10k per line per year is two to five times what buyers now spend on AOI software, so it has to be justified by the review hours it frees (about RMB 75–90k per full-time person a year), not by comparison with existing software fees.

For robot arms, China's average cobot price fell to about RMB 58,600 in 2025 (GGII) estimate, and Dobot's six-axis average fell 19.1% that year as cheaper models gained share (Huxiu) verified, even though list prices of the same models rose 5–15% in 2026 (Sina Finance). In the US, a FANUC CRX-10iA listed at $51,434 (Vention) and compilations put a UR7e at $30–48k (cobotcost) estimate. The arm is only 40–50% of a US cell (Standard Bots); in China, cameras, grippers, safety and engineering easily run two to four times a RMB 40k arm, so budget absolute engineering cost rather than a multiple estimate.

Payback: review in months, tending in years, pick-out never

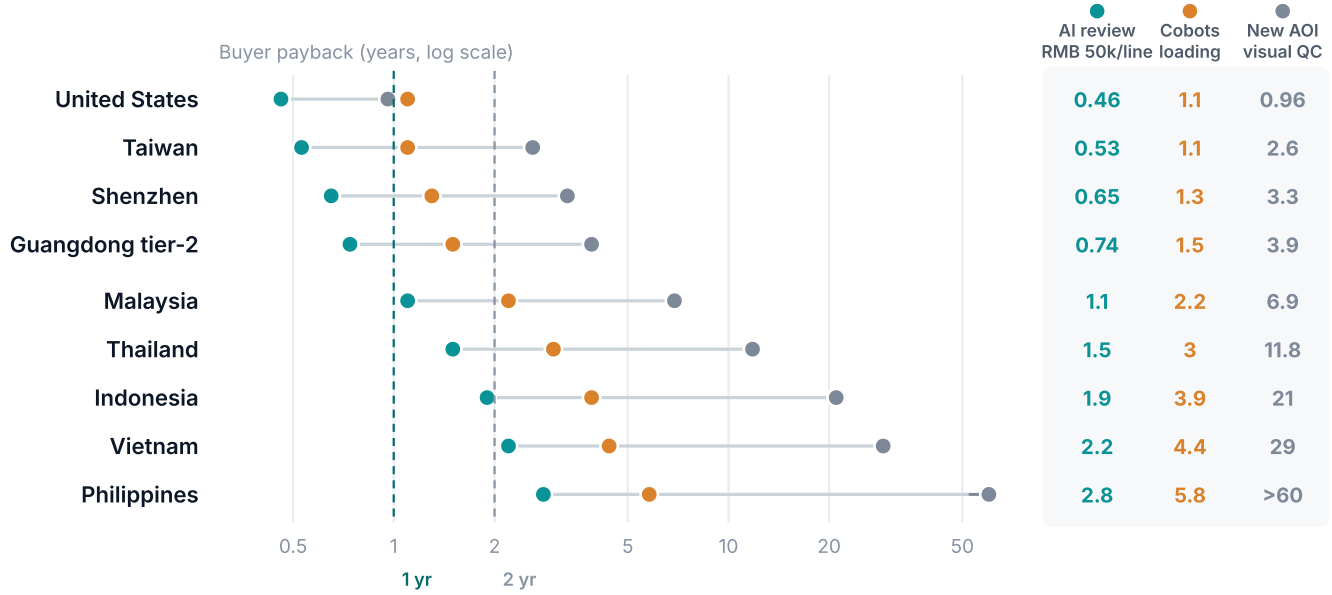
Simple payback = investment ÷ (people saved × loaded labour cost − annual running cost), for a two-shift line with no subsidies or tariffs. Taiwan equipment is priced at 1.3 times China, Vietnam, Malaysia and Thailand at 1.15 times, Indonesia and the Philippines at 1.2 times, and the US at overseas prices estimate.

Scenario	Investment (China / US)	People saved	Annual running cost
(a) AI review box or AI camera box	RMB 50k / $30k per line	1.0 per line	15%
(b) New inline 2D AOI replacing manual visual checks (0.5 reviewer per shift remains)	RMB 200k / $60k	1.0	12%
(c) Cobot loading cell (one station × two shifts)	RMB 200k / $140k	2.0	8%
(d) Integrated detect-and-pick cell (inspection 1.0 + picking 0.07)	RMB 300k / $200k	1.07	8%
(e) A robot arm added only to pick out NG boards	RMB 150k / $100k	0.07	8%

Region	(a) AI review	(b) New AOI	(c) Cobot cell	(d) Detect-and-pick	(e) Pick-out arm
Shenzhen	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)	Never
Guangdong tier-2	0.74	3.9	1.5	5.3	Never
Taiwan	0.53	2.6	1.1	3.6	Never
Malaysia	1.1	6.9	2.2	9.2	Never
Thailand	1.5	11.8	3.0	15	Never
Indonesia	1.9	21	3.9	25	Never
Vietnam	2.2 (1.6–4.5)	29	4.4	34	Never
Philippines	2.8	>60 (running cost ≈ savings)	5.8	≈100	Never
United States	0.46	0.96	1.1	3.4	Never

FIG 19 **Payback by region**

AI review pays back in months in the US, Taiwan and Shenzhen and in 1–3 years in Southeast Asia; cobot tending in years.



	AI review RMB 50k/line	Cobots loading	New AOI visual QC
United States	0.46	1.1	0.96
Taiwan	0.53	1.1	2.6
Shenzhen	0.65	1.3	3.3
Guangdong tier-2	0.74	1.5	3.9
Malaysia	1.1	2.2	6.9
Thailand	1.5	3	11.8
Indonesia	1.9	3.9	21
Vietnam	2.2	4.4	29
Philippines	2.8	5.8	>60

A robot arm just to pick out NG boards never pays back, anywhere

With first-pass yield above 90%, picking takes only 0.003–0.14 FTE per shift

Source: Author's model: RMB 50k AI review box per line; cobot tending cell at RMB 200k or $140k, two shifts.

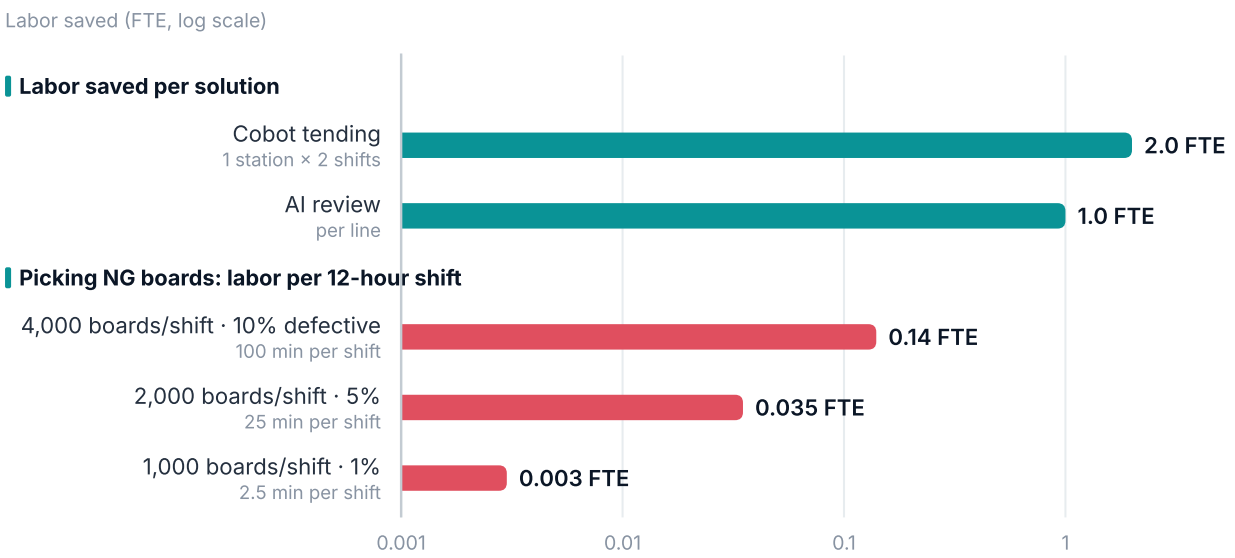
All values are estimates from the inputs above. Sensitivities: halve the people saved and every payback roughly doubles; run one shift and (b)–(d) double; 100% first-year depreciation in the US shortens payback by about 20%, while tariffs on Chinese, Japanese or Korean equipment lengthen it by about 10–40%; a 15% local subsidy in China shortens it by about 15%. I found no independent survey of buyers' payback thresholds; vendors say AOI pays back "typically in 9–18 months" (Averroes) and the Siemens case claims eight months vendor claims. I use one and two years as screening lines. The model leaves out quality gains (warranty, returns) and turnover and training costs, so it is conservative.

The counter-intuitive part of "detect and pick" is that the labour is in review, not in taking the few real NG boards off the line. At 15 seconds of handling per NG board estimate:

Boards per shift	Defect rate	Picking time per 12-hour shift	People
1,000	1%	2.5 min	0.003
2,000	5%	25 min	0.035
4,000	10%	100 min	0.14

FIG 20 **Why picking out bad boards does not pay**

Above 90% first-pass yield, picking takes only 0.003–0.14 of a person per shift.



Source: Author's estimates.

The base case (2,000 boards per shift, 5% defective, two shifts) is 0.07 of a person, worth about RMB 6,000 a year in Shenzhen, less than the arm's roughly RMB 12,000 annual running cost, so it never pays back anywhere; even the worst case (4,000 boards per shift, 10% defective) takes about 13 years in Shenzhen and 9 in the US estimate. If the worry is stopping the line to pick, a standard NG diverter or buffer conveyor solves it without an arm. People in the industry say "above 90% first-pass yield, stop the line and take the board out by hand", and **that is right for picking, but not for inspection**. The product design on an SMT line is therefore "inspect plus mark or divert", with robot arms kept for places where every part is handled (test-fixture loading, tray and magazine handling, packing), where inspection becomes a nearly free add-on and scenario (c) applies.

