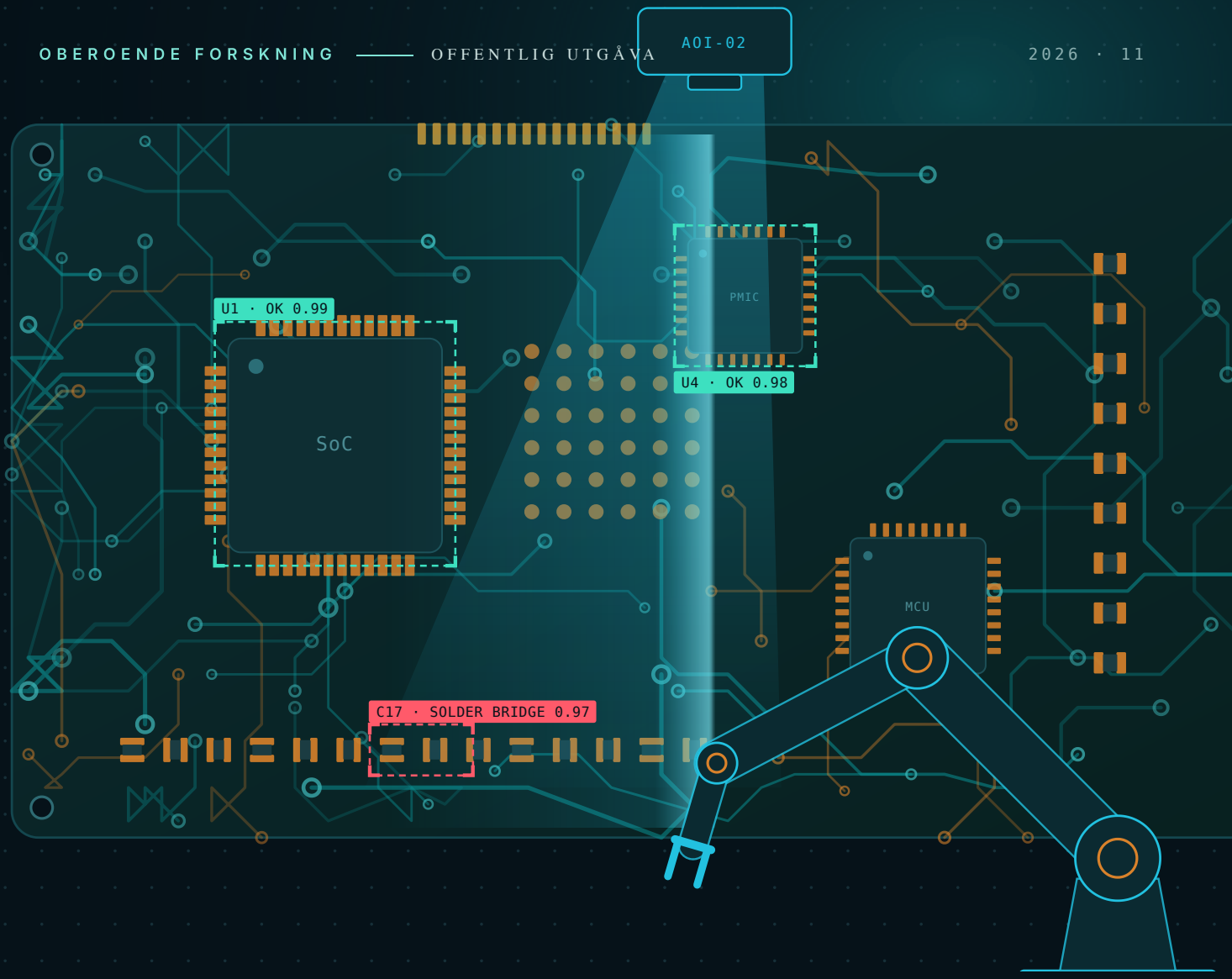


ELEKTRONIKTILLVERKNING · AI-INSPEKTION × ROBOTIK

AI-visuell inspektion och robotarmar i elektronikfabriker

2026 års bedömning

Vanliga konfigurationer · Regional genomförbarhet · Lokal installation · Lärar–elev-modeller
Avkastning på investering · Konkurrens · Vad endast ett pilotprojekt kan besvara

HUVUDFYND

AI-granskning är värd att göra, i begränsad omfattning. Förkroppsligade robotarmar kan vänta.

Tan Shuai
tanshuai.comi@tanshuai.com

Data per den 11 november 2026
Fastlandskina · Taiwan · Sydostasien · USA

INNEHÅLL

Innehåll



◆ **SAMMANFATTNING** — 5
Värt att göra, men smalt, nu och med partners

01 **Linjen idag** — 9
Vad elektronikfabriker kör idag, och vad som är tekniskt genomförbart

02 **Lokal installation** — 17
Att köra AI inne i fabriken: renrum, luftgap, hårdvara och kvantisering

03 **Lärare och elev** — 26
Lärare och elev: frontiermodeller kan förmärka inspektionsdata; för robotarmar är de verkliga lärarna fortfarande människor

04 **Kostnader och återbetalningstid** — 35
Kostnader och återbetalningstid: AI-granskning betalar sig på månader, maskinbetjäning på år, utplockningsarmar aldrig

05 **Marknad och konkurrens** — 43
Marknad, konkurrens och hur köpare betalar

06 **Förkroppsligad AI** — 52
Förkroppsligad AI: först ändrar den hur robotar lärs upp, senare hur de rör sig

07 **Öppna frågor** — 58
Vad det offentliga källäget inte kan besvara, och hur ett pilotprojekt avgör det

◆ **Om författaren** — 61
Tan Shuai

Tillförlitlighetsmarkörer i texten

- **verified** två eller fler oberoende källor, eller två oberoende primärdokument, är överens
- **single source** endast en källa, inklusive primärdokument
- **vendor claim** publicerat av säljaren av produkten eller tjänsten
- **estimate** författarens egen beräkning eller slutledning från citerade data

Så här läser du den här rapporten

Nyckelsiffror bär en tillförlitlighetsmarkör. **verified** betyder att två eller fler oberoende källor är överens, eller att två oberoende primärdokument är överens; **single source** betyder att det bara finns en källa (inklusive primärdokument); **vendor claim** betyder att det är publicerat av säljaren av produkten eller tjänsten, även om det upprepats brett; **estimate** betyder min egen beräkning eller slutledning från citerade data, med indata angivna. "Primär" beskriver källans typ (årsredovisning, prospekt, regelverk, standard, officiell statistik); det ändrar inte tillförlitlighetsklassen. Där källor strider anges siffrorna sida vid sida och medelvärdesbildas inte. Valutakonverteringar använder People's Bank of Chinas centralparitet den 9 november 2026, $1 = RMB 6.7188 (Xinhua), och för Taiwan SMBC:s fixing samma dag, $1 = NT$32.12 (SMBC). Fastlandets arbets- och subventionsdata använder Pärlflodsdeltat, där elektroniktillverkningen är som mest koncentrerad (officiella lönedata från Shenzhen), med Yangtzeflodsdeltat som jämförelse. Påståenden från branschintervjuer behandlas som påståenden att kontrollera, aldrig som fakta. Frågor som det offentliga källäget inte kan besvara är samlade i kapitel 7.

FIGURER

Figurer



01	Var AI passar på fabriksgolvet idag	7
02	En volym-SMT-linje: människorna sitter bakom AOI:n	10
03	Omkring 97–99% av AOI-larmen är falskutrop	11
04	En generell kamera är ingen AOI	13
05	En air-gapped utrullning	17
06	Ren- och ESD-zoner avgör var beräkning får stå	18
07	Hastighetsgränser per användning	20
08	On-prem beräkning: minne, pris och vad som lagligen går att köpa i fastlands-Kina	21
09	En stor arbetsstation gör ingen delad assistent	22
10	Kvantisering: 8-bit är säkert, 4-bit beror på receptet, 3-bit och lägre åker ut	23
11	Lärarkandidater: generiska poäng fördubblades på en generation	27
12	"Finns det en defekt?" är fortfarande svårt för generella modeller	28
13	Två lärare–elev-pipelines	29
14	Hur stor den gyllene uppsättningen måste vara	30
15	Edge-policylatens	32
16	Robotelever: nära produktionsgrad bara när stationen är strukturerad	33
17	Vad en inspektions- eller hanteringsarbetare kostar per år	36
18	Prisstegen	37
19	Återbetalning per region	39
20	Varför utplockning av dåliga kort inte betalar sig	40
21	USA-tull på robotar, AOI och edge-datorer	41
22	Marknadsstorlekar skiljer sig mer än tiofaldigt	44
23	Bottom-up: 40 000 linjer → 16 000 servicebara → ≈RMB 320M engångs	45
24	Retrofitfönstret är cirka fem år	46
25	Granskningskilen är omgärdad på tre sidor	46
26	Marginalen sitter i produktiserad hårdvara	48
27	Hur köpare betalar	49
28	Days sales outstanding	50
29	Embodied-AI-stacken	52
30	Vad robotutrullningar i fabrik faktiskt offentliggör	53
31	Netto ersatt arbete per robot	54
32	Tidslinje för standarder och reglering	56

SAMMANFATTNING

Värt att göra, men smalt, nu och med partners

HUVUDFYND

Domen: för elektronikfabriker är det som är moget 2026 och som fabriker betalar för AI-granskning och nischinspektion på AOI-bilder, plus programmerad robot-maskinbetjäning, inte "en stor modell som tittar på kort plus förkroppsligade robotar som plockar dem".

0.46–0.65 år

Återbetalningstid för AI-granskning i USA, Taiwan och Shenzhen, vid RMB 50k per linje

≥95% · <20 PPM

Täckning och läckage i ett ledande EMS-företags interna AI-granskning: ribban för att sälja till tier-1-EMS

100–700 miljoner RMB

Adresserbar tredjeparts-eftermonteringsmarknad i Fastlandskina, engångs, över ett fönster på ungefär fem år

0 poäng

Publicerade industridefekt-benchmarkpoäng för november 2026 års flaggskepp: lärare, inte inspektörer

1.1–1.5 år

Återbetalningstid för en cobot-cell för maskinbetjäning över två skift i USA, Taiwan och Shenzhen

0.003–0.14 FTE

Arbete som en utplockningsrobotarm sparar per skift på en SMT-linje: den betalar sig aldrig tillbaka

Inspektion
ROI

Övervakad djupinlärning som tillämpas på de bilder en befintlig AOI redan producerar sänker falsklarm eller granskningstimmar med 60–90% (vendor claims). DBG, ett stort EMS-företag, rapporterar i sin årsredovisning att dess interna system täcker mer än 95% av flaggade komponenter med ett läckage under 20 PPM och mer än 85% mindre manuell granskning; det är ribban en tredjepartsprodukt måste klara för att sälja till tier-1-EMS. Vid RMB 50,000 per linje betalar AI-granskning sig inom ett år i Fastlandskina, Taiwan och USA (Shenzhen 0.65 år, Taiwan 0.53, USA 0.46), cobot-celler för maskinbetjäning på 1.1–1.5 år och Sydostasien på 2–6 år, medan en robotarm som läggs till en SMT-linje enbart för att plocka ut dåliga kort aldrig betalar sig någonstans, eftersom plockning vid över 90% förstapassutbyte bara tar 0.003–0.14 av en person per skift estimate.

Modeller
ROBOTAR

Multimodala stora modeller kan fortfarande inte vara den primära inspektören. Per den 11 november 2026 har inget av flaggskeppen (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 och andra) ett publicerat resultat på något industridefekt- eller PCB-benchmark, och i ACM MM 2026 PCBA-utmaningen fick de fem bästa systemen, tränade på 6,600 riktiga linjebilder, bara 77–82 på "finns det en defekt". Frontiermodeller kan förmärka och triagera. För robotarmar är de verkliga lärarna fortfarande mänskliga demonstrationer, förstärkningsinlärning på riktiga robotar och simulering: elever når 90–99.4% vid fixerade, strukturerade stationer, men bara 30–80% på rörliga transportband, osedda delar och nya produkter; en frontier-agent klarade 104 av 108 kvasistatiska uppgifter på riktiga robotar, på 280 sekunder vardera.

Lokal installation
DRIFTSÄTTNING

Luftgapsisolerad installation i rena, kontrollerade fabriker är fullt praktisk. En vanlig SMT-hall är oftast ett ESD-skyddat område snarare än ett klassat renrum: en fläktlös Jetson- eller inhemsk NPU-box står vid linjen och GPU-servrar står i serverrummet. Enligt NVIDIA:s egna exportklassificeringar kan Jetson Thor och DGX Spark säljas till en vanlig fastlandsfabrik utan USA-licens; DGX Station kan inte. Precisionen är 8-bit som standard; leverantörs-kvantiseringsmedveten 4-bit är användbar med villkor; 3-bit och lägre är uteslutet.

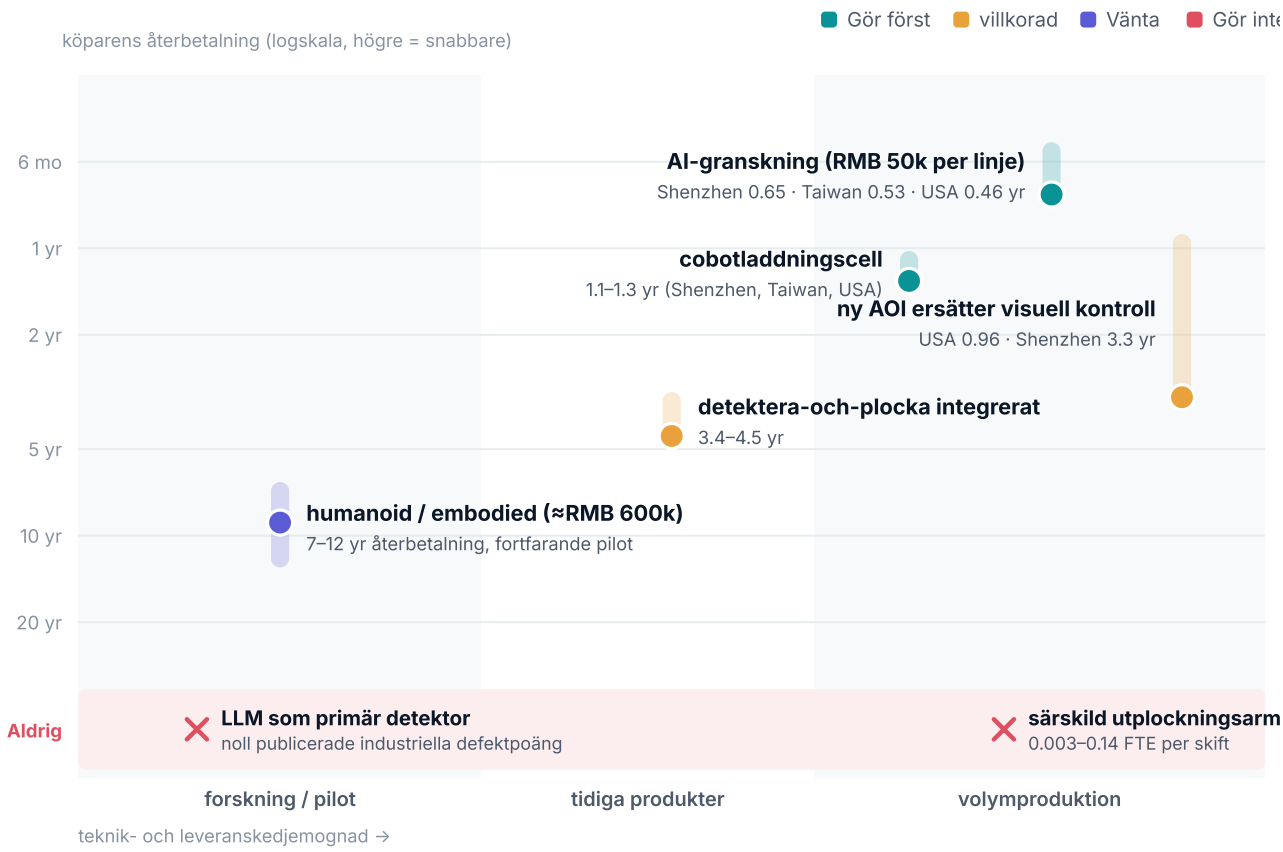
Marknad
KONKURRENS

Marknaden är reell men begränsad. Fastlandskina har cirka 35,000–46,000 SMT-linjer och cirka 26,000–66,000 inline-AOI:er som inte kör AI; den adresserbara marknaden för tredjeparts-eftermontering är cirka RMB 100–700 miljoner engångs, över ett fönster på ungefär fem år, och AOI-tillverkare har börjat bygga in granskning i nya maskiner estimate. Konkurrensen är redan stark: granskningskilen är omringad av AOI-tillverkare, interna EMS-team och plattformsleverantörer, och marginalen ligger i produktifierad hårdvara.