Subsidies, tax and tariffs move payback by 15–25%, but not the ranking

Mainland China: support mostly lowers financing cost; single machines rarely qualify directly. A national loan subsidy pays 1.5 points of interest for up to two years in 14 priority chains (Yicai) verified, about RMB 9,000 on a RMB 300k machine; Guangdong's upgrade reward of up to 20% needs a registered project with at least RMB 7 million of new equipment and excludes Shenzhen (Guangdong rules), and Dongguan actually paid 4.42% in July 2026 (Dongguan allocation) single source, primary. An AI inspection device is approved only as equipment "with supporting software" inside the buyer's registered project, so the base case assumes no subsidy.

Taiwan: incentives go to the buyer. Under article 10-1 of the Statute for Industrial Innovation, amended in May 2025 to include AI, a buyer can credit 5% of 2025–2029 spending on new smart machinery or AI (including CNN/DNN-based inspection) against that year's income tax, or 3% a year for three years (Statute, credit rules) verified; supplier-side grants are closed to mainland capital (MOEA) single source.

Southeast Asia: only Malaysia and Thailand reward buying automation. Malaysia's 200% automation capital allowance on the first RM10 million is worth about 24% of capex to a profitable buyer (MIDA, PwC Malaysia) verified; Thailand's BOI gives a three-year tax exemption up to 50–100% of an upgrade investment, the higher cap when Thai automation suppliers are involved (BOI) single source, primary, so entering Thailand works best with a local integrator. Vietnam, Indonesia and the Philippines reward building plants rather than buying automation (Vietnam Law Magazine, PwC Philippines).

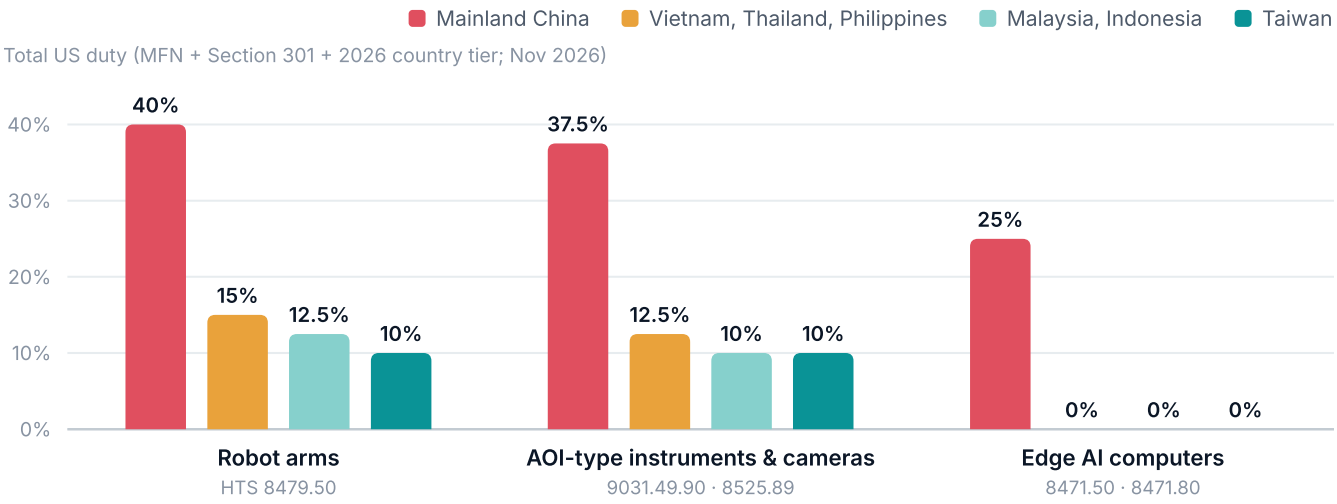
United States: depreciation shortens payback; tariffs lengthen it. Assets bought after January 19, 2025 get permanent 100% first-year depreciation, about a 21% federal tax shield for a profitable buyer (IRS Notice 2026-11) sin-

gle source, primary ; the Section 232 investigation of robots and industrial machinery ended on October 28, 2026 with negotiations rather than new tariffs, to be reassessed within 180 days (White House, Mayer Brown) verified . Final assembly in Southeast Asia cuts the surcharge on arms and AOI-type instruments from about 37.5–40% to 10–15% (Chapter 1).

Region	Main instruments	Effect on buyer payback estimate
Mainland China	1.5-point loan subsidy for up to two years; provincial and city upgrade rewards (Guangdong needs ≥RMB 7M registered projects; Dongguan actually paid 4.42%); Shenzhen 2026–2029 measures (rates not public)	About 0–5% for a single machine; 4–15% inside a large project
Taiwan	Statute §10-1 credit, 5% or 3% × 3 years	About 5% shorter, if the buyer pays enough tax
Malaysia	200% automation capital allowance (first RM10M)	About 15–25% shorter for profitable buyers
Thailand	BOI three-year tax exemption up to 50–100% of investment	About 15–25% shorter for profitable buyers outside a tax holiday
Vietnam, Indonesia, Philippines	Income-tax incentives for building plants; extra 20% equipment depreciation in the Philippines	Almost none for single automation purchases
United States	100% first-year depreciation; ≈37.5–40% combined duty on mainland-made equipment	Depreciation ≈20% shorter; tariffs ≈10–40% longer

FIG 21 **US duty on robots, AOI and edge computers**

Arms made in mainland China pay about 40%; shipped from Southeast Asia or Taiwan, 10–15%.



Source: USITC Harmonized Tariff Schedule chapter 99; Federal Register; USTR (November 2026).

By region: in mainland China, an AI review box pays back in months and a cobot cell in 1.3–1.5 years, while a new AOI bought only to cut labour takes 3–4 years and integrated detect-and-pick 4.5–5.3, both beyond a two-year screen. Taiwan does slightly better than Shenzhen (0.53 years for review, 1.1 for a cobot cell), and hiring there is harder. In Malaysia and Thailand, an AI box takes 1–1.5 years and a cobot cell 2–3, about 15–25% faster with local tax incentives for profitable buyers. In Indonesia, Vietnam and the Philippines only a very cheap AI box pays back within 2–3 years; demand there comes more from customers' quality requirements and labour shortages. In the US every option except a pure pick-out arm pays back in 0.2–3.4 years, with tariffs, the Section 232 reassessment and NRTL certification as the main variables.

05 MARKET AND COMPETITION
Market, competition and how buyers pay

Market size: published numbers differ by multiples; only a bottom-up count is useful

Every market number has to be read with its scope. China's machine-vision market was RMB 18.15 billion in 2024 (Zhiyan) verified, with 3C electronics the largest buyer; global estimates for machine vision differ about threefold and for AOI about 6.5-fold (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Bottom-up from the revenues of Koh Young, TRI, Jutze, Sinic-Tek and ViTrox, global SMT inspection equipment is on the order of $1–2 billion estimate.

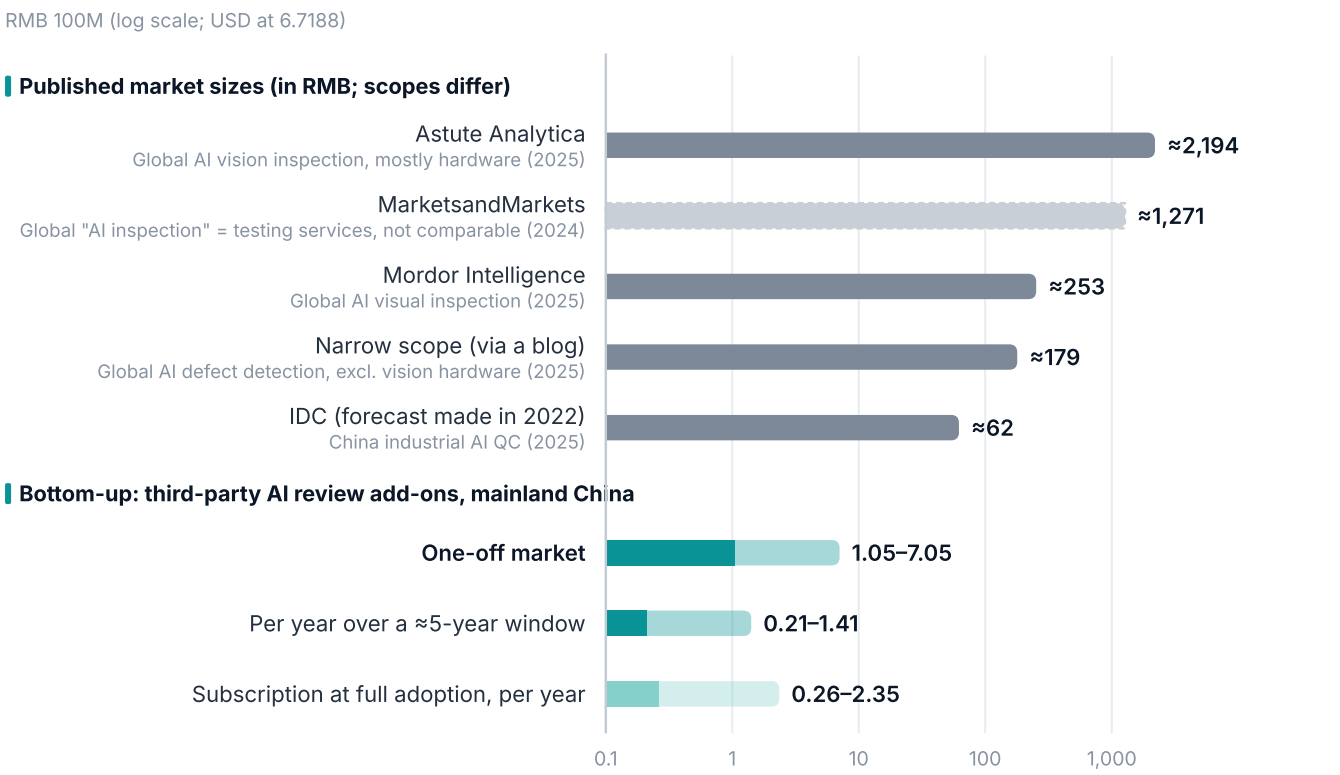
There is no standalone market for "AI visual inspection software". The published numbers differ by more than ten times and all include hardware or services:

Scope	Publisher	Figure
Global "AI industrial defect detection" (narrow, excluding vision hardware)	Unnamed firm, via an iFactory blog	$2.66 billion in 2025, 8.6% CAGR to 2035, electronics 34%
Global "AI visual inspection"	Mordor Intelligence	$3.76 billion in 2025, $12.8 billion in 2031, 22.67% CAGR
Global "AI visual inspection" (mostly hardware)	Astute Analytica	≈$32.66 billion in 2025
Global "AI inspection" (testing and certification services; not comparable)	MarketsandMarkets	$18.92 billion in 2024
China "industrial AI quality inspection"	IDC, forecast made in 2022	$958 million in 2025 (≈RMB 6.2 billion), 91.5% in communications, electronics and two other industries

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

FIG 22 **Market sizes differ by more than ten times**

Only the narrow scope and a bottom-up count are useful for decisions.



Source: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; author's bottom-up estimate.

Bottom-up, there is only one public count of SMT lines: New Venture Research estimated about 73,231 lines worldwide in 2020, 47% of them in China (19% in 2000), about 70% high-end and 30% low-end (EMSNow, NVR SMT2020 summary) single source, 2020 ; NVR now belongs to IPC and newer data is paid (NVR). Growing at 3–5% a year with a 40–47% China share gives about 87,000–98,000 lines worldwide in 2026 and 35,000–46,000 in mainland China (base 40,000) estimate .

Input	Low	Base	High	Basis
Mainland China SMT lines (2026)	35,000	40,000	46,000	NVR 2020 base × growth × share
Share without AI review	75%	80%	85%	Retrofit window, below
Share a third party can reach (not in-house, not tied to OEM AI, can pay)	40%	50%	60%	Assumption: ≈30% low-end lines plus some mixed-brand high-end lines, excluding DBG-style in-house builds
Serviceable lines	**10,500**	**16,000**	**23,500**	Product
One-off price per line	RMB 10k	RMB 20k	RMB 30k	Price ladder and ceiling (Chapter 4)
One-off serviceable market (installed base)	**RMB 105M**	**RMB 320M**	**RMB 705M**	Product
Per year over a ≈5-year window	RMB 21M	RMB 64M	RMB 141M	Window shrinks 8–12% a year
Subscription (RMB 2.5k / 5k / 10k per line per year, full adoption)	RMB 26M a year	RMB 80M a year	RMB 235M a year	Baidu-style annual fee anchor

FIG 23 **Bottom-up: 40,000 lines → 16,000 serviceable → ≈RMB 320M one-off**

The third-party AI review niche in mainland China is real, but small.



A newcomer winning 2–5% of serviceable lines in 3 years: ≈320–800 lines, ≈RMB 2–5M a year

As a subscription (RMB 5k per line a year): RMB 1.6–4M recurring, only 0.3–2.3% of IDC's China industrial AI QC forecast

Source: NVR (2020 base); author's estimates.

All of these are estimates; neither share nor the price has been measured publicly. The niche is worth only 0.3–2.3% a year of IDC's forecast for all of China's industrial AI quality inspection in 2025 (about RMB 6.2 billion). A newcomer winning 2–5% of the base serviceable lines in three years would have about 320–800 lines: RMB 6.4–16 million in total at RMB 20,000 per line, about RMB 2–5 million a year, or RMB 1.6–4 million a year in recurring revenue at RMB 5,000 per line estimate. That is enough for a business unit inside a company, not for a venture-scale standalone company. Each 1,000 installed lines in Taiwan, Southeast Asia or the US is worth roughly RMB 3–15 million one-off at mainland prices; prices are higher in Taiwan and the US, but buyers there more often use OEM 3D AOIs with AI options estimate.

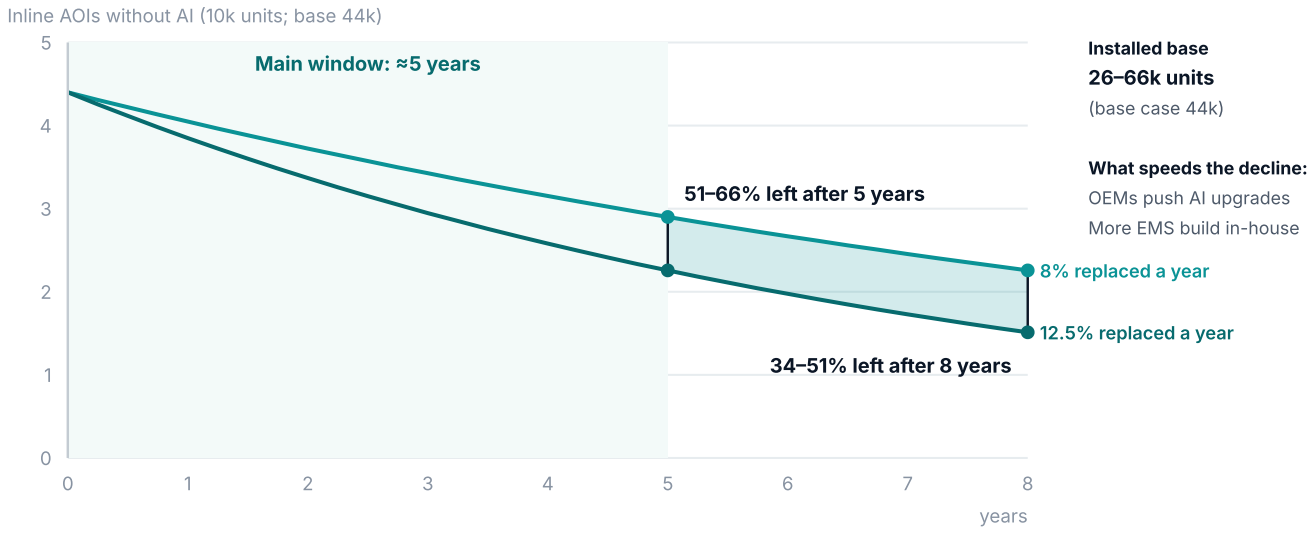
The retrofit window is about five years

No AOI maker publishes machine life; EMS companies depreciate equipment over 3–10 years (DBG 2025 annual report), and older AOIs lose software updates and spares after end of life (I-Connect007) single source. I use 8–12 years (base 10).

At 1.0–1.65 inline AOIs per line, the mainland has about 35,000–78,000 inline AOIs in service (base about 55,000), consistent with known shipments (Circuits Assembly, Sinic-Tek 2025 annual report). No source measures how many run AI review; assuming domestic machines have offered it since about 2023 and 40–60% of recent deliveries use it, **about 26,000–66,000 inline AOIs in mainland China (base about 44,000) do not run AI** estimate. If every new machine now ships with AI, that pool shrinks 8–12.5% a year: 51–66% remains after five years and 34–51% after eight, faster if AOI makers push AI upgrades to newer machines. **Most of the opportunity falls in the next five years or so**, with mid-size EMS running older domestic 2D AOIs, mixed brands and no AI upgrade path at its core.

FIG 24 **The retrofit window is about five years**

Inline AOIs without AI shrink by 8–12.5% a year as they are replaced or upgraded.



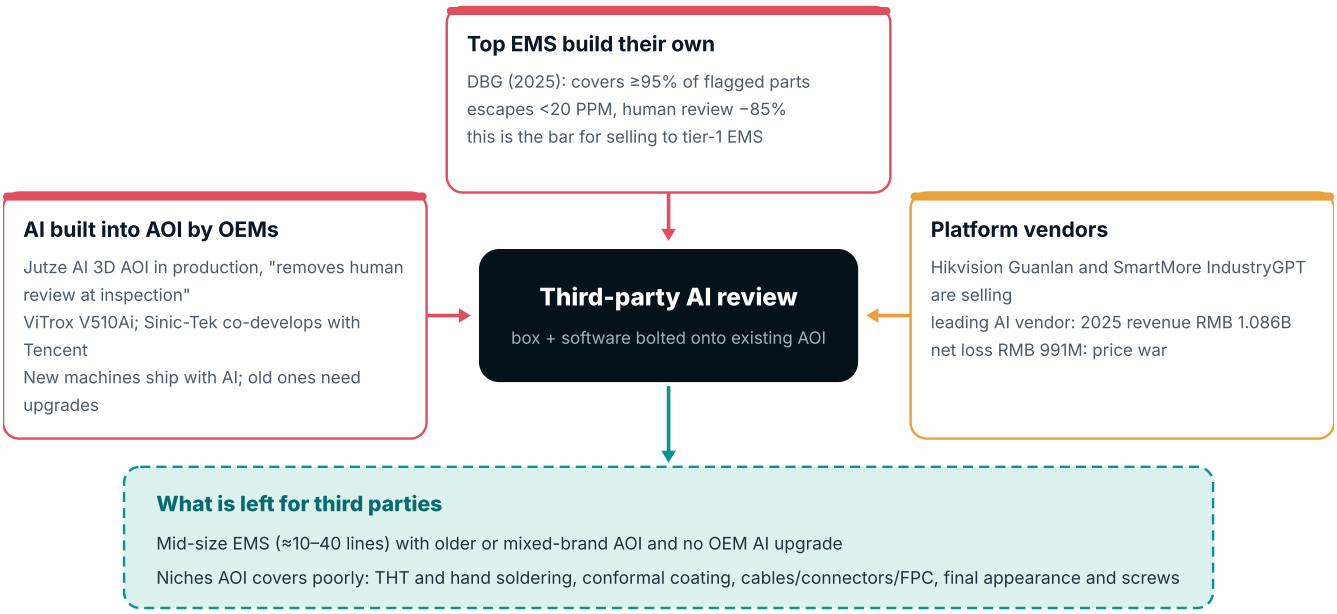
Source: Author's estimates; installed base and shares not independently verified.