Fråga	Bedömning	Nyckeltal (källor och tillförlitlighet i kapitlen)
1 **Vad kör vanliga elektronikfabriker idag?**	En volym-SMT-linje är "3D SPI + AOI före och efter reflow + stickprovsbaserad röntgen + ICT/FCT + reparation och visuell inspektion", med mänsklig granskning bakom varje AOI. AI kommer in i granskningen på tre sätt: inbyggt av AOI-tillverkare, inbyggt internt av ledande EMS, eller tillagt av tredje part. Montering och hantering använder armar, SCARA och AMR:er med inlärd perception och programmerad rörelse; förkroppsligad AI är fortfarande på pilotstadiet	Takt ≈13–22 s per kort; en ODM budgeterade RMB 30.6M AOI för 18 linjer, ≈3–5 AOI:er per linje; ≈97–99% av AOI-larmen är falsklarm
2 **Hur långt når tekniken, kan pengar köpa noll personer, och vad kan varje region använda?**	Inspektion är industrikvalitet som "dedikerad detektor + AI-granskning + några få personer"; även obegränsade pengar når inte noll personer. Fastlandet använder bara inhemska modeller eller lokala öppna vikter; privata fabriker i Taiwan kan använda nästan vad som helst, men fabbar och kritisk infrastruktur utesluter fastlandsmärken i praktiken; Sydostasien har minst restriktioner; USA har hela stacken plus tullar och certifiering	AI-granskning sänker falsklarm eller granskningstimmar med 60–90% vendor claims; en 4K-kamera över en hel panel ≈78 µm per pixel mot 8–20 µm för en AOI; fastlandstillverkade armar betalar ≈40% kombinerad USA-tull, AOI-liknande instrument ≈37.5%
3 **Kan det köras lokalt i renrum och luftgapsisolerade fabriker, på vilken hårdvara, snabbt och exakt nog?**	Ja; begränsningarna är processuella, inte beräkningsmässiga. En fläktlös box ≤100 W vid linjen, GPU-servrar i serverrummet; Thor för stationer, en Mac Studio eller inhemsk apparat för fabriks-Q&A ingen DGX Station på fastlandet. Hastigheten är bra när modellstorleken matchar uppgiften; håll 8-bit, använd leverantörs-kvantiseringsmedveten 4-bit med villkor, aldrig 3-bit	≈0.5 s per utklipp för en 8B-seendemodell på Thor estimate; Thor-utvecklingskit RMB 59,104 i Kina; DGX Station $123–140k, 139 tokens/s totalt för 32 användare på en 743B-modell
4 **Kan det nyaste flaggskeppet agera lärare över dussintals rundor och destilleras till lokala elever, även för flexibel tillverkning och robotarmar?**	Inspektion: förmärkning och triage; autoacceptera lärarens etiketter endast för klasser med "≥299 defekter, noll missar" på ett guldset. Robotar: lärarna är fortfarande mänskliga demonstrationer, RL på riktiga robotar och simulering, och elever närmar sig produktionsklass endast vid strukturerade stationer	Inget novemberflaggskepp har ett industridefektresultat; elever 97.8% på tränade kontakter, 29–67% på osedda, 37–59% på rörliga transportband; frontiermodeller som driver leder direkt 0–5.5%
5 **Vad kostar det, och vad är ROI per region?**	AI-granskning betalar sig inom ett år i Fastlandskina, Taiwan och USA; cobot-betjäning 1.1–1.5 år; Sydostasien 2–6 år; en utplockningsarm på en SMT-linje betalar sig aldrig	RMB 50k granskningsbox per linje: Shenzhen 0.65 år, Taiwan 0.53, USA 0.46, Vietnam 2.2 estimate; acceptabelt tredjepartstilläggspris ≈RMB 10–30k per linje engångs eller RMB 5–10k per år

Fråga	Bedömning	Nyckeltal (källor och tillförlitlighet i kapitlen)
⑥ **Är konkurrenterna redan starka nog, och är det bättre att börja nu eller vänta?**	Efterfrågan är inte mättad, men utbudet är trångt; OEM-AI begränsar eftermonteringsfönstret till ungefär fem år. Gå in nu med en smal produkt genom kanalpartners, utan att strida direkt mot plattformsleverantörer; vänta på produktionsdata från 2027–2028 innan du satsar på förkroppsligad AI	≈26–66k inline-AOI:er utan AI på fastlandet, krympande med 8–12% per år estimate ; den ledande AI-seende-leverantören hade 2025-intäkter på RMB 1.086B och en nettoförlust på RMB 991M; det närmaste börsnoterade jämförelsebolagets 3C-intäkter föll 79.9% år över år i januari–maj 2026

FIG 01 **Var AI passar på fabriksgolvet idag** ***AI-granskning av AOI-falskutrop och cobotmatning är mogna och betalar sig snabbt; en särskild utplockningsarm och en LLM som primär inspektör är inte värda besväret.***



Källa: Återbetalning från författarens modell (kapitel 4); mognadsaxeln är författarens bedömning.

Vad man ska bygga och vad man inte ska bygga

	Rekommendation
Bygg	AI-granskning och defektklassificering på AOI-bilder, för medelstora EMS med äldre AOI eller blandade märken och ingen AI-uppgradering från AOI-tillverkaren; inspektion som SMT-AOI:n täcker dåligt (hålmonterat och handlödning, konform beläggning, kablar, kontakter och FPC:er, slutligt utseende och skruvar); cobot-celler för lastning och lossning där varje del hanteras (testfixturer, tråg, förpackning), med inspektion som tillägg; ett guldset och skuggläge på en riktig linje som grund för varje påstående
Bygg inte	En generell "AI plus robotarm"-plattform; en multimodal LLM som primär inspektör; en robotarm på en SMT-linje enbart för att plocka ut NG-kort (en NG-avledare eller buffert gör det); humanoider och robots-as-a-service; frontiermodeller i en styrkrets eller säkerhetsfunktion; högar av GPU-minne vid stationer för att köra "fullstora" modeller; direktförsäljning till fabbar, taiwanesisk kritisk infrastruktur eller USA:s försvarselektronikfabriker

	Rekommendation
Vem man ska samarbeta med	AOI-tillverkare och distributörer (kanal- och bildgränssnitt; distributörer stod för 46.68% av Sinic-Teks intäkter 2025); cobot-tillverkare (certifierade armar och säkerhetsfunktioner); EMS-företag med 10–40 linjer (rullande expansion och förskott); lokala integratörer i Thailand och på andra håll (för att uppfylla BOI:s regler om lokala kopplingar)

Linjen idag

01 Vad elektronikfabriker kör idag, och vad som är tekniskt genomförbart

Standardvolymlinjen: flera AOI:er, stickprovsbaserad röntgen, central granskning, 13–22 sekunder per kort

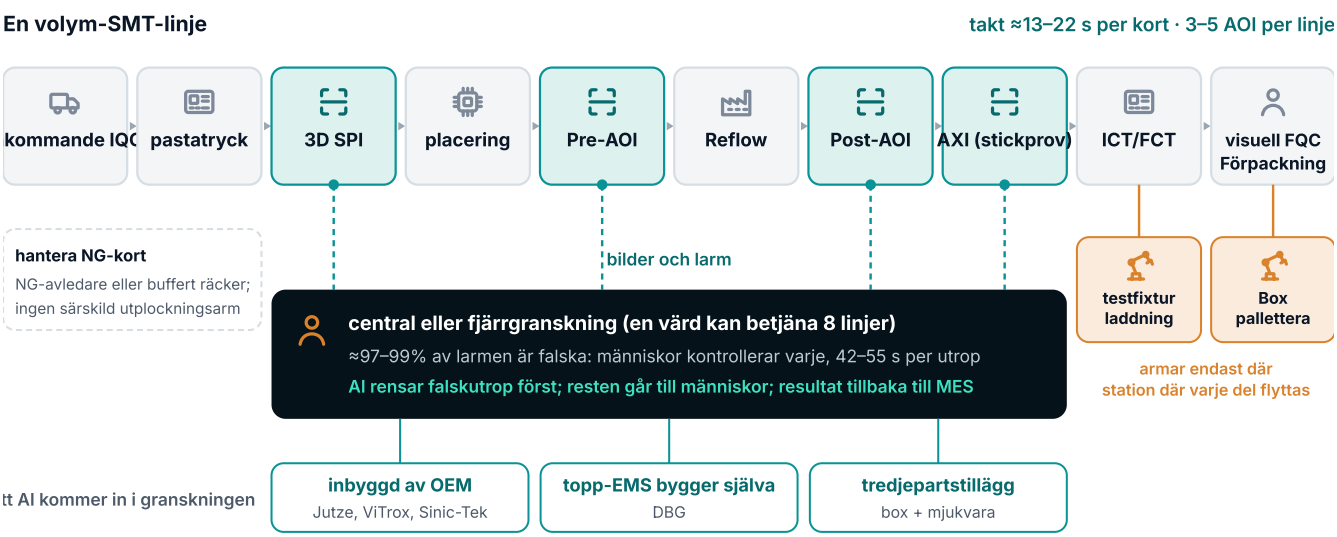
Primära inlagor från börsnoterade ODM:er visar hur en volymlinje för 2025–2026 faktiskt ser ut. Longcheers Hongkongprospekt (data till den 30 september 2025) listar sina SMT-kvalitetskontroller som "röntgeninspektion, lödpastainspektion (SPI), automatisk optisk inspektion (AOI), PCBA-funktionstest och visuell inspektion"; dess processflöde kör AOI före reflow, AOI efter reflow, stickprovsbaserad röntgen, NG-kort till en SMT-reparationsstation, och sedan en slutlig visuell kontroll. Dess fabriker i Huizhou, Nanchang och Vietnam har 34, 24 respektive 8 SMT-linjer och 67, 60 respektive 13 slutmonteringslinjer (Longcheer H-share prospekt) single source, primary filing. Huaqins bärbara fabrik från 2023 planerade 18 SMT-linjer med en AOI-budget på RMB 30.6 miljoner, cirka RMB 1.7 miljoner per linje, vilket till inhemska 3D-AOI-priser är tre till fem AOI:er per linje för inspektion före och efter reflow av dubbelsidiga kort (Huaqin A-share prospekt) primary filing + estimate. En volymlinje har därför oftast två till fyra granskningspunkter, inte en.

Maskinerna är redan höggradigt automatiserade: Huaqin rapporterar 96% automation på sina SMT-linjer (Huaqin H-share prospekt) single source. De återstående personerna finns inom granskning, reparation, visuell inspektion, test och slutmontering, och inget företag publicerar hur många granskare det har per linje (kapitel 7).

Tre oberoende uppskattningar är överens om en takt på **13–22 sekunder per kort** (en miljöinlaga från 2026 för fem Hebei-linjer (Hebei SMT-projekt EIA), Huaqins bärbara linjer och ett iFactory-fall), eller 7–11 sekunder per sida för dubbelsidiga kort estimate. Det svåra med en high-mix-linje är inte volymen utan introduktion av nya produkter och omställningar, och varje omställning är ännu en kallstart för en AI-modell.

På mjukvarusidan är AOI-tillverkarnas egna granskningsstationer redan centraliserade: Sinic-Teks reparationsstationsmjukvara låter "en värd styra utrustningen på åtta SMT-linjer" (Sinic-Tek prospekt) single source, primary, 2022, och efter att Jabil införde ViTrox V-ONE fjärrstyrd AOI-kontroll 2026 kunde en ingenjör hantera 5–8 AOI:er på distans mot 1–2 lokalt (AtlasPCB) vendor claim. En tredjepartsprodukt för AI-granskning kopplas in exakt i denna centralstation, genom AOI:ns bild- och resultatexport (IPC-CFX, IPC-HERMES-9852 eller en leverantörs eget format). AOI:erna själva får AI snabbt: Yamaha säger att dess AI-bedömning tar ungefär en sekund per anrop, hjälper inspektörer efter 30 NG-bilder och blir av "mycket hög" kvalitet efter 1,000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **En volym-SMT-linje: människorna sitter bakom AOI:n** ***Arbetet är koncentrerat till granskningsplatserna efter automatisk optisk inspektion. Det är där AI kommer in.***



Källa: Offentliga årsredovisningar och leverantörsmaterial (Huaqin, DBG, Sinic-Tek med flera); sammanställt av författaren.

Steg	Vanliga produkter och teknik	Representativa leverantörer	Hur långt AI har nått
Efter tryckning	3D SPI (strukturerat ljus, moiré, viss lasertriangulering)	Koh Young, Sinic-Tek, TRI	OEM-AI-pastabedömning i produktion; Jutze säger att vissa modeller är “helt AI”
Efter placering och reflow	2D/3D-AOI med flervinkliga, flerfärgade LED och strukturerat ljus, 8–20 µm per pixel, 8–12 MP-kameror	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek med flera	OEM-AI-granskning (Jutze i produktion, ViTrox V510Ai, Sinic-Tek med Tencent); ledande EMS bygger eget (DBG); tredjepartstillägg
Bottenterminerade och dolda fogar	Stickprovsbaserad eller inline-röntgen (AXI)	—	Hikvision hävdar att dess Guanlan-modell felbedömer röntgen-håligheter mindre än 1% av tiden vendor claim
Granskning och reparation	Central granskningsstation (en värd, åtta linjer), fjärrgranskning, reparationsstation	AOI-tillverkarnas mjukvara, ViTrox V-ONE	AI stänger falsklarm, människor tar resten
Visuell inspektion, IQC, FQC	Mestadels manuellt; vissa AI-utseendemaskiner med robotlastning	SmartMore med flera	Utseendemaskiner i produktion; ingen offentlig personaldata för IQC/FQC
Hantering och maskinbetjäning	Sexaxliga armar och cobotar, SCARA, AMR:er, 2D/3D-seende-vägledning	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Inlärd perception plus programmerad rörelse i produktion; VLA:er och humanoider fortfarande på pilotstadiet
Data och system	MES, IPC-CFX, IPC-HERMES-9852	—	Resultat skrivs tillbaka till MES; stödet i äldre utrustning är ojämnt

Källor till tabellen: inlagorna ovan samt (Koh Young, Jutze 2025 årsredovisning, Sinic-Tek 2025 årsredovisning, ViTrox, Sohu om Hengtianyi, Xinhua om Hikvision).

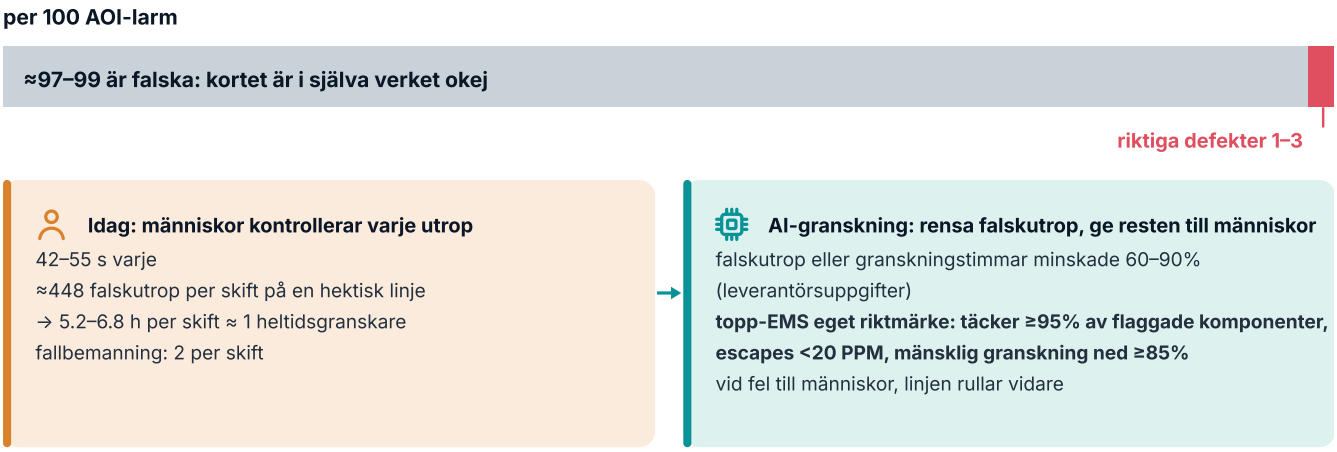
Människorna sitter bakom AOI:n, eftersom de flesta larm är falska

Inspektionskedjan på en vanlig SMT/PCBA-linje löper: 3D SPI efter tryckning för saknad, otillräcklig eller överflödig pasta, bryggor och förskjutning; AOI efter placering (valfritt) och efter reflow för saknade, förskjutna, roterade, vän-

da eller gravstensresta delar och lödmängd; röntgen när det finns BGA:er eller andra bottenterminerade delar; sedan in-circuit- eller flygprobs-test, funktionstest och manuell visuell inspektion. Varje AOI har en granskningsplats och en reparationsstation bakom sig (Sinic-Tek prospekt, JLCPCB) verified. AOI efter reflow är den vanligaste och mest värdefulla, men den ser bara form och area; fogar, håligheter och vätning under en BGA kräver röntgen eller elektriskt test (JLCPCB).