The competitors are already strong: the review wedge is encircled

The AI review wedge is narrowing from three directions. AOI and vision leaders are building AI into new platforms: Jutze has "fully adopted AI algorithms" (Jutze investor relations, June 2026), OPT's industrial AI revenue grew 96.77% in 2025 (OPT 2025 annual report) single source, primary, and ViTrox claims up to 95% accuracy for autonomous review (ViTrox) vendor claim. Large manufacturers build their own (iThome). And venture-backed AI companies sell at scale: SmartMore's edge-sensor customers grew from 65 to 251 in two years (36Kr) single source, Hikvision sells its Guanlan model for inspection (Xinhua), and "eye, brain and hand" integration is already on offer from SmartMore, Hikrobot and Mech-Mind (36Kr, Securities Star).

FIG 25 **The review wedge is encircled on three sides**

Room for third parties is in mid-size EMS and in inspection niches the AOI covers poorly.



Source: Jutze, Sinic-Tek and DBG annual reports; Hikvision and SmartMore public material; compiled by the author.

The space left for a newcomer is narrow: mid-size EMS with older 2D/3D AOIs and no AI upgrade from the maker; lines mixing several AOI brands (one retrofit vendor already supports Omron, VI, Koh Young and Mirtec (Sohu on Hengtianyi)); niches SMT AOIs cover poorly, such as through-hole and hand-soldered joints, conformal coating, cables, connectors and FPCs, and final appearance and screws; and small plants that cannot afford a RMB 230–280k AOI.

The closest listed comparable gives a warning. Weiyi Zhizao, which sells AI inspection equipment and robots, saw 3C revenue fall 79.9% year on year in January–May 2026, acknowledged "intensifying competition and pricing pressure", postponed its Hong Kong listing on October 25, 2026, and disclosed 15 projects terminated before completion, 12 of them embodied-robot projects (Weiyi Zhizao prospectus, postponement notice) single source, primary. AInnovation ended 2025 with 29.8% of its receivables provisioned (AInnovation 2025 annual report). The 3C inspection wedge is crowded, and pricing power per line is weak.

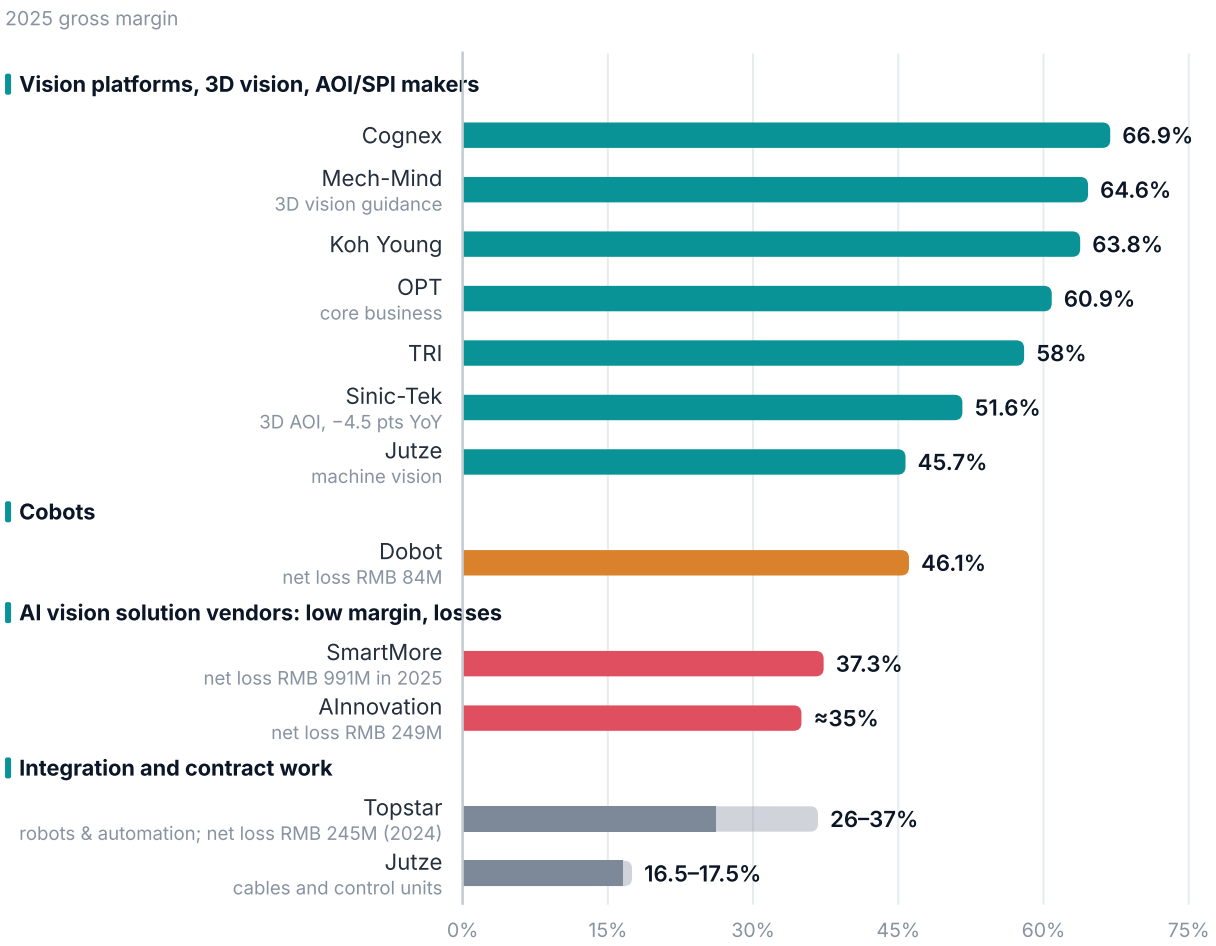
The margin sits with productized hardware

Position in the chain	Representative companies and 2025 gross margin	Profit and collections
Vision platforms	Cognex 66.9%	Profitable; divesting ≈$22M of low-margin revenue
AOI/SPI equipment	Koh Young 63.8%; TRI 58%; OPT core business 60.85%; Sinic-Tek 3D AOI 51.62% (−4.53 points YoY); Jutze machine vision 45.74%	Profitable; receivable days 139 (Jutze) to 247–274 (OPT)
3D vision guidance	Mech-Mind 64.6%	—
AI vision solution vendors	SmartMore 37.3%; AInnovation ≈35%	Loss-making: SmartMore net loss RMB 991M in 2025, AInnovation RMB 249M
Cobots	Dobot 46.1%	Dobot net loss RMB 84M; UR revenue from $326M (2022) to ≈$232M (2025, forecast)
System integration	Topstar robotics and automation 26.2–36.7%	Net loss RMB 245M (2024); receivable days 180–240
Contract hardware	Jutze cables and control units 16.5–17.5%	—

Sources: (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 **The margin sits in productized hardware**

Vision platforms and equipment makers earn 46–67% gross margin; AI solution vendors 35–37% and lose money.



Source: 2025 annual reports or prospectuses.

A newcomer keeps its margin only by turning inspection into a product, equipment or software; project-by-project tuning pushes margins to about 35% with persistent losses. The mechanics also work against small companies: revenue is recognized only after on-site acceptance, prepayments are small (Jutze, OPT), pilots are usually free (A3/Mindtrace), and brand owners often specify the inspection equipment on their contract manufacturers' lines.

How buyers pay: 30% at signing, 60% after acceptance, 10% held for a year

The payment terms mainland electronics plants use for inspection equipment are spelled out in 13 named buyer contracts disclosed in Sinic-Tek's listing inquiry reply (Sinic-Tek inquiry reply 8-1-1) single source: one primary filing, several independent buyers, 2020–2022 :

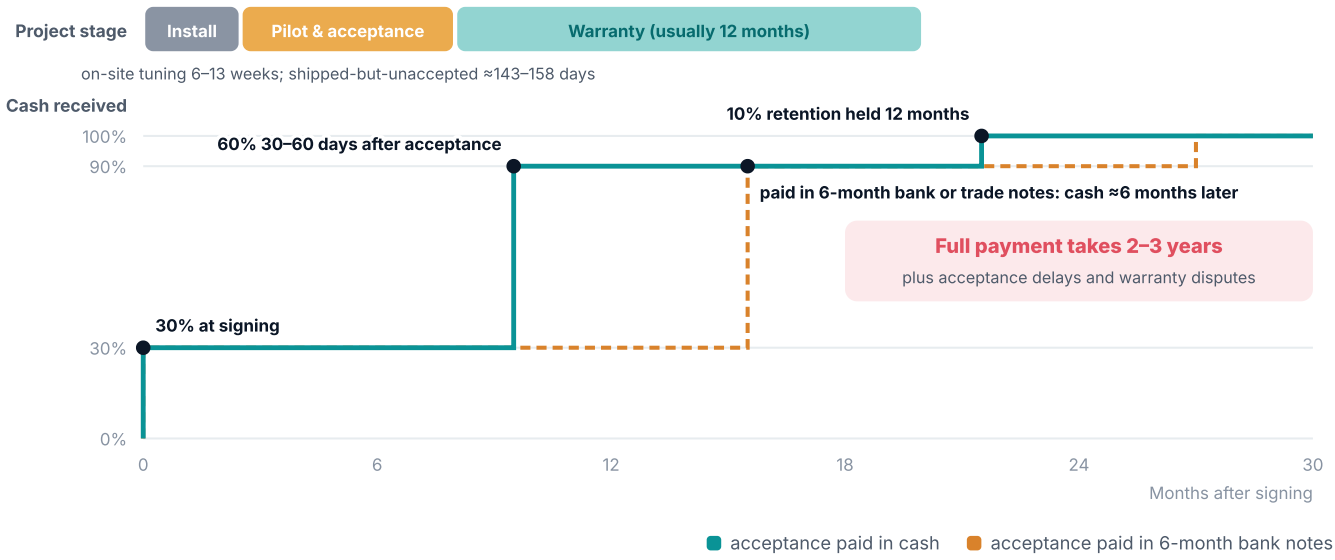
Buyer (year)	Signing	Delivery	Acceptance	Retention or balance
A Hikvision unit (2021)	30% within 15 days	—	65% within 30 days of final acceptance	5% warranty retention paid 12 months after acceptance; part paid in 6-month+ bank acceptance notes
Nanchang Huaqin (2020–2021)	30% within 10 working days	—	60% at 60 days after acceptance	10% after a one-year warranty
Guangdong Yinuo (2021, 3D AOI)	30% within 10 working days	—	60% within 30 days	10% after three months of fault-free running