Arbetskraften sitter i granskningen eftersom de flesta larm är falska. En brittisk EMS noterade **54,019 falsklarm mot 544 verkliga defekter över 917 artikelnummer, cirka 99% av larmen** (arXiv 2505.03576) single source, preprint; vid Siemens Rastatt-fabrik var **97.5% av de kort AOI:n avvisade bra** (Siemens fallstudie) vendor claim. En hektisk linje ser cirka 448 falsklarm per skift på 42–55 sekunder vardera, 5.2–6.8 timmar per skift, ungefär en heltidsgranskare (iFactory) vendor claim; timmarna är en uppskattning.

FIG 03 **Omkring 97–99% av AOI-larmen är falskutrop*Idag bekräftar en person varje utrop; AI-granskning rensar först de falska och skickar resten till människor.***



Källa: Siemens fallstudie, iFactory (leverantörsuppgift), DBG:s årsredovisning 2025; granskningstimmar uppskattade av författaren.

Från "granskningstimmar per skift = kort × falsklarm per kort × sekunder per larm" kan en granskare täcka cirka 0.7–1 volymlinje med en regelbaserad AOI. När AI väl tar bort 60–90% av falsklarmen skulle enbart bedömningstiden tillåta 2–14 linjer per person, men gång, hantering av NG-buffert och reparationsöverlämningar begränsar lokal täckning till två till fyra angränsande linjer; bara central eller fjärrgranskning närmar sig det teoretiska taket estimate; sekunder per larm kommer från en källa. Eftersom vanliga AOI:er skannar 60–80 cm² per sekund (Yunheng 2026 AOI-guide) tar en 20×15 cm-sida 4–5 sekunder; AI:n behöver bara bearbeta de få flaggade regionerna inom en takt, inte skanna om kortet.

Vad som redan fungerar i produktion: övervakad klassificering på AOI:ns egna bilder

Det bevisade mönstret är övervakad djupinlärningsklassificering på bilder från den AOI en fabrik redan äger. Fujitsus tillägg på ADLINK:s AOI i Taiwan sänkte falsklarmen med **mer än 90%** (Fujitsu-fall); OKI har använt AI sedan juli 2026 för att sänka manuell bekräftelse efter AOI med cirka **80%** i sin egen EMS-verksamhet (Circuits Assembly); Siemens rapporterar 60% mindre manuell analys, €160,287 sparade under det första året och en återbetalningstid på åtta månader (Siemens fallstudie) alla vendor claims. Två saker är gemensamma för dessa fall: nästan ingen publicerar en läckagegrad, och alla behåller människor på de återstående larmen. Det enda fall som anger en läckagesiffra (0.8% ner till 0.3%, mot ett avtalat 0.5%) är internt motsägelsefullt (iFactory). AI sänker granskningstimmar, inte själva granskningsjobbet.

AI-granskning har gått från tredjepartstillägg till OEM-funktioner och interna system, vilket ändrar vad en nykomling mäts mot. Jutzes årsredovisning för 2025 listar sin "3D-AOI-uppgradering" som i volymproduktion och säger att den "använder AI-algoritmer för att ta bort manuell granskning vid inspektionsänden" (Jutze 2025 årsredovisning);

Sinic-Tek har utvecklat ett "multimodalt AI-assisterat manuellt granskningssystem" sedan 2019 och utvecklar tillsammans med Tencent en topp-och-botten-3D-AOI, med mål om 99.5%+ defekttäckning och 98% igenkänningsnoggrannhet (Sinic-Tek 2025 årsredovisning) vendor claims in primary filings. Stora EMS-företag bygger eget. DBG:s årsredovisning för 2025 listar "AI-bedömning för visuell inspektion" som ett slutfört projekt som "täcker mer än 95% av AOI-flaggade komponenter", med en "läckagegrad under 20 PPM" och "mer än 85% mindre manuell granskning", fullt spårbart (DBG 2025 årsredovisning) single source, primary. Det är den enda elektronikfabrikssiffra jag fann som anger täckning, läckage och minskning tillsammans. **En tredjepartsprodukt som säljs till tier-1-EMS måste matcha DBG:s 95% täckning, 20 PPM läckage och 85% minskning**; under tier 1 är riktmärket den minskning på 50–85% som leverantörer annonserar.

Anomalidetektering som bara tränats på bra delar är snabb (EfficientAD-S tar 2.2 ms per bild (EfficientAD, WACV 2024)) men svag på "logiska" anomalier som saknade eller felaktiga delar, och de bästa metoderna ligger i genomsnitt under 60% AU-PRO på det svårare MVTec AD 2 (MVTec AD 2) single source. Saknade och felaktiga delar kontrolleras bättre mot CAD och BOM.

Taket till varje pris: lager av maskiner, mycket få människor, men inte noll

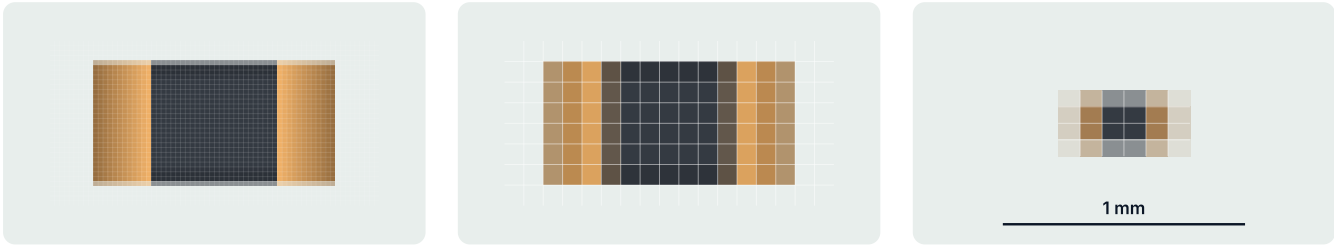
Jag fann ingen elektronikfabrik som helt tog bort manuell granskning och nådde noll läckage. Praxis för IPC-A-610 Klass 3-produkter (hög tillförlitlighet) låter fortfarande utbildade inspektörer avgöra de gränsfall som automatiska trösklar inte kan klassificera, och en PCB-montör kallar "automatiserad inspektion kan ersätta mänsklig sign-off" för en "vanlig missuppfattning" (PCBCart) single source. Den mänskliga baslinjen behöver också korrigeras: med hänvisning till en Sandia-studie från 2012 fångar en enskild visuell inspektör i bästa fall cirka **80%** av defekterna, två i serie cirka 96% (Instrumental) single source, citerad av en leverantör. Det försvarbara påståendet är "AI plus en liten mänsklig granskningsloop slår helt manuell inspektion", inte "AI ersätter inspektion".

Utan budgetbegränsning är taket 3D SPI, 3D-AOI, röntgen, AI-granskning och AI-utseendestationer staplade tillsammans, med människor kvar bara för gränsfall, nya produkter och omställningar, nya defekttyper och kundsign-off. Flaggskeppsfabriker rapporterar 85–92% mindre manuell inspektion (I-Connect007, 2021, WiAdvance) vendor claims; min uppskattning av taket utan budget är granskningsarbete på 8–15% av dagens estimate. Slutmontering är fortfarande arbetsintensiv: automatiseringen av iPhone 16-knapparna lades ner på grund av höga defektgrader (9to5Mac) single source.

En generell högupplöst kamera är inte en AOI

Produktions-AOI:er avbildar ruta för ruta vid 8–20 µm per pixel, med 8–12 MP-kameror, programmerbar flervinklig flerfärgad belysning och strukturerat ljus i 3D (Koh Young, Saki) verified. Branschen noterade för länge sedan att en 01005-del bara täcker cirka 8×16 pixlar vid 25 µm per pixel, "möjligen inte tillräckligt" för tillförlitlig bedömning (Circuits Assembly, 2007). Med samma beräkning estimate får en 4K-kamera över en 300 mm-panel cirka **78 µm per pixel**: en imperial 0402-del är cirka 13×6 pixlar, tillräckligt för närvaro och grov förskjutning men inte fogform, och en 01005 är cirka 5×3 pixlar, i praktiken osynlig. Att nå 15 µm per pixel över en hel panel skulle ta cirka 330 megapixlar, vilket är anledningen till att AOI:er flyttar en kamera över kortet. När någon säger "0402", kontrollera vilken de menar: metrisk 0402 är imperial 01005, långt bortom varje vidvinkelkamera.

FIG 04 **En generell kamera är ingen AOI** ***En 4K-kamera över en 300 mm-panel ger cirka 78 µm per pixel: en 0402-del krymper till ≈13×6 pixlar och en 01005 försvinner.***



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

4K-kamera över en 300 mm-panel: ≈78 µm/px
0402 krymper till ≈13×6 px: räcker för närvaro och grov förskjutning

samma kamera på en 01005
≈5×3 px: i praktiken osynlig

Källa: Circuits Assembly; författarens beräkning för en 300 mm-panel och en 4K-sensor. Pixelrutnäten är illustrativa.

På linjer som redan har AOI:er är den bästa platsen för AI ovanpå AOI:ns bilder. Bara linjer utan AOI, eller ändstationer, motiverar en inkapslad kamerastation med kontrollerad belysning; en öppen högupplöst kamera över linjen är mest utsatt för omgivande ljus och skuggor (CSDN).

Montering och hantering: inlärd perception plus programmerad rörelse är i produktion

IFR:s rapport för 2026 säger att "maskinseende för inspektion och robotvägledning är moget och av industrikvalitet", medan "allmän fysisk AI och inlärning från simulering och erfarenhet ännu inte är av industrikvalitet" (IFR World Robotics 2026 sammanfattning) single source, primary. Elektronik är den största robotköparen (142,235 enheter installerade världen över 2025, 59% av alla installationer i Kina (IFR, IFR Kina-release)), ändå har Kina bara 166 robotar per 10,000 tillverkningsarbetare mot 307 i USA (IFR, april 2026). Mönstret i produktion är inlärd perception plus deterministisk rörelse; änd-till-änd-inlärd rörelse är nära produktionsklass endast vid strukturerade stationer (kapitel 3 och 6).

Vad varje region kan använda

Fastlandskina förlitar sig på inhemska modeller och lokala vikter; i Taiwan avgör kundens identitet; i USA avgör tullar och certifiering.

Element	Fastlandskina	Taiwan (privata fabriker)	Sydostasien	USA
Claude (vilken kanal som helst)	Inte tillgänglig	Villkorat: leverantören får inte vara majoritetsägd eller kontrollerad av personer eller enheter i Fastlandskina, Hongkong eller Macao	Samma	Samma
GPT- och Gemini-moln	Inte tillgängliga	Tillgängliga	Tillgängliga (listor inte kontrollerade land för land)	Tillgängliga
Kinesiska molnmodeller (Qwen, DeepSeek, Doubao, GLM, Kimi)	Tillgängliga (registrerade modeller; att erbjuda en tjänst kräver registrering)	Lagligt, men fastlandshosting klarar sällan kundens säkerhetsgranskning; förbjudet inom offentlig sektor, statliga företag och kritisk infrastruktur	Lagligt (internationella noder)	Lagligt, strikt granskning

Element	Fastlandskina	Taiwan (privata fabriker)	Sydostasien	USA
Lokala öppna vikter av kinesiskt ursprung (Qwen, GLM, DeepSeek)	🟢 Förstahandsval; privat installation i fabriken kräver ingen modell- eller algoritmregistrering (enligt advokatbyråers läsning)	🟢 Tillgängliga (offentlig sektor begränsad)	🟢 Tillgängliga	🟡 Villkorat: ursprungsgranskning
Lokala öppna vikter av icke-kinesiskt ursprung (t.ex. Gemma 4)	🟢 Tillgängliga	🟢 Tillgängliga	🟢 Tillgängliga	🟢 Tillgängliga
Klassiska icke-generativa seendemodeller	🟢 Tillgängliga (utanför reglerna för generativ AI)	🟢 Tillgängliga	🟢 Tillgängliga	🟢 Tillgängliga
Jetson Thor/T5000/T4000, DGX Spark	🟢 Tillgängliga: NVIDIA klassificerar dem som 5A992.c, ingen USA-licens behövs för en vanlig civil fabrik; Spark har en Kina-SKU; kanalpriser cirka 40–60% över USA:s listpris	🟢 Tillgängliga	🟢 Tillgängliga	🟢 Tillgängliga
RTX 5090, RTX PRO 6000, DGX Station	🔴 Inte tillgängliga: de två första är 4A090.a, DGX Station 5A992.z; lagliga alternativ är 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	🟢 Tillgängliga	🟢 Tillgängliga, men köpare med huvudkontor i eller ytterst ägda från Kina räknas som kinesiska; Malaysia kräver tillstånd för export och omlastning	🟢 Tillgängliga
Mac Studio M5 Ultra 512 GB	🟢 Tillgänglig (Apples Kinabutik)	🟢 Tillgänglig	🟡 Förmodat tillgänglig	🟢 Tillgänglig
Huawei Ascend och andra kinesiska acceleratorer	🟢 Tillgängliga	🔴 Hög risk	🔴 Hög risk	🔴 Inte användbara i praktiken
Fastlandstillverkade fasta armar, AOI:er, industrikameror	🟢 Tillgängliga	🟡 Importerbara; sedan den 1 december 2025 kan myndigheterna agera direkt mot fastlandsmärken, och myndigheter, statliga företag, vissa stiftelser och leverantörer av kritisk infrastruktur kan begränsas eller förbjudas att använda dem	🟢 Tillgängliga	🟡 Lagligt, men kombinerad tull är cirka 40% på armar, 37.5% på AOI-liknande instrument och kameror, 25% på edge-datorer
Utlandstillverkade AMR:er och mobila manipulatorer	🟢 Tillgängliga	🟢 Tillgängliga	🟢 Tillgängliga	🟡 Nya modeller behöver villkorat godkännande från USA:s krigsdepartement
Hikvision- och Dahua-kameror	🟢 Tillgängliga	🟡 Privat bruk tillåtet, offentlig sektor förbjuden	🟢 Tillgängliga	🔴 Nya FCC-godkännanden blockerade; förbjudna för federala entreprenörer

Element	Fastlandskina	Taiwan (privata fabriker)	Sydostasien	USA
Robotsäkerhet och certifiering	● GB 11291.1-2011 och GB 11291.2-2013 obligatoriska, under revidering sedan april 2026; CR-certifiering frivillig; GB/T 39404-2026 täcker styrenheters cybersäkerhet	● Inhägnad som standard; inhägnadslöst samarbete måste uppfylla CNS 14490 eller ISO 10218 med en skriftlig bedömning	● Per kund och exportdestination, bekräftat skriftligt före leverans	● Elektrisk utrustning måste vara "godkänd" (i praktiken ett NRTL-märke eller fältbedömning); UL 1740 kräver att robotceller uppfyller ISO 10218-2 sedan den 11 augusti 2025
Fabbar och fabriker för kontrollerade produkter	—	● TSMC:s inköpsavtal kräver SEMI E187; fastlandsutrustning tas bort från 2 nm-fabbar	—	● ITAR-tekniska data inkluderar "fotografier"; avslagspolicy för Kina; fastlandskontrollerade leverantörer kan i praktiken inte betjäna försvarselektronikfabriker

Källor till tabellen: (Anthropic: regioner som stöds, OpenAI: länder som stöds, Gemini-tillgänglighet, CAC tillfälliga åtgärder för generativ AI, NVIDIA export-klassificeringsdata, eCFR 15 CFR 774, NVIDIA Kina DGX Spark, Apple Kina, Covington om nya BIS-regler, Greenberg Traurig, juni 2026, Fulcrum, Tom's Hardware om Ascend-vägledning, Taiwan BOFT importlista, Taiwan Cyber Security Management Act-historik, Taiwan granskningsregler för produkter som äventyrar nationell cybersäkerhet, Taipei Times, USITC HTS kapitel 99, Latham & Watkins, IPVM, Kinas nationella standardplattform, GB 11291.2, Intertek UL 1740-uppdatering, eCFR 22 CFR 126.1).

Fastlandskina: en komplett stack, men varken stängda frontiermodeller eller de starkaste enskilda maskinerna kan komma in. Registrerade inhemska molnmodeller, lokala öppna vikter (Qwen under Apache-2.0, GLM och DeepSeek under MIT), klassiska seendemodeller och inhemska AOI:er, kameror och cobotar är alla användbara. NVIDIA:s egna exportklassificeringar listar Jetson AGX Thor-kitet, T5000/T4000-modulerna och DGX Spark som 5A992.c, så en vanlig EMS- eller PCBA-fabrik behöver ingen USA-licens, medan RTX 5090 och RTX PRO 6000 (4A090.a) och DGX Station (5A992.z, TPP 58,484) inte kan köpas på fastlandet (NVIDIA exportklassificeringar, eCFR) primary; den juridiska läsningen är min och inte juridisk rådgivning. RS China listar Thor-kitet till RMB 59,104 inklusive skatt, cirka 1.4 gånger USA:s listpris (RS China).

Fyra advokatbyråer är överens om att en privat installation i fabriken för fabrikens eget bruk inte kräver någon modell- eller algoritmregistrering, medan en hostad tjänst för flera hyresgäster ligger i en gråzon (Covington, FPF, Jun-He) verified: enhetliga läsningar, inget officiellt utslag. GB 11291.1-2011 och GB 11291.2-2013 förblir den obligatoriska robotsäkerhetsbaslinjen och har varit under revidering sedan den 28 april 2026; GB/T 39404-2026 om styrenheters cybersäkerhet trädde i kraft den 1 november 2026 (revisionsplan 20262542, GB/T 39404-2026) single source, primary. Tills den nya GB 11291 anländer (2028–2029 estimate), utforma enligt ISO 10218:2025 och behåll GB 11291-kompatibilitetsdokumentation.

Taiwan: privata fabriker kan använda nästan vad som helst; kundens identitet avgör. Frontier-API:er, öppna vikter och fastlandsutrustning är alla lagliga (data.gov.tw). Men sedan den ändrade Cyber Security Management Act trädde i kraft den **1 december 2025** kan myndigheterna agera direkt mot produkter som är "fastlandsmärken" och begränsa dem hos statliga organ, statliga företag och leverantörer av kritisk infrastruktur (Laws & Regulations Database, granskningsregler) verified; listan över leverantörer av kritisk infrastruktur är konfidentiell; och TSMC har skrivit in SEMI E187 i sina inköpsavtal (CNA) och rapporteras ta bort fastlandsutrustning från sina 2 nm-fabbar (TrendForce) single source. Privata EMS-, monterings- och små PCBA-fabriker är öppna; fabbar, parkverk och myndighetsrelaterade kunder är i praktiken stängda. En fastlandskontrollerad leverantör bör utforma för E187, leverera en programvaruförteckning (SBOM), erbjuda offline-uppdateringar och aldrig fjärransluta från fastlandet.

Sydostasien: minst restriktioner; problemen är långsam återbetalningstid och kinesiska moderbolag. Allt är användbart, men USA:s vägledning av den 31 maj 2026 kräver licens för att exportera avancerade beräkningsartiklar till enheter med huvudkontor i, eller ytterst ägda från, Kina även när mottagaren finns i ett tredje land (Greenberg Traurig) single source, och Malaysia har egna tillstånd för avancerade chip (Fulcrum). Efterfrågan flyttar hit: TrendForce förväntade sig att Thailand, Vietnam och Malaysia närmar sig 50% av server-ODM:s SMT-kapacitet 2026 (TrendForce) single source, en prognos från 2023.

USA: hela stacken, till priset av tullar, certifiering och kontrollerade data. Enligt kapitel 99 i USA:s tullschema betalar en fastlandstillverkad robotarm cirka 40% i kombinerad tull, AOI-liknande instrument och kameror cirka 37.5% och edge-datorer 25%; tillverkade i Malaysia eller Indonesien cirka 10–12.5%, i Vietnam, Thailand eller Filippinerna 12.5–15%, i Taiwan 10%, med datorer på 0% utanför fastlandet (USITC HTS kapitel 99, Great Lakes Customs Law) primary schedule + estimate. USTR:s meddelande av den 10 november förlängde 178 gamla Section 301-undantag till den 9 januari 2027, varav inget täcker robotar, inspektionsinstrument, kameror eller datorer (Federal Register 2026-22925) verified.

OSHA kräver att elektrisk utrustning på arbetsplatsen är "godkänd" (i praktiken ett NRTL-märke, eller en fältbedömning på cirka $3,000 till $20,000+ per cell), och sedan den 11 augusti 2025 för UL 1740 in hela robotceller under ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). Försvarselektronik är i praktiken stängd för fastlandskontrollerade leverantörer: ITAR behandlar produktionsfotografier som tekniska data och tillämpar en avslagspolicy för Kina (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). En USA-bunden cell bör använda en fast arm, obegränsade kameror och lokala eller USA-molnmodeller, med slutmontering i Sydostasien där det verkligen ändrar ursprung.

Till varje budget skiljer sig taket per region men når ingenstans noll personer estimate. Fastlandet kan stapla inhemsk 3D SPI, 3D-AOI, röntgen, AI-granskning och ett internt AI-team för att föra ner granskningsarbetet till 8–15%, men utan stängda frontiermodeller eller DGX Station måste fabriksomfattande LLM-Q&A köras på lokala öppna vikter och inhemsk beräkningskraft (en fullstor modell för många användare behöver minst två Ascend-servrar med åtta kort). Taiwan och Sydostasien kan tekniskt använda det bästa av allt, inklusive frontier-API:er och DGX Station, begränsat av kundidentitet, fastlandsmärken och licensiering av kinesiska moderbolag. USA har inga tekniska gränser, till priset av tullar, NRTL-certifiering och personalseparation för kontrollerade data. Överallt finns människor kvar för gränsfall, nya produkter och omställningar, nya defekttyper och kundsign-off.

En produkt som säljs över regioner behöver därför två modellstackar: en fastlandsstack på registrerade inhemska modeller eller lokala öppna vikter, och en utlandsstack på frontier-API:er plus en lokal modell av icke-kinesiskt ursprung. Anthropics policy som träder i kraft den 12 november 2026 bedömer utifrån ägande och kontroll, så en fastlandskontrollerad leverantör kan inte bygga på Claude ens genom ett utländskt dotterbolag, en molnplattform eller en återförsäljare (Anthropic) single source, primary.

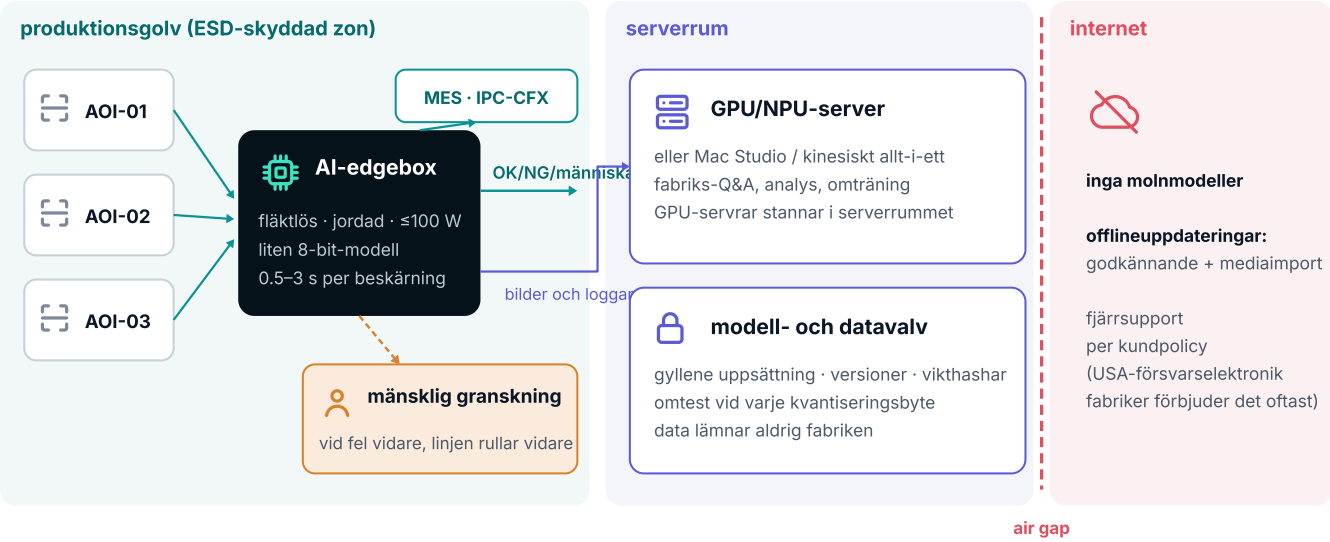
Lokal installation

02 Att köra AI inne i fabriken: renrum, luftgap, hårdvara och kvantisering

I en luftgapsisolerad fabrik är begränsningen process, inte beräkningskraft

Helt offline-ansluten visuell inspektion är rutin; begränsningarna är processuella. Kinas MIIT-guide från 2024 om säkerhet för industriella styrsystem efterfrågar zonindelning, brandväggar eller datadioder, granskat tidsbegränsat fjärrunderhåll, skannade medier och testade patchar (guidetext) single source ; halvledarkunder lägger till SEMI E187 och E188, som TSMC skrev in i sina inköpsavtal 2023 (SEMI, TXOne) verified . Resultat går tillbaka till MES över IPC-CFX och kortets identitet färdas över IPC-HERMES-9852 (Hermes-standarden). Den vanliga fällan är licensiering som förutsätter ett nätverk: Roboflows offline-läge cachar vikter på ett 30-dagars leasingavtal (Roboflow). En offline-produkt behöver offline-aktivering, signerade uppdateringar på skannade medier, en programvaruförteckning (SBOM), USB-kontroll, lokala granskningsloggar och en CFX-adapter.

FIG 05 **En air-gapped utrullning** ***En liten box vid linjen bestämmer; serverrummet svarar på frågor; vid fel faller man tillbaka på människor och data lämnar aldrig fabriken.***



Källa: Författarens syntes (kapitel 2).

Stängda frontiermodeller kommer i praktiken inte in i en fabriks serverrum. Gemini erbjuds luftgapsisolerat endast genom Google Distributed Cloud för myndigheter och stora företag (Google Cloud); Microsofts Foundry Local öppnar stora modeller bara för "kvalificerade kunder" och nämner ingen GPT-modell (Microsoft); Claude erbjuds bara i molnbaserade myndighetsregioner, utan lokal installation (Claude för myndigheter); ingen av de tre erbjuds i Fastlandskina. Fabriksutrustning måste utformas kring öppna vikter; frontiermodeller kan på sin höjd etikettera eller förklara bilder på huvudkontoret, offline, när kunden tillåter att anonymiserade bilder lämnar fabriken.

Serverrumshårdvara bör behandlas som kontorsklass: Mac Studio är klassad för 10–35 °C och DGX Spark för 5–30 °C, med rapporter om termiska avstängningar under ihållande inferens (Apple, NVIDIA-forum) verified . Linjen får fläktlösa industriella boxar, som en Jetson T4000 industri-PC klassad för −25 till 60 °C (AAEON) vendor claim . Eftersom en timmes driftstopp kostar en stor fabrik $36,000 till $2.3 miljoner (Siemens True Cost of Downtime 2024)

vendor survey måste varje AI-steg på linjen **öppna vid fel (fail open)**: vid varje fel faller det tillbaka till den befintliga AOI-plus-människor-processen.

Ren- och ESD-zoner: en vanlig SMT-hall är ett ESD-skyddat område, inte ett klassat renrum

Det som begränsar var beräkningskraften får stå är statisk elektricitet och luftflöde, inte renhetsklass. De flesta SMT-montageområden är "inte formellt partikelkontrollerade", och fabriker i världsklass "kan nå eller närma sig ISO 8" (Circuitnet expertpanel) single source. Medicinsk PCBA och kameramoduler ligger typiskt i ISO 7, nakna chip och optisk inriktning i lokal ISO 5–6 (PCBSync) single source; renrumsbyggare anger ISO 7–8 för SMT generellt (Angstrom Technology) vendor claim.

Det universella kravet är ESD-kontroll: ANSI/ESD S20.20-2021 kräver isolerade ledare under 35 V och håller stark-fältsisolatorer minst 30 cm från känsliga enheter (ANSI, ESDA, Desco) verified. Fabbar placerar beräkningskraften i sub-fabben eller datacentret, i vätskekylda rack som uppfyller SEMI S2/S8/F47 (Prodrive) vendor claim.

Huruvida utrustning får komma in i ett rent område testas för partikelemission enligt ISO 14644-14; robotar har sådana certifikat (Fraunhofer IPA), men jag fann ingen industri-PC, edge-AI-box eller GPU-server med ett. CE, FCC och IEC 60068 visar att en box tål sin miljö, inte att den får komma in i ett ESD-skyddat område; kundens ESD-samordnare kontrollerar det. GPU-servrar hamnar aldrig på linjen: en DGX Station drog 1,292 W vid GPU:n i tester (StorageReview).

FIG 06 **Ren- och ESD-zoner avgör var beräkning får stå** ***En vanlig SMT-hall är en ESD-skyddad zon, inte ett renrum. GPU- och NPU-servrar hör hemma i serverrummet.***



Källa: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, SEMI-standarder och leverantörsmaterial; sammanställt av författaren.

Zon	Typisk miljö	Beräkningskraft vid linjen	GPU/NPU-servrar
Vanlig SMT/PCBA-hall (konsument, industri)	ESD-skyddat område enligt S20.20 eller IEC 61340-5-1, oftast inte ett klassat renrum (≈ISO 8-ekvivalent); 40–60% RH	Fläktlös box i ett jordat metallhölje (Jetson T4000/Thor, kinesisk NPU-box) bredvid AOI:n eller granskningsplatsen; plastmonitorer och kapslingar utanför området eller ≥30 cm från kort	Serverrum eller luftkonditionerat IT-rack
Medicin, kameramoduler, EV-elektronik (ISO 7–8 renrum)	Renrum plus ESD-skydd	Fläktlös, IP65/66-förseglad, slät och lätt att torka av; begär ISO 14644-14-data eller ett IPA-certifikat	Utanför renrummet, nätverksansluten in

Zon	Typisk miljö	Beräkningskraft vid linjen	GPU/NPU-servrar
Nakna chip, wire bonding, optiska stationer (lokal ISO 5–6)	Mini-miljö	Endast kameror och ljus inuti; beräkningsboxen utanför	Serverrum
Fab-renrum (ISO 5 och renare)	Fab-specifika ESD- och elektrostatiska attraktionsplaner, utrustning enligt SEMI E78; cybersäkerhet enligt SEMI E187/E188	Endast verktygsintegrerade, fab-kvalificerade styrenheter; inga tilläggs-PC:ar med fläktar	Sub-fab-vätskekylda rack eller fabbens datacenter

Tabellen är min syntes av källorna ovan estimate. För en produkt innebär det en fläktlös linjesidig box under cirka 100 W, ett metallhölje kopplat till skyddsjord, 12–24 V DC-ingång, bredtemperaturdrift vid verkligt kapslingsluftflöde, och en ISO 14644-14-partikelemissionsrapport redo för kunder vid ISO 7 och renare. Varje utformning där "en GPU-server betjänar fyra linjer" placerar servern i serverrummet och lämnar bara en liten box eller en kamera vid linjen.