Buyer (year)	Signing	Delivery	Acceptance	Retention or balance
Xiamen Hongxin, Jingmen Hongyi (2022)	30% within 10 working days	—	60% within 30 working days of final acceptance, 6-month bank notes	10% after a one-year warranty
AKM Panyu, Suzhou (2021–2022)	30% within 7–10 working days	—	60% a month after acceptance and invoicing, 180-day bank notes	10% after 360 days
BYD Precision, Shenzhen, Huizhou (2020)	—	30% on arrival	70% after acceptance and invoicing	All in 6-month commercial notes
Shenzhen H&T (2020)	—	—	90% after acceptance and invoicing	10% after six months of fault-free use

Other primary filings show the same structure: Mech-Mind's large projects pay 30% at signing, 30% before delivery, 30% after acceptance and 10% warranty retention over 12–15 months (Mech-Mind prospectus); GKG says retention in the equipment industry is usually 5–10% of the contract, paid a year after acceptance (GKG inquiry reply) single sources, primary. The norm, then, is about 30% at signing (20–30%, zero at some plants such as BYD), 60–65% after acceptance (paid 30–60 days later, often in six-month bank or commercial notes, which delays usable cash by another half year), and 5–10% retention for 3–12 months (12 most common) verified across several contracts.

FIG 27 **How buyers pay**

30% at signing, 60% after acceptance, 10% held for a year: full payment takes 2–3 years.



Source: Sinic-Tek inquiry reply, Mech-Mind, GKG, JT and other filings; schematic on usual terms, timing estimated.

Acceptance and trials are slow and paid for by the seller. Sinic-Tek's shipped-but-unaccepted goods took about 143–158 days to turn over, because large EMS and brand customers need whole-line integration testing and an extended trial run (Sinic-Tek inquiry reply 8-1-2); Weiyi's AI projects take 4–8 months from evaluation to deployment (Weiyi Zhizao prospectus) single sources, primary. There are no conversion data for AI review pilots; Sinic-Tek converted only about 18–19% of its demo machines directly into sales. I plan on one in three to one in five for free pilots and about 80% for paid pilots after shortlisting estimate.

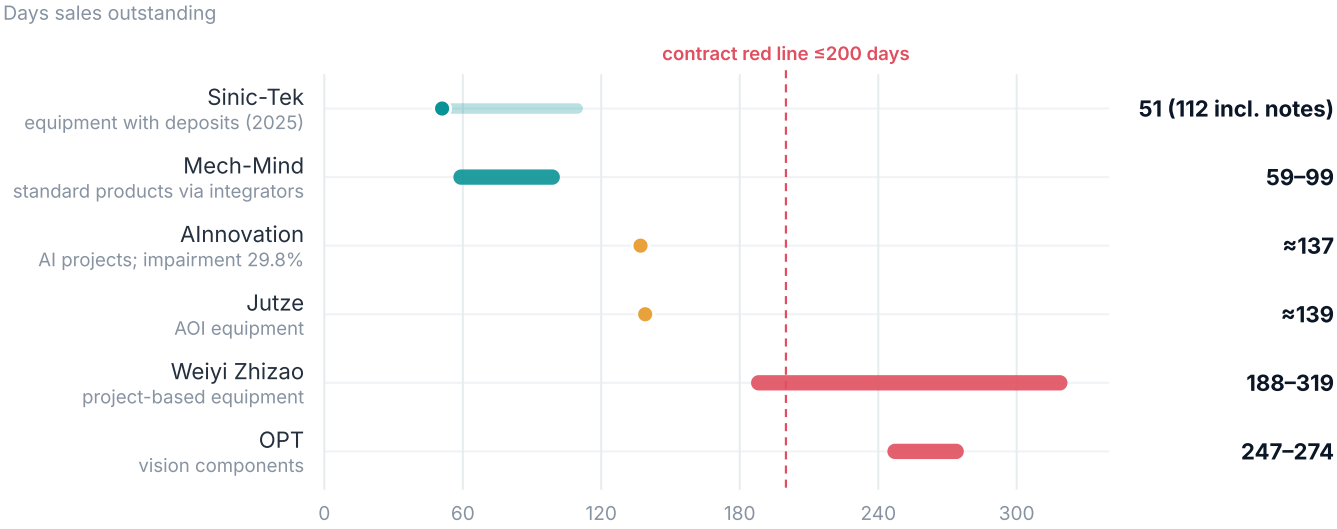
Company	Business model	Days sales outstanding
Sinic-Tek (2025)	Equipment with deposits	≈51 (≈112 including notes)

Company	Business model	Days sales outstanding
Mech-Mind	Standard products through integrators	59–99
Jutze	AOI equipment	≈139
AInnovation (2025)	AI projects	≈137 at year end, with 29.8% impairment
Weiyi Zhizao	Project-based equipment	188–319
OPT	Vision components	247–274

Sources: the annual reports and prospectuses above; receivables (including contract assets) ÷ revenue × 365 primary + estimate.

FIG 28 **Days sales outstanding**

Standard products 51–99 days; project-based equipment 188–319 days.



Source: Company filings; receivables (incl. contract assets) ÷ revenue × 365.

AI review licences sold through AOI makers or integrators should collect like Mech-Mind's 60–100 days; direct project sales like Weiyi's 190–320 days. Put together as a cash timeline: free pilot one to three months, qualification at a large ODM up to about twelve months, build, delivery and acceptance three to seven months; 30% arrives at the order, 60% about four to nine months after it and the last 10% about 16–21 months after it. **From the first pilot at a large EMS to full payment can take two to three years**; mid-size EMS without formal tenders can shorten qualification to weeks estimate. The acceptance criteria in the disclosed contracts cover only whole-line integration, throughput (one Taiwanese project required 100 units per hour) and months of fault-free running; none specifies escape or false-call rates (Anda inquiry reply) single source, primary. Written acceptance criteria for AI review have to be agreed before the pilot.

Four patterns of failure

Documented failures fall into four patterns: shipping before the product was reliable (Rethink (The Robot Report)); a software layer on a leader's hardware that the leader's ecosystem later replaced (READY Robotics (The Robot Report)); running out of money at the bottom of a cycle (Dextrous Robotics (Robots & Startups)); and model companies ending in acqui-hires (Covariant (GeekWire)). The lessons: productize narrowly instead of tuning project by project; hold working capital for acceptance-based revenue and long collections; provide local service for plants that run 24 hours; and do not scale before the product is proven.

Timing: inspection gains nothing from waiting; arm hardware gets cheaper

The arguments for waiting are weak. Hardware did not get cheaper in 2026: robot makers raised prices 5–15%, some products 25–50% (Sina Finance) single source, and NVIDIA kits rose a third time (36Kr). Production-grade VLAs will not arrive within the window (IFR, Instrumental). Meanwhile every AOI replacement removes a retrofit target and every OEM AI deployment locks in a line's data and process; model capability is commoditizing, and what is scarce is access to plants and labelled factory data. Local support has windows too: Shenzhen's 2026 "AI pioneer city" rules, with up to RMB 2 million for AI application demonstration projects and model vouchers covering 30% of purchases, run until **June 30, 2027** (Yicai, May 2026) terms single source .

What is still getting cheaper fast is humanoids and dexterous hands (hands fell from about RMB 50,000 to under 10,000 in two years (Sina Finance)), and entry cobots as the product mix moves down; that argues for deferring the embodied part, not the inspection part. **Overall: inspection now, as a narrow product; robot cells now, through partners and integration; embodied AI later, with only an R&D station today.**

EMBODIED AI

06 Embodied AI: first it changes how robots are taught, later how they move

What is changing is perception and programming; motion stays deterministic

Every robot-arm leader now markets “physical AI”, but what has shipped falls into three layers: learned perception without teaching (FANUC’s AI Masterless Vision, shipping in January 2027 (FANUC); KUKA’s neural grasp selection (KUKA)); large models that write programs and plan tasks (KUKA’s iiQWorks.Copilot, which warns that AI-generated code should be verified in a digital twin (KUKA); Yaskawa with Gemini Robotics-ER, where “generative AI decides what to do” and the robot decides “how to move” (Yaskawa)); and compute and data platforms (UR’s AI Accelerator with a Jetson Orin (Universal Robots)) vendor claims . No leader publishes production success rates for its AI features, and the Japanese leaders partner with Google, SoftBank, Microsoft and others rather than building frontier VLAs (FANUC news index): the learned policy layer will come from model companies, the certified arms and controllers from the arm makers.

FIG 29 **The embodied-AI stack**

Large models plan offline, small policies run in the cell, motion and safety stay deterministic.