Hastighetsgränser per användning: noggrannhet är svårare än hastighet

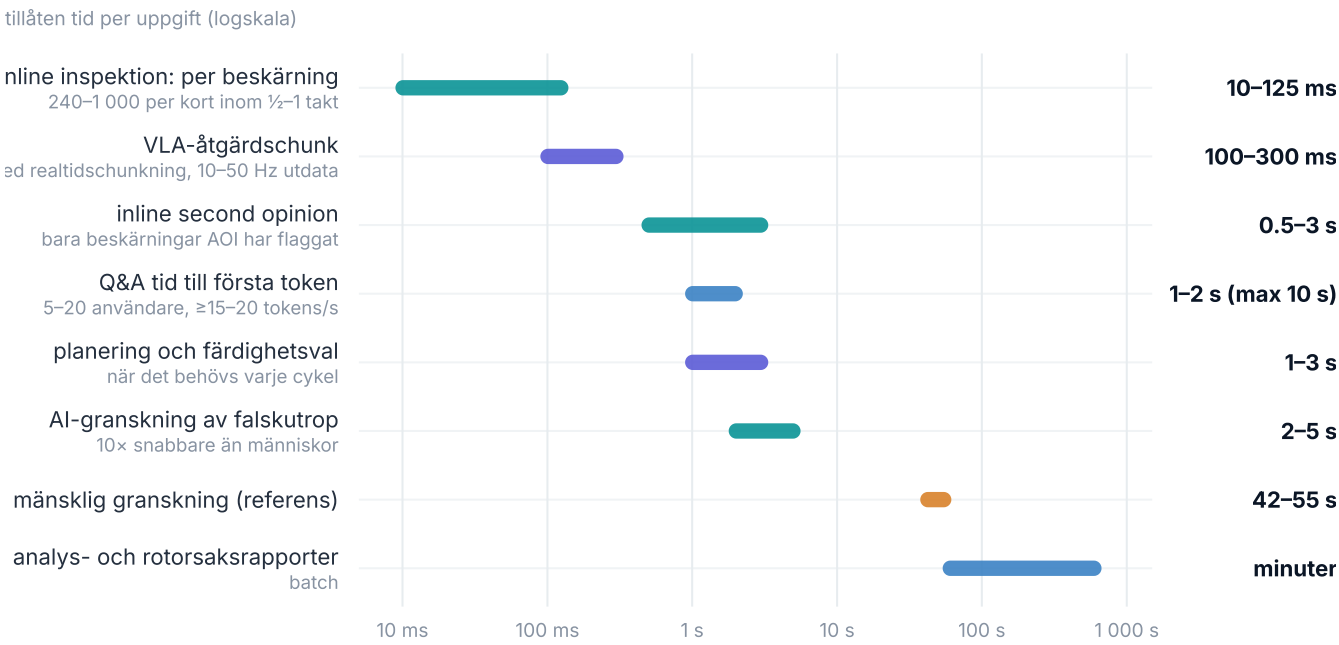
Fabriker och leverantörer publicerar nästan aldrig AI-latens- eller tillgänglighetsmål. Trösklarna nedan härleds från takt, mänsklig granskningstid, robotstyrningshastigheter och interaktionsforskning estimate:

Användning	Hastighetsgräns	Noggrannhetsgräns	Tillgänglighet
Inline-inspektion av varje del	Inom ½–1 takt: 240–1,000 utklipp i en 10–30 s-takt, 8–100 utklipp per sekund	Läckage inte sämre än dagens "AOI plus människor" (avtalsexempel <0.5%)	Öppna vid fel, stoppa aldrig linjen; mål ≥99.9% (ingen offentlig SLA)
Inline andra utlåtande enbart på AOI-flaggade utklipp	Flaggor (oftast ≤5–20 per kort) inom ½ takt: 0.5–3 s per utklipp	Samma; tränad på linjens egna bilder	Faller tillbaka till mänsklig granskning
AI-granskning av falsklarm	2–5 s per larm, mer än tio gånger snabbare än en persons 42–55 s	Läckage inte sämre än en mänsklig granskare	Försämring lägger till människor, stoppar aldrig linjen
Robotstyrning (VLA)	100–300 ms per åtgärdsblock med realtidsindelning, 10–50 Hz utdata	≥99% uppgiftsframgång för att slå programmerad rörelse	Säkerhetssystem oberoende av modellen
Robotplanering och färdighetsval	1–3 s vid behov varje cykel; ≤1 min om bara vid undantag eller omställningar	Utdata måste vara verifierbar kod eller ett färdighets-ID	Faller tillbaka till det fasta programmet
Q&A (SOP:er, MES-frågor)	Första token helst ≤1–2 s, högst 10 s; ≥15–20 tokens/s per användare; 5–20 samtidiga användare	Korrekta citat; trogen över lång kontext	Kontors-IT-klass (≈99–99.5%)
Analys- och rotorsaksrapporter	Minuter är okej (batch)	Högst; trogen över lång kontext	Kontors-IT-klass

Physical Intelligence mätte 97 ms inferens för en VLA i 3B-klass (cirka 108–139 ms med nätverk och förbearbetning), fann att realtidsindelning håller genomströmningen oförändrad upp till 200 ms extra fördröjning, och klarade fina uppgifter som att tända en tändsticka eller koppla en nätverkskabel vid mer än 300 ms (Physical Intelligence) vendor research; interaktionsforskning sätter gränsen för att behålla en tankegång vid en sekund och för uppmärksamhet vid tio (NN/g). Noggrannhet är det svåra: öppen Qwen3-VL-8B ligger i genomsnitt på bara 56.8

zero-shot över fem industriella anomalidatamängder (IndusAgent) single source. Ingen hårdvara räddar en seende-språkmodell som inte har finjusterats på linjens egna bilder.

FIG 07 **Hastighetsgränser per användning***Inline-inspektion mäts i millisekunder, Q&A i sekunder, analys i minuter. Människor tar 42–55 s per utrop.*



Källa: Författarens syntes (kapitel 1–2).

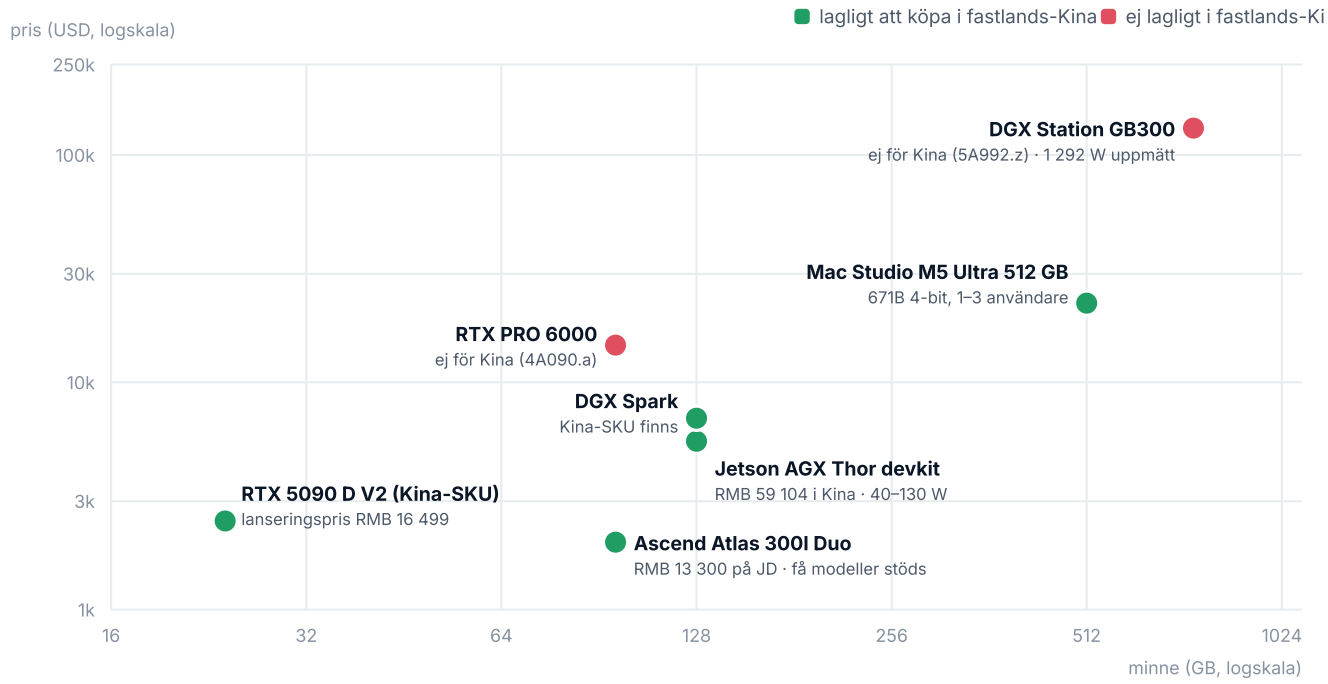
Hårdvara: Thor för stationen, en Mac Studio eller inhemsk apparat för fabriks-Q&A, och ingen DGX Station på fastlandet

Enhet	Minne och bandbredd	Pris (november 2026)	Vad den är bra för	Fastlandskina
Jetson AGX Thor dev kit / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 vid lansering → $5,499 (juli 2026); T5000 $4,999, T4000 $2,999; RS China dev kit RMB 59,104 inkl. skatt	2–8B-seendemodeller som omkontrollerar utklipp; realtids-VLA:er i 3B-klass	Ja (5A992.c, ingen licens för civila fabriker)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; ASUS GX10 Kinamodell med samma chip RMB 32,999 på JD (okt 2025)	Fleranvändar-8–35B MoE	Ja (5A992.c, Kinamodell, auktoriserade kanaler)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (juni 2026) → ≈$16,005 genomsnittlig eBay-listning i november	Fleranvändar-27–35B FP8; 120B-klass MoE i 4-bit	Nej (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB unified, 1.2 TB/s; 480 W max ihållande	Från $22,299 i USA; ≈RMB 173–186k uppskattat i Kina	DeepSeek 671B i 4-bit för 1–3 användare	Ja
DGX Station GB300	748 GB (252 GB HBM3e vid 7.1 TB/s + 496 GB LPDDR5X vid 396 GB/s); 1,292 W uppmätt vid GPU:n	$123–140k gatpris i november (≈RMB 830–940k före tull); $95–117k i augusti	Hög samtidighet på 120–235B MoE som ryms i HBM; bara några användare på 433–743B-modeller	Nej (5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	RMB 13,300 från en JD-handlare	Lågvolym-batchgranskning; få stödda modeller	Ja (inhemsk)

Enhet	Minne och bandbredd	Pris (november 2026)	Vad den är bra för	Fastlandskina
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Inte erhållet; full DeepSeek behöver två vid W8A8, fyra vid BF16	Fabriksomfattande Q&A och analys för flera användare	● Ja (inhemsk)
RTX 5090 D V2 (Kinamodell)	24 GB	RMB 16,499 vid lansering (aug 2025), upp sedan dess med minnespriser	8B-seendemodeller, små VLA:er	● Ja

Källor: (NVIDIA China Thor, NVIDIA Thor-lansering, CNX om prishöjningar, NVIDIA exportklassificeringsdata, RS China, Implicator, IThome om ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, Apple-specifikationer, TechRadar, IThome om M5 Ultra Kinapriser, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, vLLM-Ascend stödmatris, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). Fastlandspriser för 512 GB Mac Studio extrapoleras från USA:s konfigurationstillägg och är lågtillförlitliga estimate.

FIG 08 **On-prem beräkning: minne, pris och vad som lagligen går att köpa i fastlands-Kina*Exportklassificeringar avgör shortlistan före benchmarks.***



Källa: NVIDIA:s exportklassificeringar, Apple, RS China, JD.com och recensionspriser (november 2026).

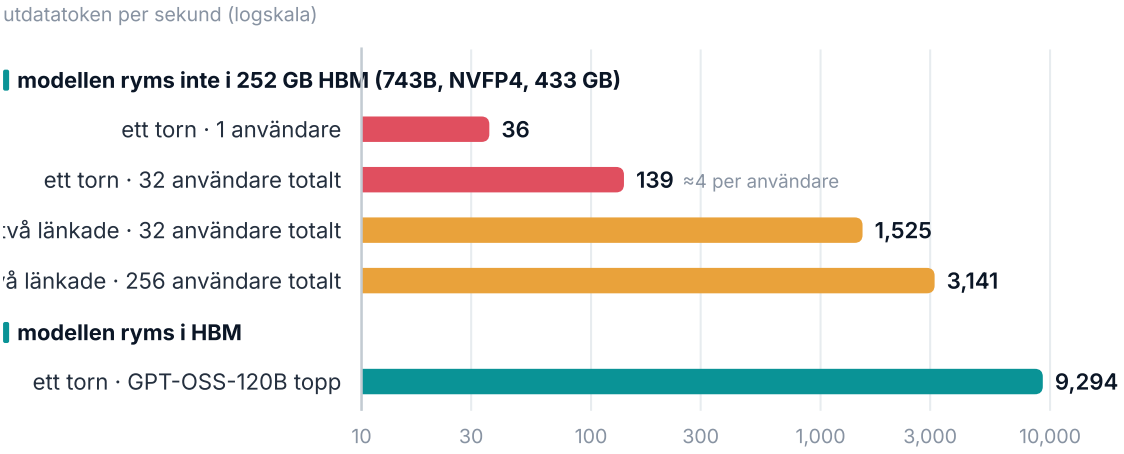
Jetson Thor får stationsnivå-distribution att fungera. NVIDIA:s TensorRT Edge-LLM-tester visar Qwen3-VL-8B (NVFP4) som prefillar 2,048 tokens på 104.5 ms och avkodar 44.4 tokens/s på Thor (TensorRT Edge-LLM) vendor claim; för ett AOI-utklipp och ett 20-tokenssvar är det cirka 0.5 s med 8B-modellen och 0.2 s med 2B, eller 8.5–21 utklipp per sekund batchat estimate. Utdatalängd, inte parameterantal, sätter latensen, så begränsad avkodning till en klasstoken eller kort JSON är en del av utformningen. GR00T N1.7 (3B) körs på 25.1 Hz på Thor (Jetson AI Lab) vendor claim. En Orin NX tar cirka 1.8 s per utklipp och en Atlas 300I Duo har en p95 över 23 s, lämplig bara för batcharbete (CSDN) single source. På fastlandet budgeteras NVIDIA-produkter till 40–60% över USA:s listpris estimate.

En 512 GB Mac Studio gör en fullstor modell på en maskin möjlig, för några användare. M5 Ultra (augusti 2026) erbjuder upp till 512 GB vid 1.2 TB/s från cirka $22,299 (Apple, TechRadar); DeepSeek 671B i 4-bit (≈380–404 GB) avkodar 17–20 tokens/s på en M3 Ultra (Pinggy) verified och uppskattat 25–29 på M5 Ultra. Långa promptar är den svaga punkten: en hämtningsprompt på 10,000 tokens tar ungefär 17–70 s estimate, bortom interaktionsgränsen på 10 sekunder.

DGX Station är den starkaste enskilda maskinen, men inte tillgänglig på fastlandet och fem till sex gånger priset på en Mac. Den levererade GB300 har **748 GB**, varav bara 252 GB är snabb HBM (Wccftech, Tom's Hardware, aug

2026) verified. Oberoende tester visar cirka 9,300 utdatatokens/s på GPT-OSS-120B, som ryms i HBM, men en 743B-modell som spiller över i LPDDR5X gav 32 användare bara 139 tokens/s totalt, cirka 4 var; två sammanlänkade enheter nådde 1,525 (StorageReview, aug 2026, StorageReview, okt 2026) single source. Gatpriser steg till $123–140k i november (CyberPowerPC, monitors.com). På fastlandet är de realistiska alternativen för samma jobb Mac Studio, KTransformers-liknande CPU+GPU-servrar (KTransformers) vendor test eller inhemska Ascend-servrar.

FIG 09 **En stor arbetsstation gör ingen delad assistent** ***När modellen inte ryms i höghastighetsminne får 32 samtidiga användare cirka 4 tokens per sekund var.***



Källa: StorageReview (aug och okt 2026); ServeTheHome. Oberoende tester från en källa.

Minnesbrist flyttade samma delar 20–100% på ett år (DGX Spark gick från $3,999 till $6,950 (TechRadar)), så utrustningsofferter bör indexeras snarare än fastställas.

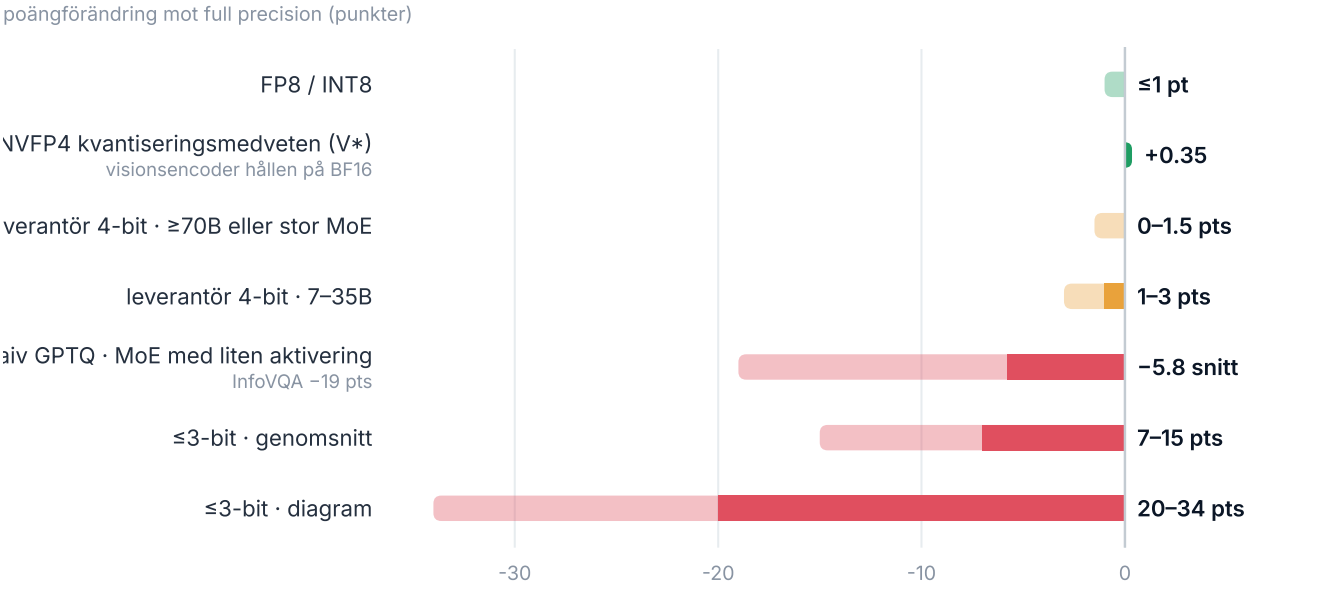
Inhemska acceleratorer kör fullstora modeller; mjukvaruunderhåll är flaskhalsen. Ascend 910B har 64 GB HBM, men vLLM-Ascend markerar DeepSeek V3/R1 som "inte längre underhållen" och stöder bara några få modeller på mindre kort (vLLM-Ascend) single source, primary, och varje ny modell tar en till två månader att anpassa (Caijing/Anue); leverera en frusen, validerad modellversion. Att betjäna full DeepSeek 671B till 20–100 användare behöver minst två Atlas 800I A2-servrar med åtta kort (China Telecom Cloud) verified, ändå uppgick offentliga upphandlingar för LLM-apparater i Kina till bara RMB 27.5 miljoner i januari–september 2025 (C114) single source: fabriker betalar för integration och tjänst, inte för en box med beräkningskraft.

Kvantisering: 8-bit är säkert, 4-bit beror på receptet, 3-bit och lägre är uteslutet

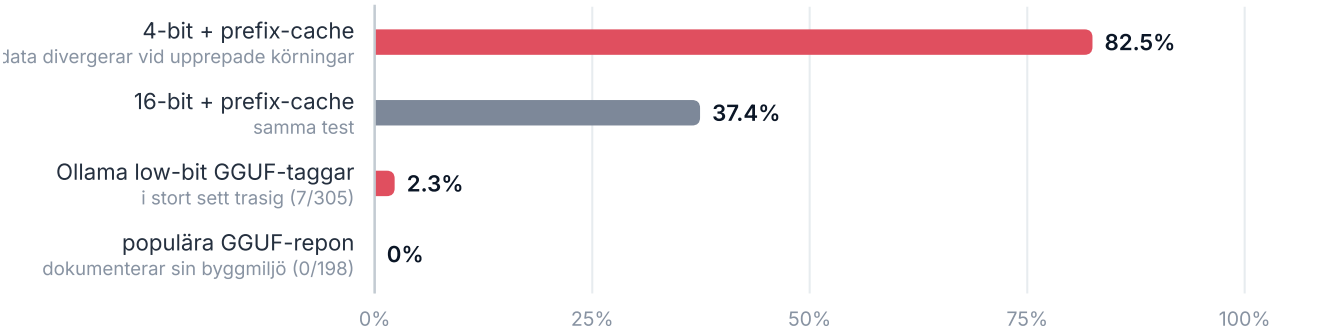
8-bit (FP8 eller INT8) är i huvudsak förlustfritt för seende-språkmodeller (Red Hat, Hugging Face) verified. Vid 4-bit beror förlusten på storlek och recept: Qwens 72B AWQ-bygge förlorar under en poäng på MMMU medan 7B förlorar 2.8 (Qwen AWQ) verified, och detaljtunga uppgifter faller mer (naiv 4-bit GPTQ kostar Qwen3-VL-30B-A3B 19 poäng på InfoVQA (VEQ)). Vid 3-bit och lägre är den genomsnittliga förlusten 7–15 poäng och diagram förlorar 20–34 (VLMQ) verified; bortom 64k tokens förlorar 4-bit upp till 59%, värre på icke-engelska indata (arXiv 2505.20276). Håll seende-encodern på FP16/BF16.

FIG 10 **Kvantisering: 8-bit är säkert, 4-bit beror på receptet, 3-bit och lägre åker ut*Ingenjörsmissar i servingstacken kostar mer träffsäkerhet än själva kvantiseringen.***

träffsäkerhet: 8-bit är säker, 4-bit beror på receptet, ≤3-bit åker ut



ingenjörsmissar väger tyngre än kvantiseringsfel



Källa: NVIDIA Nemotron-modellkort, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 med flera.

Två varningar väger tyngre. För det första har **ingen studie eller modellkort per den 11 november 2026 mätt kvantisering på PCB/SMT-defektbilder**, så omvalidera på guldsetet. För det andra är betjäningsstackfel ofta större än kvantiseringsfelet: samma modell fick cirka 30 poäng lägre under Ollama än under referensimplementationen (SiliconBench); 7 av Ollamas 305 officiella kodmodellstaggar var i grunden trasiga trots att filerna var intakta (arXiv 2609.05881); inget av 198 populära GGUF-repositories noterade sitt bygge (arXiv 2610.19498); och med prefix-cache-återanvändning divergerade 82.5% av episoderna vid 4-bit mot 37.4% vid 16-bit (arXiv 2609.04748) single sources. Leverantörs-kvantiseringsmedveten 4-bit är nära förlustfri på detaljuppgifter (NVIDIA Nemotron 3.5) vendor claim, men 4-bit VLA:er förlorar 1.5–4.3 poäng under störning och körs långsammare på Thor (LoopQ) single source. Hantera en kvantiserad modell som ett nytt mätinstrument: fastnåla vikthashar, notera bygget, återställ prefix-cache och omvalidera efter varje förändring av precision, motor eller mall.

Precision	≥70B dense eller ≥100B MoE	7–35B	≤4B eller detaljtunga uppgifter	Dom
FP8 / INT8	Säkert (≤1 poäng)	Säkert	Säkert; standard för inspektion	Använd

Precision	≥70B dense eller ≥100B MoE	7–35B	≤4B eller detaljtunga uppgifter	Dom
4-bit AWQ/GPTQ (leverantörsrecept, seende-encoder FP16)	0–1.5 poäng	1–3 poäng; naiv GPTQ på small-active MoE −5.8 i genomsnitt, −19 på InfoVQA	2–3 poäng; validera detaljuppgifter först	Villkorat
NVFP4 (leverantörsrecept)	≈0.3–1 poäng under FP8 (leverantörsdata)	Kvantiseringsmedvetna byggen nära förlustfria på V*, OCRBench	Ingen offentlig data	Villkorat (föredra kvantiseringsmedvetna byggen)
GGUF Q4_K_M	Otestat	Hög varians på domänuppgifter	Inte för liten text eller defekter	Inte för detaljbedömning
GGUF Q5_K_M och högre	Förmodat säkert	Nära förlustfritt i två tester	Golvet för detalj på llama.cpp	Använd
≤3-bit (inklusive 1T-modeller pressade till 1.8–2 bitar på en maskin)	Inte för beslut	Nej	Nej	Nej

Vilken konfiguration för vilket jobb

Konfiguration	Inline-inspektion	AI-granskning av falsklarm	Robotplanering och styrning	Q&A och analys
Dedikerade detektorer eller anomalimodeller (YOLO, RF-DETR, PatchCore-klass), INT8/FP16, vilken NPU- eller GPU-edge-box som helst	● Ja: huvuddomaren, 2–12 ms per utklipp	● Ja (klassificerare finjusterad på linjen)	● Endast perception	—
2–4B-seendemodeller, NVFP4/AWQ/FP8 med FP16-seende-encoder, på Thor/T4000 eller DGX Spark	● Villkorat: omkontrollera AOI-flaggade utklipp på 0.2–0.3 s vardera	● Villkorat, under 1 s efter finjustering	● VLA-stomme (2–3B VLA vid ≈40–130 ms per block)	● Svag
7–9B-seendemodeller (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 eller leverantörs-4-bit, på Thor, Spark eller ett 24–32 GB-kort	● Villkorat: ≈0.5 s per flaggat utklipp	● Ja: 0.5–1 s per anrop, bäst värde	● Planering och färdighetsval (1–3 s), inte i styrkretsen	● Grundläggande SOP-Q&A för en eller några användare
26–35B MoE (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● Nej (för tung)	● Ja, som förklarare eller för svåra fall	● Planering	● Ja (8 användare på en Spark, ≈0.1 s första token)
106–241B MoE i 4-bit på en Mac Studio, RTX PRO 6000 eller Spark	● Nej	● Nej	● Offline-planering	● Ja, för en eller några användare
DeepSeek 671B eller Kimi K2.x 1T i 4-bit på en 512 GB Mac, KTransformers-server eller DGX Station	● Nej	● Nej	● Endast offline-programgenerering	● Villkorat: 1–3 användare och korta promptar; långa promptar tar 16–170 s

Konfiguration	Inline-inspektion	AI-granskning av falsklarm	Robotplanering och styrning	Q&A och analys
Vad som helst vid ≤3-bit; GGUF Q4_K_M för detaljarbete med små modeller; Ollama för många användare	● Nej	● Nej	● Nej	● Nej

Källor: TensorRT Edge-LLM-, Jetson AI Lab-, Apple ML Research- och KTransformers-data ovan, samt (DGX Spark-tester, RTX PRO 6000-tester, Hardware Corner om Atlas, Ultralytics YOLO26, RF-DETR) **community- och leverantörstester, mestadels single source** .

Så, till de två vanliga orosmomenten: **det behöver inte vara långsamt** om modellen matchar jobbet (små och medelstora seendemodeller på Thor eller Spark klarar varje inline- och granskningsbudget; 671B–1T-modeller hålls borta från linjen), och **kvantisering behöver inte göra det fel** med FP8/INT8 som standard, eller leverantörsrecept-4-bit över 7B validerad på guldsetet. De större riskerna är den ungefär 57 poäng låga zero-shot-industrinoggrannheten hos generella modeller och betjäningsstackfel. Ge inspektionsförfrågningar fasta batchstorlekar och en egen kö (temperatur 0 är inte bit-reproducerbar under batching (vLLM-dokumentation)), håll en N+1 varm standby, och gör varje steg på linjen felsäkert öppet (fail open).

Lärare och elev

03 Lärare och elev: frontiermodeller kan förmärka inspektionsdata; för robotarmar är de verkliga lärarna fortfarande människor

November 2026 års lärarkandidater: generiskt seende fördubblades, industridefektpoäng saknas

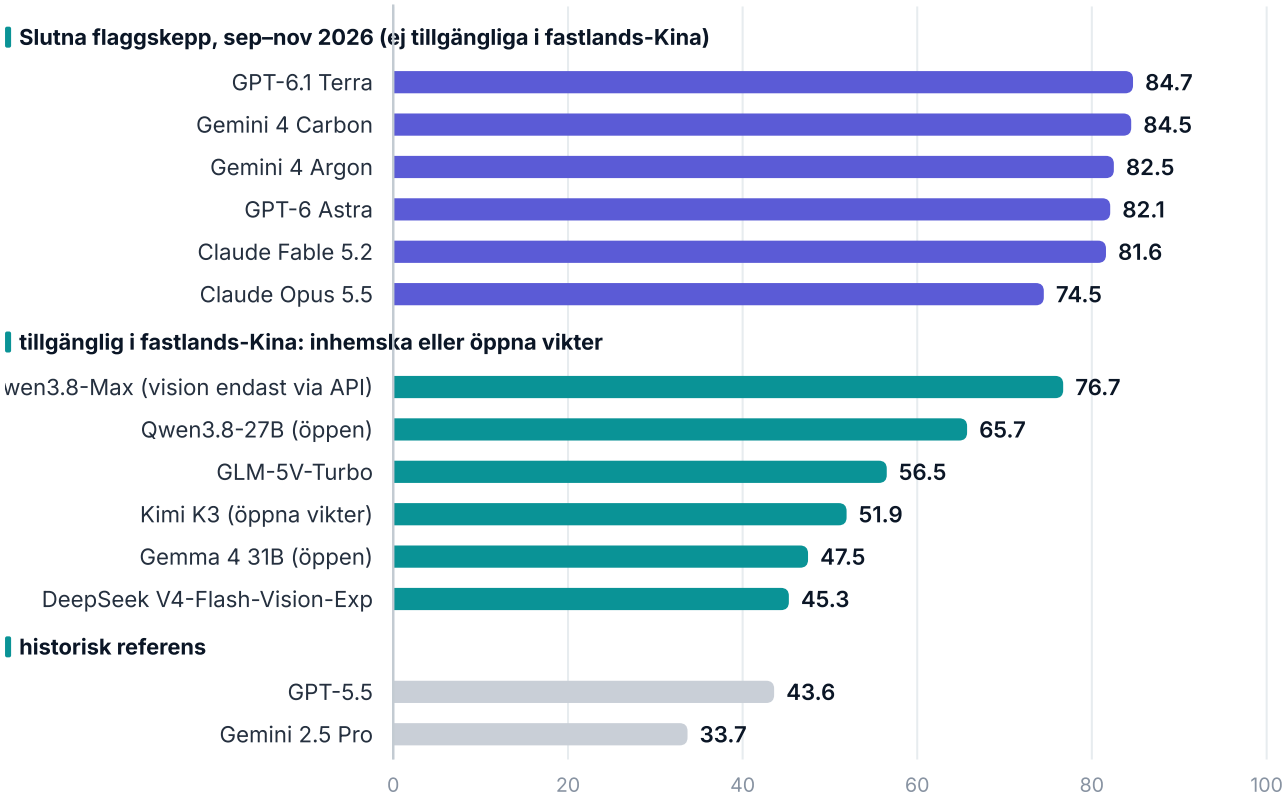
Varje stängt flaggskepp som skulle kunna agera lärare i november 2026 tar emot bilder: Claude Fable 5.2 (allmänt tillgänglig 9 november, $8/$40 per miljon tokens) och Opus 5.5 ($4/$20) (Anthropic, Claude-modellöversikt) verifi-ed; GPT-6 Astra ($10/$50) och GPT-6.1 Terra ($1/$5) (OpenAI-prissättning); och Gemini 4 Argon och Carbon, öpp-nade i etapper (Google Carbon). Användbara i Fastlandskina är Qwen3.8-Max (seende via API) och öppna modeller som Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp och Gemma 4 (Roboflow-blogg, Hug-ging Face).

Den centrala luckan måste sägas tydligt: ingen av dessa modeller har ett publicerat resultat på MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB eller någon PCB/SMT-datamängd. MMAD har fortfarande ingen topplista (MMAD GitHub), och de två nyaste industriella anomalipapperen (augusti–september 2026) använder fortfarande Gemini 2.5 Pro och Claude Sonnet 4 som sina nyaste generella baslinjer (InspectorGPT, Anomaly-LR). En arXiv-fulltextsök-ning den 11 november fann noll träffar för "Fable 5.2" eller "Gemini 4 Carbon" i detta sammanhang (arXiv API) single source, absence of evidence. De närmaste ersättningarna:

- Den nyaste frontiermodellen med ett MMAD-resultat är GPT-5.5: 69.95% på "defekt eller inte" och 76.20% i genomsnitt över sju uppgifter på ett 500-bilders delurval, mot 95.24% och 86.65% för mänskliga experter (AnomalyClaw) single source; GPT-4o fick 68.63% och 74.92% 2024 (MMAD, ICLR 2025), ungefär 1.3 poängs framsteg på två år.
- På LDU-Bench, riktiga SEM-granskningsbilder från IC-litografi och det enda elektronikbenchmark med 2026-modeller, är bra-eller-dålig-triage hygglig (makro-F1 0.932 för GPT-5.4, 0.884 för GLM-5V-Turbo, 0.876 för Claude Opus 4.6, 0.504 för Gemma 4 31B), men klassificering av 11 defektmorfologier når bara 0.20–0.41 (LDU-Bench) single source.
- På Roboflows oberoende seende-utvärderingar (uppdaterade 10 november 2026) är objektdetektering mAP@50 84.5 för Gemini 4 Carbon, 84.7 för GPT-6.1 Terra och 81.6 för Fable 5.2, mot 33.7, 43.6 och 38.6 för Gemini 2.5 Pro, GPT-5.5 och Claude Opus 4.8 (Roboflow Vision Evals) single source, oberoende utvärderare, privata datamängder.