Layer	What it does, and how far the evidence goes	Time scale · who owns it
Planning L4	**Frontier models (offline)** write or change robot programs offline, changeovers, explain faults Rewrites ABB RAPID programs at 77–99.7%, deployed after simulation Control loop only 0.2–0.4 Hz: stays out of real time	**seconds to minutes** Model giants
Skills / policy L3	**ACT, diffusion policies, small VLAs** Fixtured, structured stations 90–99.4% Edge chips: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**action chunks 100–300 ms** Chip vendors · robot makers
Motion control L2	**Robot controller** trajectories, servo, force control, collision checks deterministic code; learned policies only set goals	**milliseconds (real time needs >83 Hz)** Robot makers
Safety L1	**Safety PLC, scanners, force and speed limits** certified to ISO 10218-1/-2:2025 and local rules No ML in safety functions: EU notified-body review from Jan 20, 2027	**hard real time** Certified robot + integrator

Small firms win in integration and sensing hardware: put learned perception and small policies into certified robot cells, one station at a time

Source: Author's synthesis (Chapters 3 and 6); ISO 10218:2025; EU Machinery Regulation.

How far from unstructured, flexible, small-batch work: structured stations in 1–3 years, flexible assembly 5+

Real factory numbers have started to appear, but they are vendor-reported, narrow and from structured stations. AgiBot’s eight humanoids on Longcheer’s tablet final-test line logged 64,828 tasks at 99.99% over more than 64 hours (AgiBot, Yicai Global) vendor and customer , but this is tending work a conventional arm with vision can do. UBTech says its Walker S2 works at 30–50% of human efficiency (Irish Times/FT) single source . Figure 02’s 11

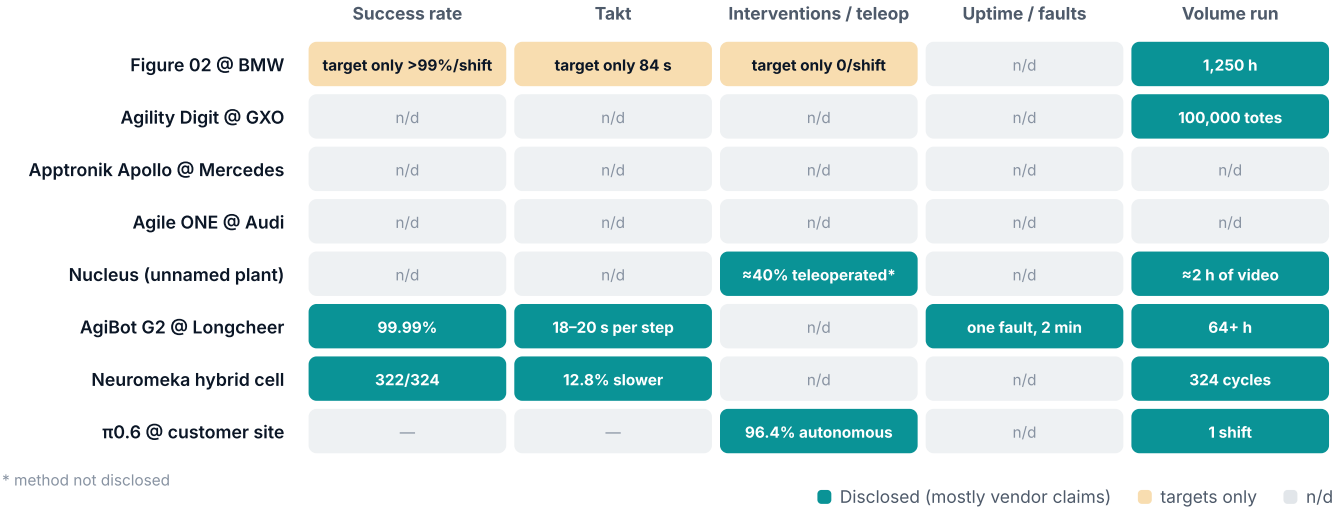
months at BMW Spartanburg produced volume figures (1,250+ hours, 90,000+ parts), while its cycle, success and intervention numbers are targets, not results (Figure, BMW Group) verified, numerators only. Nucleus says about 40% of its factory video is teleoperated (Interesting Engineering) vendor claim, and a Reuters investigation quoted about 20% of a worker's output for some Chinese humanoids (Notebookcheck) single source. No deployment publishes success rate, cycle time and intervention rate together, and none has been audited, yet those three are exactly what a payback calculation needs.

Deployment	Success rate	Takt	Interventions or teleoperation	Uptime or failure interval	Volume run
Figure 02 @ BMW Spartanburg	Target only (>99% per shift)	Target only (84 s)	Target only (0 per shift)	Undisclosed (forearm top failure point)	1,250 h
Agility Digit @ GXO	Undisclosed	Undisclosed	Undisclosed	Undisclosed	100,000 totes
Apptronik Apollo @ Mercedes-Benz	Undisclosed	Undisclosed	Undisclosed	Undisclosed	Undisclosed
Agile ONE @ Audi	Undisclosed	Undisclosed	Undisclosed	Undisclosed	Undisclosed
Nucleus (unnamed plant)	Undisclosed	Undisclosed	≈40% teleoperated (method undisclosed)	Undisclosed	≈2 h of video
AgiBot G2 @ Longcheer	99.99% of 64,828 tasks	18–20 s per step	Undisclosed	One communication fault, ≈2 min to recover	64+ h
Neuromeka hybrid cell	322/324	12.8% slower than a person	Undisclosed	Undisclosed	324 cycles
Physical Intelligence π0.6 at a customer site	—	—	96.4% autonomous over a shift	Undisclosed	One shift

Sources: this chapter and Chapter 3 mostly vendor claims.

FIG 30 **What in-plant robot deployments actually disclose**

Most disclose only volume; none discloses success rate, takt and intervention rate together.



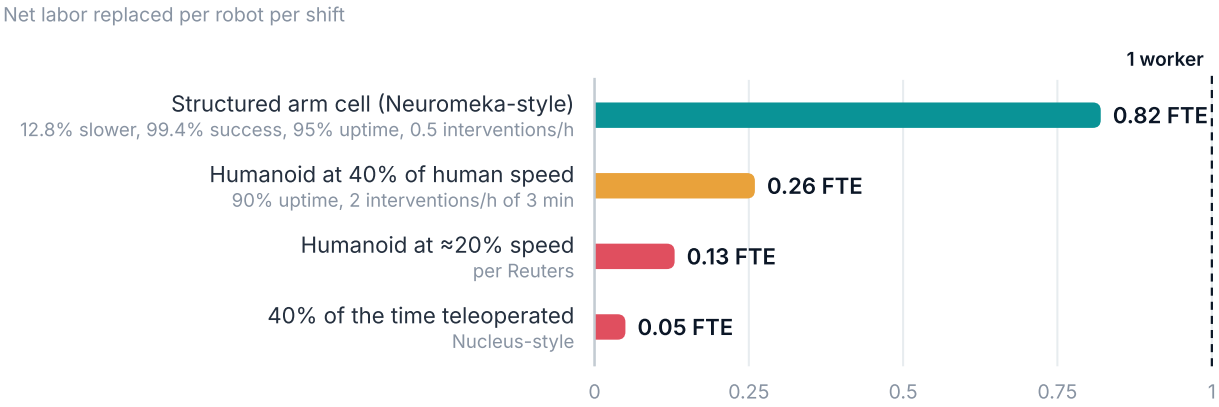
Source: Public material from BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence and others.

Teleoperation matters more to the economics than success rate. Net people replaced per robot per shift ≈ (human cycle ÷ robot cycle) × uptime × success, minus intervention time and teleoperation estimate, my formula. A Neuromeka-style structured station (12.8% slower, 99.4% success, 95% uptime, 0.5 interventions of 2 minutes an

hour) replaces about 0.82 people; a humanoid at 40% of human speed, 90% uptime and two 3-minute interventions an hour about 0.26; at Reuters' 20%, about 0.13; and if, like Nucleus, one operator teleoperates it 40% of the time, about 0.05, almost nothing. At about RMB 600k per humanoid and RMB 85k per person per year on the mainland, the last three pay back over decades; only the structured arm station has a realistic payback (Chapter 4) estimate.

FIG 31 **Net labour replaced per robot**

A structured arm cell replaces about 0.82 of a worker; a humanoid teleoperated 40% of the time, about 0.05.



Source: Author's formula-based estimates.

Away from structured stations, general models still score 40–90% on new tasks: Gemini Robotics 2 reaches 74.2% on lab pick-and-place, 78.9% on tool kitting and 44% on tying a bag (eesel/DeepMind); on RoboChallenge Table30, 30 real-robot tasks, π0.5 scores 43.7% and the best open model 50.3% (RoboChallenge, QbitAI).

Outside forecasts differ, but none supports widespread use within two years. The IFR says general physical AI is "not yet industrial-grade" and most humanoids are "still prototypes or pilots" (IFR); Interact Analysis estimates that only about 10% of humanoids produced in 2025 went into real use, with the commercial turning point "after 2032" (automation.com); UBS sees a humanoid "EV moment" as unlikely before 2030 (Yicai Global); Chinese insiders estimate two to three more years before factory-scale use (Gasgoo) single sources. My judgement estimate: in well-resourced electronics and automotive plants with the vendor on site, **structured station tasks** (tending, OK/NG sorting, fixtured insertion) will reach ≥99% at close to human speed in about 1–3 years; **mixed-part kitting** with general VLAs in about 3–5; **free-routed harnesses and multi-step flexible assembly** in more than 5. Humanoid payback does not work today: a Walker S2 at about RMB 600k (10jqka) single source, at 30–50% of human efficiency over two shifts, equals about 0.6–1.0 workers, a simple payback of 7–12 years at mainland labour costs before integration, maintenance and downtime estimate.