FIG 11 **Lärarkandidater: generiska poäng fördubblades på en generation*Men inget av 2026-flaggskeppen har en publicerad poäng på en industriell defektbenchmark.***

Roboflow generisk detektering mAP@50



industriella defektbenchmarks (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): noll publicerade poäng för dessa 2026-flaggske
generiska detektionspoäng fördubblades på en generation, men det är ingen defektbedömning; testa på din egen gyllene uppsättning

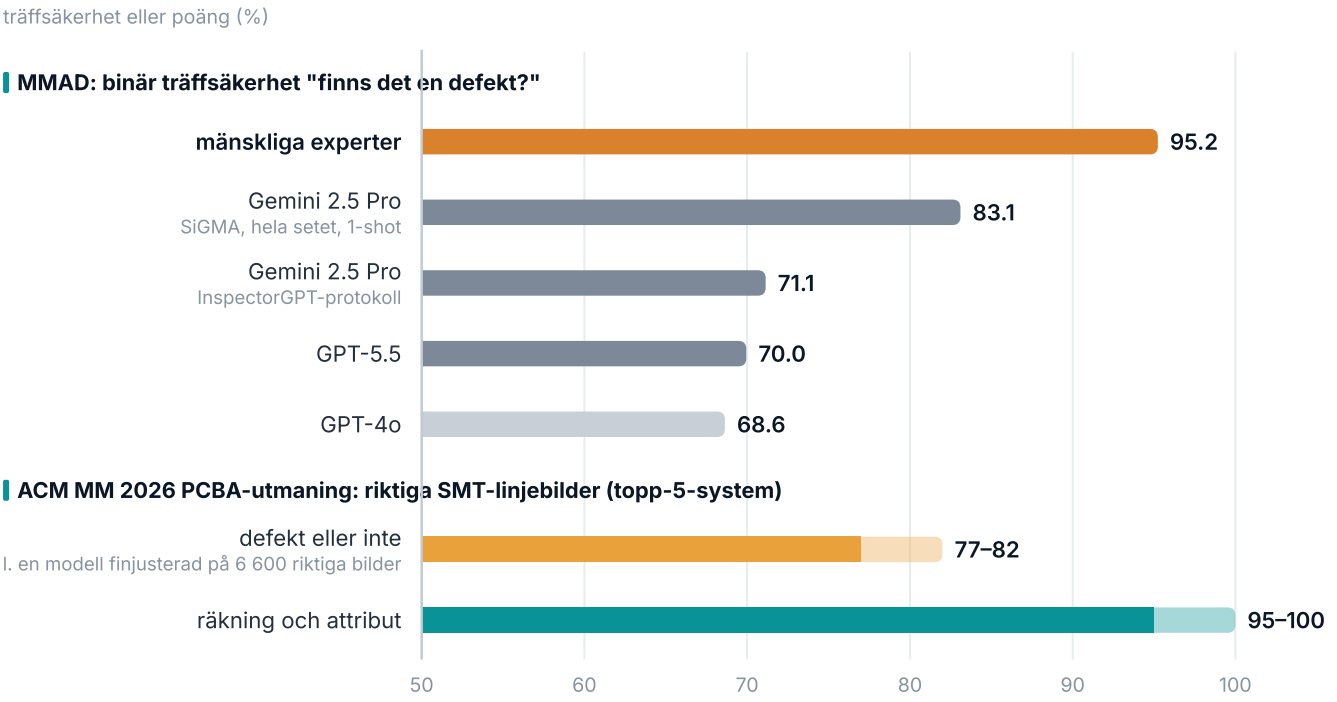
Källa: Roboflow Vision Evals; modellrelease-noter; MMAD, MVTec AD 2 och andra benchmarksidor (per 11 nov 2026).

Lärarkandidat (nov 2026)	Pris per M tokens (in/ut)	Roboflow detektion mAP@50 / räkning	Industridefekt-benchmark	Användbar i Fastlandskina
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Inga	Nej
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Inga	Nej
GPT-6 Astra	$10 / $50	82.1 / 80.2	Inga	Nej
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Inga	Nej
Gemini 4 Carbon	$2 / $10 (kampanj)	84.5 / 90.1	Inga	Nej
Qwen3.8-Max (seende endast via API)	$2 / $6	76.7 / 81.1	Inga	Ja
Qwen3.8-27B (öppen)	Lokal	65.7 / 64.9	Inga	Ja
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench triage 0.884	Ja
Gemma 4 31B (öppen)	Lokal	47.5 / 51.4	LDU-Bench triage 0.504	Ja (lokala vikter)
Referens: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD binär 69.95% / 68.63%	Nej

Lärarkandidat (nov 2026)	Pris per M tokens (in/ut)	Roboflow detektion mAP@50 / räkning	Industridefekt-benchmark	Användbar i Fastlandskina
Mänskliga experter	—	—	MMAD binär 95.24%	—

Nyare papper testar fortfarande föregående generation, men pekar åt samma håll. På hela MMAD-setet får mänskliga experter 95.2% på defekt-eller-inte och Gemini 2.5 Pro 83.1% (SiGMA), ändå fick samma modell bara 71.14% i ett annat papper (InspectorGPT): testprotokollet flyttar poängen lika mycket som en modellgeneration gör. I CVPR 2026 VAND 4.0-utmaningen nådde färdiga seende-språkmodeller 61.88 AP mot 90.84 för övervakade metoder, och arrangörerna drog slutsatsen att de "för närvarande inte kan ersätta finjusterade klassificerare" (VAND 4.0) single source. I ACM MM 2026 PCBA-utmaningen fick de fem bästa 76.92–82.47 på defekt-eller-inte medan komponenttyp och pin-räkning nådde 95–99.5 (arXiv 2609.21276) single source: att känna igen är mycket lättare än att bedöma.

FIG 12 **"Finns det en defekt?" är fortfarande svårt för generella modeller** ***Den bästa generella modellen ligger mer än 12 poäng efter mänskliga experter på MMAD; tunade system gör det bättre på riktiga linjebilder.***



Källa: MMAD; SiGMA; InspectorGPT; ACM MM 2026 PCBA-utmaning.

Tre slutsatser följer. Att se och räkna fördubblades på en generation, men att bedöma avvikelser rör sig långsamt, beror kraftigt på protokollet och har systematiska snedheter; hur mycket de nya modellerna vinner kan bara mätas på en fabriks egna bilder. Inom sin träningsdistribution slår finjusterade 7B-specialister fortfarande generella modeller (Anomaly-LR, UniPCB). Och riktiga linjedata kollapsar poängen: på monterade elektronikpaneler föll bildnivå-AUROC från 97–98 på MVTec AD till 31.5–78.6 (arXiv 2606.01992) single source. För fastlandsföretag ligger den starkaste tillgängliga läraren, Qwen3.8-Max, cirka 8 poäng under den globala fronten.

Att köra dussintals eller hundratals rundor: pengar är inte problemet, etikettkvalitet är

Pengar är inte problemet. Till november 2026 års listpriser kostar 1,000 enmegapixelbilder cirka $3 (GPT-6.1 Terra) till $32 (GPT-6 Astra) per runda, och hundra rundor över 100,000 bilder cirka $32,000–135,000 före batchrabatter estimate (Claude seende-dokumentation, OpenAI seende-guide, Roboflow Vision Evals). För enkla etiketter kostar

en person ungefär lika mycket som Terra; lärarens fördel är hastighet och skala, inte pris. Kundernas NDA:er håller oftast produktbilder inne i fabriken, så guldsetet etiketteras på plats.

Det verkliga hindret är kvalitet, och fler rundor åtgärdar inte systematiska fel. Roboflow kör varje uppgift tre gånger och poängen rör sig bara ±0.1–1.4 poäng (Roboflow), så en modell upprepar mest sina misstag och majoritetsröstning kan inte rädda den. Rundor hjälper bara när varje runda tillför information: en eller två bra referensbilder lägger till cirka 3–4 poäng, fler tillför inget (AD-Copilot); att låta en modell argumentera mot en bra referens över flera vändor lägger till 6.23 poäng av tvärdomän-AUROC för GPT-5.5 och 2.95 på MMAD, med topp efter ungefär tre anrop (AnomalyClaw) single sources. "En elev som tränas på lärarens etiketter närmar sig mänskliga etiketter" har bara visats på naturliga bilder: en ResNet-101 som tränats på Qwen3-VL-etiketter av COCO och Objects365 nådde 95.7–99.4% av resultatet med mänskliga etiketter, och författarna säger att domändata är otestad (arXiv 2602.20972) single source.

FIG 13 **Två lärare–elev-pipelines** ***För inspektion för-märker en frontmodell och människor avgör; för robotarmar är de riktiga lärarna fortfarande mänskliga demonstrationer.***



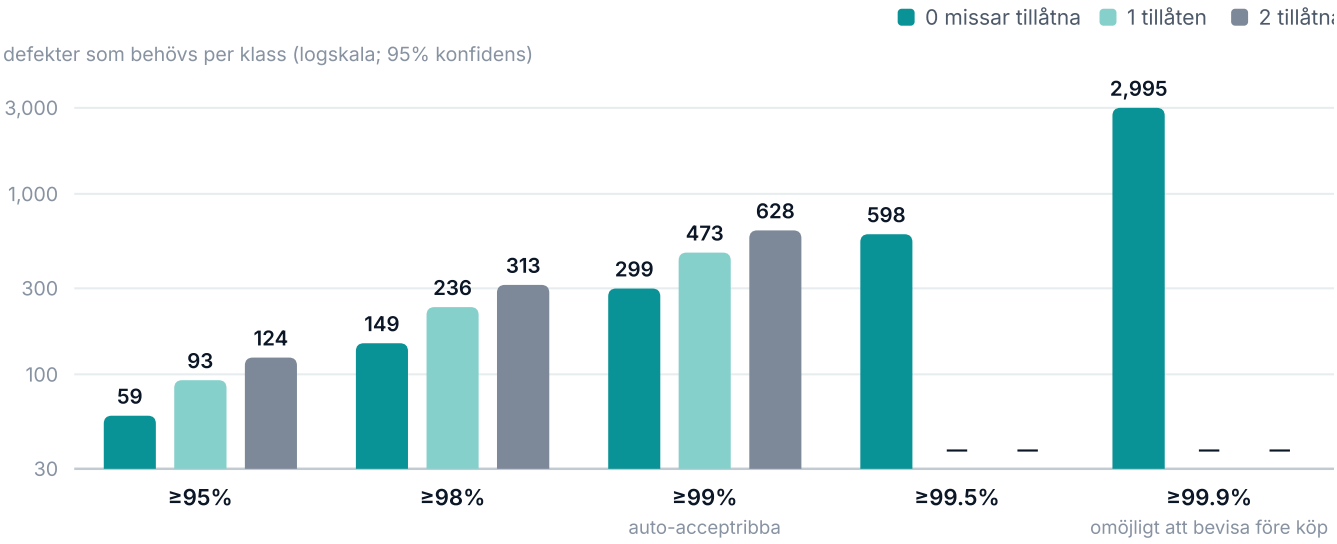
Källa: Författarens syntes (kapitel 3); kostnader och stickprovsstorlekar är uppskattningar.

Ett fungerande protokoll estimate:

1. **Samla guldsetet från fabrikens eget AOI-arkiv och sparade NG-kort**, stratifierat i 12 sanna defektklasser (saknad, förskjutning, gravsten, polaritet, fel del, brygga/kortslutning, otillräckligt löd, överflödigt löd, lödkula, lyft ben/öppen, skadad del, främmande material) och fyra falsklarmsorsaker (skugga, fluss, silkscreen, tolerans), plus en "drift-slice" på minst 300 defekter från en andra linje eller annan belysning, kortfärg eller artikelnummer.
2. **Etikettera blint med två eller tre IPC-A-610-utbildade inspektörer**, med en senior ingenjör som avgör och röntgen, elektriskt test eller tvärsnitt som avgör tvetydiga prover. Acceptera en klassdefinition endast om inspektörsöverensstämmelsen når $\kappa \geq 0.75$ (≥0.90 för kritiska klasser); mänsklig överensstämmelse om 11 elektronikdefektmorfologier i LDU-Bench var bara 0.727, så vissa klasser behöver slås samman eller omdefinieras. AIAG-liknande attributöverensstämmelse (effektivitet ≥90%, miss ≤2%, falsklarm ≤5%, $\kappa > 0.75$; >0.90 för release) är ett rimligt minimum (Fabrico, iFactory).
3. **Dimensionera med exakta binomiala (Clopper–Pearson) ensidiga 95%-gränser** (binomiala konfidensintervall, rule of three:

Recall att bevisa	0 missar tillåtna	1 tillåten	2 tillåtna
≥95%	59 defekter	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **Hur stor den gyllene uppsättningen måste vara*Att bevisa 99% recall med noll missar kräver 299 defekter per klass; 99.9% kan inte bevisas före köp.***



Källa: Exakta binomiska konfidensintervall (95%).

Sextio defekter per klass räcker för en godkänd/underkänd-screening; 300 behövs för att besluta om automatisk acceptans. Hela setet, cirka 3,600 defekter plus 3,500 falsklarmsutklipp, tar cirka 37–57 arbetstimmar att etikettera och avgöra, och att köra en lärarkandidat över det kostar $23–196; den verkliga kostnaden är två till sex veckors insamling, satt av den sällsyntaste klassen estimate. Skicka enkomponents-AOI-utklipp, inte hela kort, med målet minst 42–48 pixlar tvärsöver (arXiv 2610.09313).

1. **Acceptera lärarens etiketter automatiskt endast för en klass med noll missar på ≥299 guldset-defekter**, modell–sanningsöverensstämmelse inte sämre än människa–människa-överensstämmelse, och minst 50% av falsklarmen rensade. Kritiska klasser som kortslutningar, saknade delar, polaritet och fel delar behöver 99.5–99.9% (598–2,995 defekter); annars går allt AI:n inte kan rensa med hög konfidens till en person. En läckagegrad under 0.1% av defekterna kan inte bevisas före köp (2,995 defekter med noll missar; 0.01% skulle ta 29,956), så linjen behåller mänsklig stickprovstagning av AI-rensade anrop, nedströms ICT/FCT som skyddsnät och SPC på läckage. Verifierade etiketter destilleras sedan till en konventionell detektor eller klassificerare och omvalideras på varje ny produkt.

Villkor och dataregler avgör också läraren. Anthropic förbjuder att använda sina tjänster för att träna konkurrerande modeller (Anthropic kommersiella villkor); OpenAI tillåter att utdata tränar interna klassificerare som inte säljs till tredje part (OpenAI affärsvillkor); DeepSeek tillåter uttryckligen destillering (DeepSeek-villkor) primary terms. Ingen definierar en "konkurrerande modell", så kommersiell användning behöver juridisk granskning, och kundernas dataregler håller oftast bilder på plats. Den säkraste utgångspunkten är en lokalt körd lärare med öppna vikter, med molnbaserade frontiermodeller bara när kunden tillåter det skriftligt.

Inspektionselever: små och snabba; flaskhalsarna är optik, data och validering

Eleven är inte flaskhalsen: YOLO26n tar 3.80 ms per bild i INT8 på en $249–399 Orin Nano Super (Ultralytics) vendor test, NVIDIA rapporterar 98.51% på tre PCB-defektklasser efter att ha anpassat en grundmodell på cirka 600 bilder (NVIDIA) vendor claim, och syntetiska defekter lägger bara till 0.9–2.5 mAP-poäng (arXiv 2605.04635). Begränsningarna är optik, linjedata, validering och licenser: Ultralytics YOLO är AGPL-3.0 eller företagslicensierad (Ultralytics), så stängda produkter behöver en licens eller en Apache-2.0-detektor som RF-DETR.

Robotelever: strukturerade stationer är nära produktionsklass; flexibelt och dynamiskt arbete är det inte

I november 2026 finns tre sorters robotlärare, och bara en kan bli en lokal elev: öppna, finjusterbara VLA:er (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Stängda robotmodeller (π0.6/π0.7 genom ett beta-API, Gemini Robotics 2 för betrodda testare) kan inte distribueras på plats (The AI Insider, DeepMind). Frontiermodeller av generell typ är långsamma och dåliga på finmanipulation: i Artificial Analysis test fick de 46–69 med 28–45% greppframgång och cirka 30 minuter per uppgift, "alldeles för långsamt och dyrt för någon produktionsanvändning" (AA-Robotics); GPT-6 Astra klarade ett precisionsspår bara 2 gånger av 20 (RoboCurve); och att driva leder direkt ger 0–5.5% på LIBERO vid 0.2–0.4 Hz mot cirka 83 Hz som behövs (Anthropic) verified across three sources. Fable 5.2:s 64.9% på det simulerade RoboDojo-benchmarket är en intern körning (Anthropic) vendor claim.

Tredjepartsstudier sedan oktober är överens: cirka 33–44% på LIBERO-10 för de tre ledande agenterna (arXiv 2610.30051); agenter som falskeligen påstår framgång på 24–39% av deluppgifterna (arXiv 2611.08209); och 104 av 108 kvasistatiska uppgifter på riktiga robotar för Opus 5.5, men på 280 sekunder vardera och utan insättning eller montering (AMP1) single sources. Frontiermodeller hör hemma offline: skriva program, orkestrera uppgifter och utarbeta framgångskriterier.