Where it is stuck

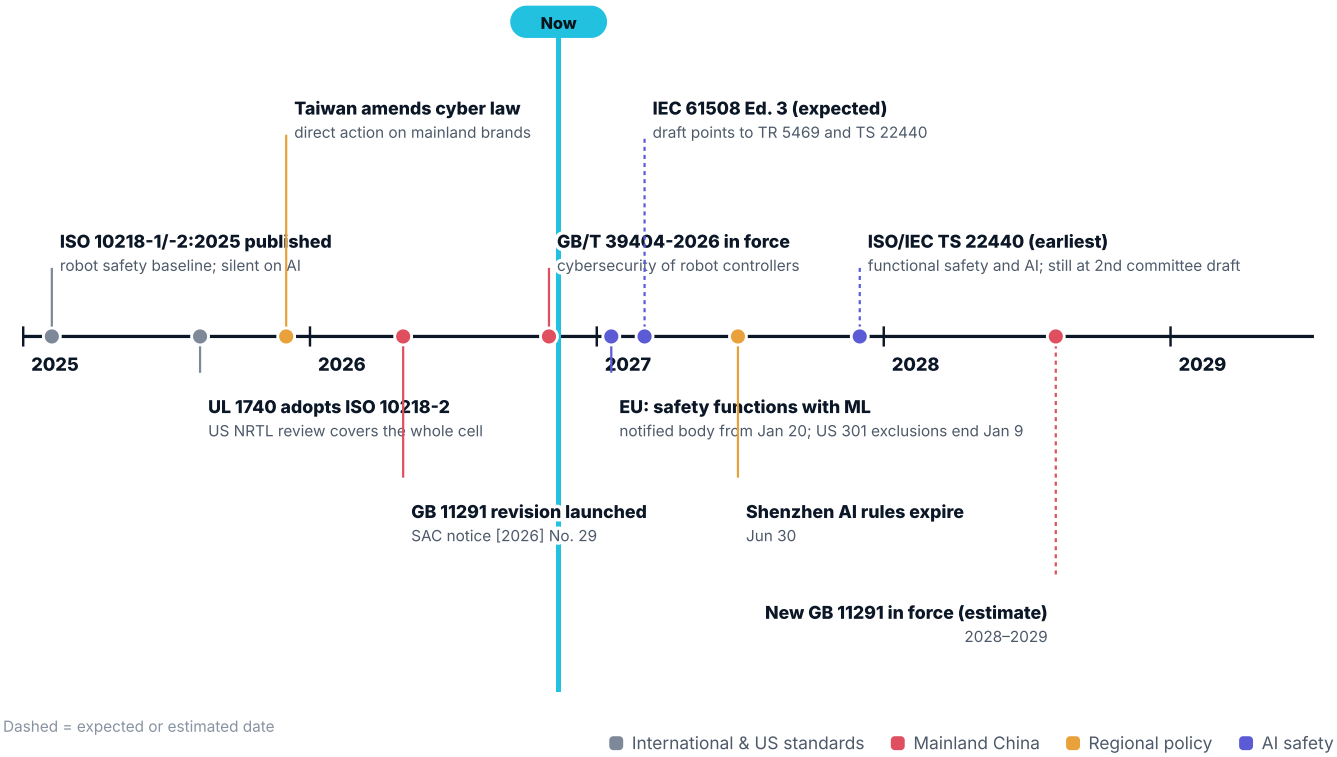
Link	Evidence in 2026
Data	The industry estimates at least 10 million hours of high-quality real data are needed and only about 500,000 had been collected by early 2026 (Gasgoo); Reuters reports about 500,000 hours against about 100 million needed (Notebookcheck); a novice teleoperator needs about 300 attempts per usable motion; plants will not share process data; simulation-trained policies drop sharply on real robots
Reliability	Structured stations 98–99.99% (vendor claims, failure undefined); general models 40–90% on new tasks; unseen connectors 29–67%
Speed	Humanoids at 30–50% of human speed (UBTech), some at about 20% (Reuters); frontier agents about 280 s per quasi-static real-robot task; LLM control loops at 0.2–0.4 Hz against about 83 Hz needed (Anthropic)
Cost	Walker S2 averages about RMB 600k; 7–12-year payback estimate

Link	Evidence in 2026
Safety certification	ISO 10218-1/-2:2025 makes functional safety more explicit and absorbs collaborative requirements, but says nothing about AI (The Robot Report); ISO/IEC TS 22440 (functional safety and AI) was still a second committee draft in November 2026 (ISS, ISO Open Data); IEC 61508-3 guidance advises against AI above SIL 1 (Analog Devices); under the EU Machinery Regulation, safety functions with machine learning need notified-body assessment from January 20, 2027 (EUR-Lex)
Hands and tools	Dexterous hands manage about 30 N (≈3 kg) per finger against about 10 kg for a human hand (Gasgoo); tool changes are still manual
Integration	Sensor misreads, protocol mismatches and MES links on site; intervention rates almost never disclosed

The safety implication is direct: the standards neither forbid learned policies nor recognize them. ISO 10218-1/-2:2025 says nothing about AI; ISO/IEC TS 22440 (functional safety and AI) is still at committee draft and unlikely before late 2027 or 2028 estimate ; ISO/IEC TR 5469:2024 is guidance only, recommending non-AI fallbacks, supervisory constraints and redundant voting; IEC 61508 edition 3 is expected in early 2027 (ISO Open Data, BSI, 61508 Association) verified . No published standard yet offers an accepted route to certifying machine-learning perception as a safety function. Regional rules are moving too: China is revising GB 11291 (in force about 2028–2029) (revision plan 20262543); UL 1740 has brought ISO 10218-2 into US NRTL review since August 2025; the EU requires notified-body assessment of safety functions with machine learning from January 20, 2027; and no humanoid is certified for fenceless factory use while ISO 25785-1 is still in draft (i-SCOOP). The defensible architecture: machine learning makes only process decisions (“which part”, “is this board good”), while a certified deterministic safety layer (typically PL d / Category 3) and a safety PLC decide speed, force and stops. Treat any “camera AI for fenceless safety” claim as unproven until a PL d or SIL 2 certificate names the machine-learning component.

FIG 32 **Standards and regulation timeline**

A certification path for AI inside safety functions will not settle before 2027–2029.



Source: ISO, IEC, UL, EUR-Lex, SAC and national regulation databases.

Which layer is the opportunity: models belong to the giants, edge boxes to the chip makers, small firms to integration and sensing hardware

The model layer needs capital and data and belongs to the giants (AgiBot trained GO-2 on "tens of thousands of hours" (iFeng Tech); the strongest models are closed or cloud-only). The edge-box layer belongs to chip makers: Jetson Thor is the de facto standard, and on the mainland D-Robotics' RDK boards (D-Robotics) and Rockchip co-processors (CSDN) compete, mostly without published VLA results.

Small firms have two openings. One is sensing hardware that directly buys reliability: touch lifts ContactDP's insertion success from 58.7% to 97.8%, and force/torque sensors and wrist cameras do similar work; data-collection kits sell too (10jqka), though hands, sensors and arms are all falling in price. The other is the integration layer between arm makers and model companies, owning each station's task data, safety envelope and MES link, which arm makers do not sell as a product today. Competitors fall into four groups: arm leaders (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot and others), vision-plus-arm firms (Mech-Mind, Hikrobot, SmartMore), embodied AI companies (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) and automotive-chip entrants.

Buyer acceptance: no number without a denominator

Since no vendor publishes intervention rates, failure intervals or long-run success counts, these belong in the acceptance terms of the contract. Statistically, 3,000 consecutive production cycles with zero failures prove a failure rate of at most 0.1% at 95% confidence; 10 failures in 3,000 prove only ≤0.57%; proving ≤1%, ≤0.5%, ≤0.1% and ≤0.01% with zero failures takes 299, 598, 2,995 and 29,956 cycles estimate, exact binomial. By that standard, Neuromeka's 322/324 has a two-sided 95% interval of 97.8–99.9% and does not prove ≥99%; AgiBot's roughly 6 failures in 64,828 tasks imply 99.98–100%, but only if "failure" is defined the way the buyer defines it estimate. These are the acceptance requirements I would put to any robot-arm or embodied cell; the thresholds draw on Figure's own targets at BMW and on the Neuromeka and AgiBot cases, and they are judgement, not a standard estimate:

Item	Require	Suggested pass threshold
Definitions	Signed definitions of cycle, success, failure (partial success, re-grasp, drop, quality escape), intervention (touch, remote teleoperation, reset, e-stop) and how autonomy is computed	Definitions come before any metric; no percentage without a formula
Interventions and teleoperation	Per robot-hour, by type and duration; mean time between interventions; teleoperation minutes as a share of run time; operators per robot	Structured stations ≤0.5 interventions per robot-hour (≥2 h apart); zero teleoperation during acceptance
Long-run success	≥3,000 consecutive production cycles over ≥3 shifts, ≥2 part numbers and normal lighting and supply variation; failures classified; exact lower bound	Lower bound ≥99.5% for line-stopping inline stations, ≥99.0% offline; quality escapes reported separately
Cycle time	P50/P90/P99 per part including recovery, against ≥30 timed human cycles at the same station	≤1.15 × human P50, or within line takt
Uptime and faults	Run hours, downtime log, faults by component, part-swap time	≥95% uptime over ≥2 weeks; MTTR ≤20 min with spares on site
Recovery	Scripted fault injection: drops, occlusion, lighting change, conveyor stop, MES or network loss, power loss, e-stop	≥90% autonomous recovery, back in production within 2 min
Changeover	Hours, demonstrations, operator skill and data-collection cost per new part number or fixture	Structured stations: ≤1 shift and ≤100 demos per new part number

Item	Require	Suggested pass threshold
Safety	Arm and cell to ISO 10218-1/-2:2025, cell risk assessment, local rules (GB 11291 on the mainland; CNS 14490 plus written assessment in Taiwan; ANSI/A3 R15.06-2025 and UL 1740 in the US)	No safety function implemented by machine learning; a certified safety layer (typically PL d / Cat 3) limits speed and force; no fenceless legged robots until ISO 25785-1 is published and assessed
Compute and offline behaviour	Inference hardware, P99 latency, behaviour without network, model versioning, revalidation after every update	Takt-critical inference does not depend on the cloud; full re-acceptance after any model, quantization or engine change
Evidence	Unedited video and raw logs for the whole acceptance window; witnessed by the customer's quality team or a third party	No highlight reels or vendor-only summaries
Commercial terms	Payment milestones tied to acceptance metrics; availability penalties when billed per robot-hour; data ownership and remote-access rules	Final payment and retention released only after the 3,000-cycle acceptance and ≥4 weeks of stable running
Economics	Net people replaced = speed ratio × uptime × success − intervention time − teleoperation share	Reject any proposal whose net replacement cannot support its claimed payback

Sources: (Figure, Neuromeka, Yicai Global, i-SCOOP); thresholds are my suggestions.

OPEN QUESTIONS

07 What the public record cannot answer, and how a pilot settles it

The questions public sources cannot answer fall into five groups: what AI review actually sells for and what buyers will pay; how AOI makers bundle AI and upgrade old machines; how old the installed AOIs are and how many already run AI; how many people and hours review really takes per line; and how pilots convert and how buyers pay. On top of those comes how the newest models, quantized builds and robot arms actually perform under a given plant's conditions. Each can be turned into a test with a clear threshold. The thresholds below are decision rules I derived from the evidence in this report estimate ; recalculate them once the first plant data comes in.

Open question	How to settle it	What a good result looks like
Price and willingness to pay for AI review	A standard request for quotation (for example four lines with mixed older 2D and newer 3D AOIs, two shifts, about 2,000 boards per shift, 12 defect classes, on-premises only, MES or IPC-CFX integration) sent to several third-party vendors; a four-question price-sensitivity survey of 20–30 EMS plants	Most quotes within RMB 40–120k one-off or 20–40k a year for four lines; at least half of plants accept RMB 10–30k per line with payback within 12 months
AOI makers' AI bundling and upgrades for older machines	Short structured interviews with about a dozen AOI makers and distributors	Whether AI review is standard, optional or absent on 2025–26 models; whether pre-2020 machines can be upgraded and at what price
Age of installed AOIs and share running AI	On-site survey of about 30 EMS plants, stratified by size and region, with equipment lists reviewed under NDA	Median AOI age ≥5 years and ≥60% without AI means the retrofit pool is real
Review labour and hours	Interviews and 2–4 weeks of AOI alarm, review, repair and shift logs at two or more plants, with work sampling	≥1 dedicated reviewer per line per shift and ≥3 review hours per line per shift, with at least 0.5 FTE per line the plant is willing to redeploy
Performance of AI review itself	Golden set plus shadow mode on a real line (Chapter 3): data agreement in week 0, baseline and labelling in weeks 1–2, shadow running in weeks 2–4, optional assisted running in weeks 4–8	False calls or review hours cut ≥50–60%; per-class escape bounds no worse than the human baseline; zero misses on critical classes; p95 latency ≤½ takt; automatic fail-open to people
Teacher-label quality of frontier models	Run two or three candidates on the same golden set, three prompt wordings each, one good reference image, single-component crops ≥42–48 px	Automatic acceptance only for classes with zero misses on ≥299 defects and human-level agreement; otherwise pre-labelling and triage only
Quantized builds	Change only the precision; pin weight hashes, template, preprocessing, decoding, engine and batch size; disable or reset prefix caching; check the reference build reproduces vendor results; score pairs on the golden set and drift slice	No new misses on critical classes; ≤1 new miss in 1,000 defects; clearance down ≤2 points; ≥99.5% identical decisions over three runs; otherwise fall back to FP8/INT8
Line-side hardware, cleanliness and ESD	≥2 units of each candidate box stress-tested at twice the expected alarm rate; 72-hour soak at real cabinet airflow; ESD check by the customer's coordinator; ISO 14644-14 test for ISO 7 and cleaner customers	p95 ≤½ takt; no crashes or throttling over 72 hours; ESD sign-off; particle report at the right class
Payment terms, pilot rules and conversion	Payment clauses from recent inspection-equipment contracts at the surveyed plants; interviews on pilot rules, acceptance criteria and qualification time	About 30/60/10 with cash within 60 days of acceptance; pilots ≤8–12 weeks with a named sign-off owner; ≥1 in 3 trials converting; a framework covering ≥5–10 lines in year one
Data, IT and liability	Interviews and contract review with legal counsel	A local box may connect and access data; liability limited to software defects

Open question	How to settle it	What a good result looks like
Regional compliance before going abroad	Taiwan: customer's written status under the Cyber Security Management Act, E187 self-assessment; US: customs ruling on how a cell is classified, NRTL quotes, ITAR check; Southeast Asia: BOI local-linkage and Malaysian allowance eligibility, customer's parent; mainland: CCC scope for bundled PCs and power supplies	The customer is not a controlled body, and payback stays ≤2 years with duty and certification included
Acceptance of a robot or embodied cell	The 12-item acceptance table in Chapter 6: signed definitions, ≥3,000 consecutive production cycles over ≥3 shifts and ≥2 part numbers, scripted fault injection, ≥30 timed human cycles, witnessed	Lower bound ≥99.5% inline (≥99.0% offline); cycle ≤1.15 × human; ≤0.5 interventions per hour and zero teleoperation; ≥95% uptime; MTTR ≤20 min; ≥90% recovery
ACT and VLA policies on Chinese edge chips	Four weeks: port and micro-benchmark (≥1,000 inferences, latency distribution, power, 60-minute soak); closed-loop simulation with measured latency injected; real-robot A/B against a GPU; 8–24 hour endurance (≥3,000 cycles)	ACT chunks p95 ≤100 ms, VLA ≤300 ms; real-robot success within 3 points of the GPU and cycle ≤1.2 × human; otherwise keep the policy on Jetson or a GPU
Cobot, AMR and integration prices	Model-by-model quotes from several cobot distributors, and full-cell quotes from two or three integrators for test-fixture loading (one station, two shifts)	Full cell ≤RMB 200k and payback ≤1.5 years over two shifts in Shenzhen
SMT line counts in Taiwan, Southeast Asia and the US	IPC (formerly New Venture Research) data or local equipment distributors' installed base	Enough serviceable lines per region to justify local application engineering

Of all these, four decide the most sensitive numbers in this report: the price buyers will pay, the labour AI review actually frees, whether AI review really misses no more than people do, and how fast cash comes back. They are worth settling before any significant commitment, by a plant considering a purchase or by a supplier considering a product.

A pilot sequence that works for both sides

For a plant, and for a vendor that wants a credible reference, the order is the same:

1. **Weeks 0–2: agree on data and definitions.** Sign the data agreement, define the defect classes and what counts as an escape, and record a baseline: false calls per board, review seconds per call, escapes found downstream.
2. **Weeks 1–4: build the golden set** from the AOI archive and retained NG boards, labelled blind and adjudicated (Chapter 3).
3. **Weeks 2–4 (or longer): shadow mode.** The model logs its decisions while people decide as before; compare per class.
4. **Weeks 4–8: assisted running on random shifts,** with the model clearing only classes that passed the gate and everything else going to people.
5. **Acceptance, in writing, before scaling:** the per-class results, latency, fail-open behaviour, and who owns models and labels. Scale line by line, and revalidate on every new product, model version, quantization or engine change.

Limitations of this report

This is desk research. I conducted no primary interviews with plants for this edition; much line data comes from vendors and has not been independently audited. Several government websites, tender databases and recruitment platforms could not be accessed, so Shenzhen's 2026 upgrade subsidy rates, Yangtze River Delta wage guidance, transaction prices for AI review, reviewers per line and pilot conversion rates are estimates or analogies; each appears above as an open question. Exchange rates are People's Bank of China central parity on November 9, 2026 ($1 = RMB 6.7188) and, for Taiwan, SMBC's fixing on the same day ($1 = NT$32.12); the Vietnamese dong uses May 2026 data. All thresholds are decision rules derived from the cited evidence, not standards.



ABOUT THE AUTHOR

Tan Shuai

Tan Shuai is a technologist and founder with more than twenty years of work across chips, hardware and system software, and 20 granted patents. His work has taken him through Silicon Valley, New York, Chicago, Taiwan, the Greater Bay Area (Shenzhen and Hong Kong), Vietnam and Malaysia, with a long focus on advanced manufacturing and supply chains. Today he works as an AI researcher and developer, concentrating on how AI is actually used on the factory floor.

He welcomes conversations with people working on AI in manufacturing, whether building a team, looking for a partner or comparing notes: i@tanshuai.com · tanshuai.com

Notes and disclaimer

This report is independent desk research based on public documents available up to November 11, 2026. It is provided for information only and is not investment, legal, tax or engineering advice. Product names, prices, model releases, export classifications, tariffs and regulations change quickly; verify anything you rely on against the primary source. Figures marked vendor claim or estimate should be read as such. Company and product names belong to their owners and are used for identification only; no company reviewed or sponsored this report.

© 2026 Tan Shuai. You may share this report unchanged with attribution.

ELECTRONICS MANUFACTURING · AI INSPECTION × ROBOTICS

AI Visual Inspection and Robot Arms in Electronics Factories: 2026 Assessment

Every key number carries its source and a confidence tag: verified, single source, vendor claim or estimate. What the public record cannot answer is collected in Chapter 7, with how a pilot settles each question.

Tan Shuai

tanshuai.com • i@tanshuai.com

PUBLIC EDITION · DATA AS OF NOVEMBER 11, 2026