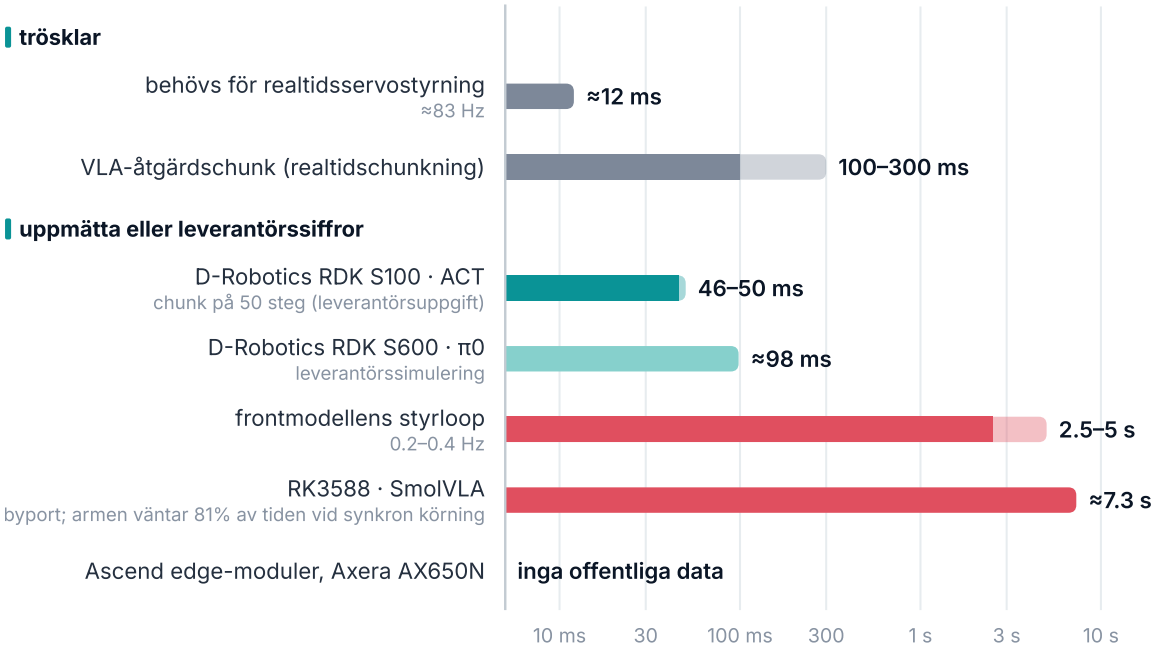
Elevkvalitet har en tydlig kant: bra inom träningsdistributionen, kollapsar utanför den. π0.7 "överstiger ofta 90%" på sedda uppgifter och når 60–80% på osedda (π0.7-papper) vendor research; ContactDP når 97.8% på tränade kontakter men 28.6–66.7% på osedda (ContactDP); små kamera- eller poseförändringar sänker 95% under 30% i LIBERO-Plus (LIBERO-Plus, CVPR 2026); inlärt grepp från rörliga transportband toppar på 58.89% (ForeTime-VLA) single sources, ett jobb som klassisk seende-tracking har gjort i produktion i åratal.

Bevisen för nära produktionsklass kommer alla från strukturerade stationer, mänskliga demonstrationer och korrigering på riktiga robotar. Neuromeka behöll en tillståndsmaskin och förinlärd rörelse och lärde sig bara grepp, kabelinsättning och lödning, var och en från under 20 minuters data: 322 av 324 i produktion, 12.8% långsammare än en person (Neuromeka) single source, vendor. Fraunhofer nådde 92.8% över fyra kontakter från 68 minuters demonstrationer (Fraunhofer); Berkeleys HIL-SERL lyckades 100 gånger av 100 på RAM-, SSD- och USB-insättning efter 1–2.5 timmars RL på riktiga robotar (HIL-SERL); AgiBot körde 18–20 sekunder per steg inom en 20-sekunders takt på en surfplatta-testlinje (IThome) vendor claim.

Destillering köper hastighet, inte generalisering: π0.5 körs på 0.70 Hz på Jetson Orin rå och 10.30 Hz efter asynkron schemaläggning (EagleVLA); lagerdestillering tar den från cirka 3 Hz till cirka 10 Hz (Shallow-π) verified. En elev ärver bara det läraren täckte; varje ny produkt behöver fortfarande uppgiftsdata.

FIG 15 **Edge-policylatens***En liten ACT-policy körs på cirka 50 ms på ett edge-kort; en VLA portad till ett hobbyschip tar 7.3 s.*

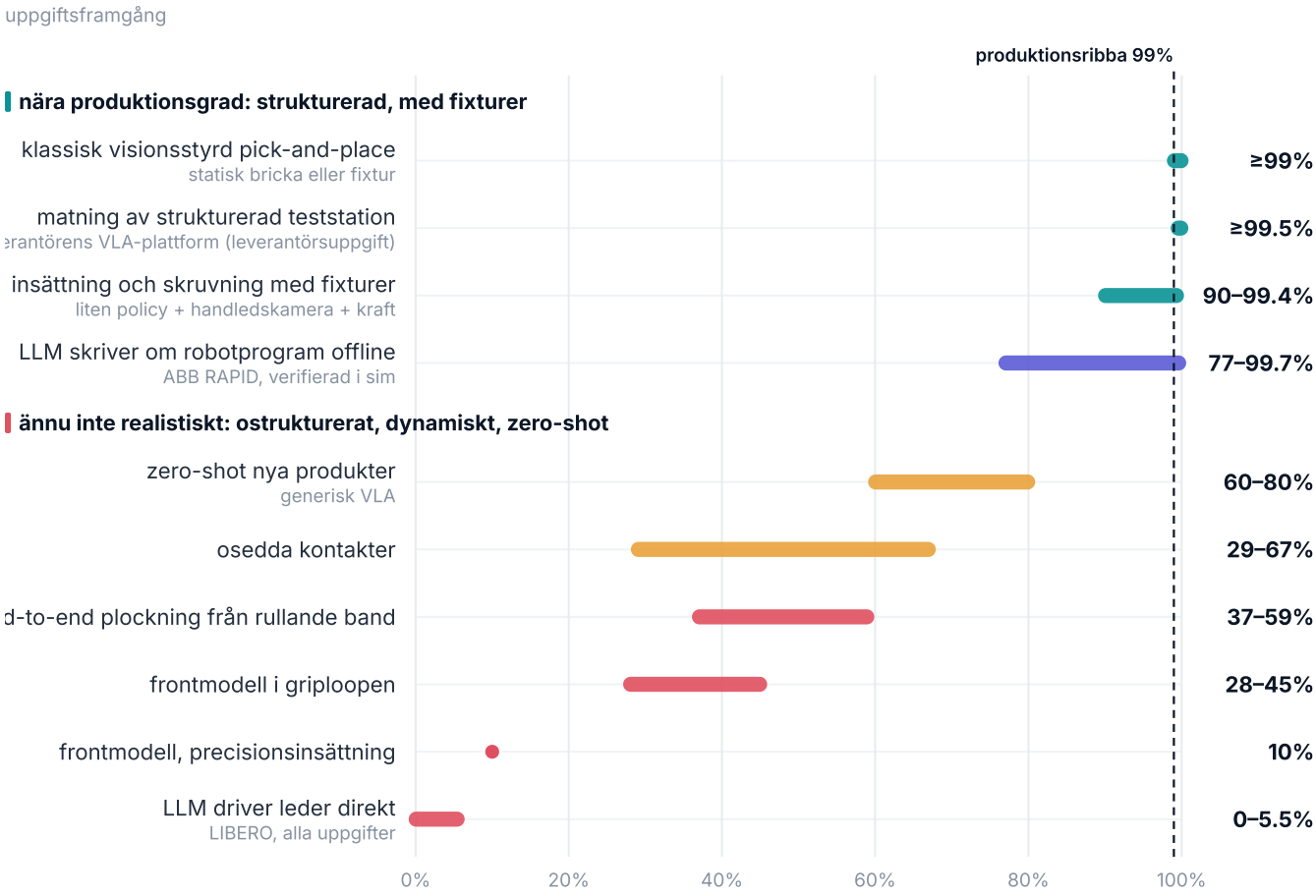
inferenslatens per åtgärdschunk eller styrcykel (logskala)



Källa: D-Robotics dokumentation och forum (leverantörssiffror); community-portar på GitHub; Anthropic-forskning.

På kinesiska edge-chip är den enda uppmätta inlärda policyn en ACT-policy på D-Robotics RDK S100 vid 46–50 ms per 50-stegsblock, utan publicerad framgångsgrad (D-Robotics RDK S-dokumentation) vendor claim; en hobbyport av SmolVLA på en RK3588 tar cirka 7.3 s per block (GitHub). Kvantisering köper inte hastighet gratis: inget 4-bit π0.5-bygge var snabbare än 16-bit vid batchstorlek 1 (LoopQ). Demonstrationer är dyra (Reuters rapporterade cirka 300 försök per användbar rörelse för en nybörjare inom teleoperation (Notebookcheck)), och validering kostar mer: att bevisa en felfrekvens ≤0.1% vid 95% konfidens tar cirka 3,000 på varandra följande framgångar, 17–22 robottimmar estimate. Ingen inlärd manipulationspolicy har säkerhetscertifiering.

FIG 16 **Robotelever: nära produktionsgrad bara när stationen är strukturerad** ***Stationer med fixturer når 90–99%+; ostrukturerade, dynamiska och zero-shot-uppgifter är ännu inte realistiska.***



Källa: Artiklar och leverantörsmaterial citerade i kapitel 3; vissa siffror är leverantörsuppgifter.

Elektronikuppgift	Fungerande elev	Förväntad framgång	Hastighet vs en person	Kostnad per omställning	2026–2028 dom
High-mix pick-and-place, kittning, maskinbetjäning (statiskt tråg eller fixtur)	Inlärd perception plus programmerad eller planerad rörelse	≥99% med klassisk seende-vägledning	Samma som konventionella stationer	Timmar	Genomförbart nu
Fixerad kontakt-, FFC- och kabelinsättning, skruvdragning, lödning	Tillståndsmaskin + liten BC/diffusion/ACT-policy + handledskamera + kraft/moment (helst beröring) + mänsklig korrigering	90–99.4%	10–15% långsammare	10–300 demos (≈15–70 min) + 1,000–3,000 valideringskörningar (6–22 robottimmar)	Genomförbart, FoU-tungt; månader för den första stationen
Strukturerad teststationsbetjäning och OK/NG-sortering	Leverantörs-VLA-plattform + digital tvilling + RL på riktiga robotar	≥99.5% vendor claim, conflicting	18–20 s per steg, inom en 20 s-takt	≤4 h omskolning + 15 min kalibrering	Genomförbart på en leverantörsplattform

Elektronikuppgift	Fungerande elev	Förväntad framgång	Hastighet vs en person	Kostnad per omställning	2026–2028 dom
Omställningsprogrammering	LLM skriver eller redigerar robotprogram offline, verifierat i simulering	77–99.7% på ABB RAPID-omskrivningar	Ingen effekt på takt	Minuter till timmar per program	**Genomförbart som ingenjörsverktyg**
Änd-till-änd-inlärt plock från rörliga transportband	VLA-elev	37–59%; ≤37% på snabba band	—	≥458 demos per uppsättning	**Inte före 2028; använd klassisk tracking plus inlärd perception**
Nya produkter utan ny data (zero-shot)	Generisk VLA	60–80%; osedda kontakter 29–67%	—	—	**Inte realistiskt**
Frontier-LLM i loopen som driver armen	GPT-6-, Claude-, Gemini-agenter	Grepp 28–45%; precisionsinsättning 10%	≈30 min eller 280 s per uppgift vs ≈20 s	—	**Inte realistiskt**
Fritt dragna kabelhärvor och flerstegs flexibel montering	—	Inga produktionsbevis 2026	—	—	**Inte före 2028**

Källor: studierna ovan samt (University of Passau om RAPID-omskrivning, GraspVLA, Komax) **mestadels single source och leverantörsdata** . Under 2026–2028 är en stor modells värde för robotar som en förtränad prior och en offline-programmerare, inte som något som "lär" en elev. Varje fall nära produktionsklass jag fann (Neuromeka, AgiBot hos Longcheer, Fraunhofer, ContactDP) förlitar sig på mänskliga demonstrationer, kraft- eller beröringsavkänning, simulering eller digitala tvillingar och övervakad RL. Inget förlitar sig på etiketter från en frontiermodell.

Kostnader och återbetalningstid

04 Kostnader och återbetalningstid: AI-granskning betalar sig på månader, maskinbetjäning på år, utplockningsarmar aldrig

Arbetskostnad: USA är cirka 5.5 gånger Shenzhen, Taiwan cirka 1.5 gånger, större delen av Sydostasien under hälften

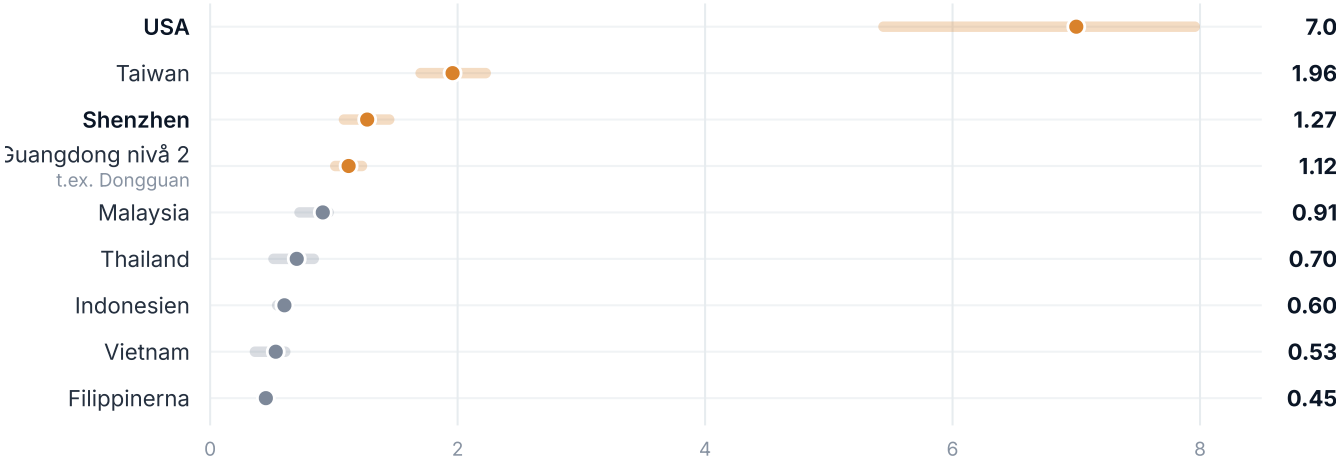
Avkastningen på att ersätta arbetskraft börjar med vad arbetskraften kostar. Shenzhens officiella lönevägledning sätter 2024 års median för en kvalitetsinspektör inom tillverkning till **RMB 74,074** per år; löner utgör 86.92% av den totala arbetskostnaden inom elektroniktillverkning, så den fullt belastade kostnaden är cirka RMB 85,000. JETRO:s undersökning från 2025 sätter årskostnaden för en kinesisk fabriksarbetare till $12,667, cirka RMB 85,100, inom 0.2% (Shenzhens lönevägledning, JETRO 2025) verified .

Region	Fullt belastad årskostnad för en inspektör eller hanterare (låg / bas / hög)	Grund
Fastlandskina, Shenzhen	RMB 70k / **85k** / 100k ($10.4k / **12.7k** / 14.9k)	Officiell median × 1.15; matchar JETRO verified
Guangdong tier-2-löneområden (Dongguan etc.)	RMB 65k / **75k** / 85k	Guangdong 2023 median × 1.15 plus två års höjningar estimate
Yangtzeflodsdeltat (Suzhou, Kunshan, Wuxi)	Som Shenzhen	Ingen grund för billigare inspektionsarbete; Jiangsus högsta minimilön är cirka 5.6% över Shenzhens estimate
Taiwan	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	Arbetsmarknadsdepartementets yrkeslöneundersökning plus ≈1.18 lagstadgade påslag; JETRO som högänden estimate
Malaysia	$6.8k / **9.1k** / 10.0k	JETRO och minimilöner estimate
Thailand	$4.7k / **7.0k** / 8.8k	Samma
Indonesien	$5.0k / **6.0k** / 6.3k	Samma
Vietnam	$3.2k / **5.3k** / 6.5k	Samma
Filippinerna	$4.0k / **4.5k** / 4.9k	Samma
USA	$54k / **70k** / 80k	BLS maj 2025 inspektörsmedian $48,570 ÷ löner vid 65.5% av kompensation estimate

Källor: (JETRO 2025, Guangdong lönevägledning, Wuxi HR-myndighet, Taiwan yrkeslöneundersökning, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **Vad en inspektions- eller hanteringsarbetare kostar per år** ***Fullt belastade är USA ungefär 5.5× Shenzhen och Taiwan ungefär 1.5×.***

årlig full kostnad per QC- eller hanteringsarbetare ($10k; punkt = bas, stapel = låg–hög)



Källa: Shenzhens officiella löneriktlinjer, Taiwans arbetsmarknadsdepartement, JETRO, BLS; författarens uppskattningar.

Tre korrigeringar sätter grunden. För det första höjde Shenzhen sitt pensionsavgiftsgolv från RMB 4,775 till 5,521 per månad i juli 2026 (Shenzhen Bendibao) single source, så en Shenzhen-inspektör 2026 kostar mer troligt RMB 80–90k. För det andra är en post inte en person: två 12-timmarsskift behöver cirka 2.0 personer (sex dagar i veckan) till 2.4 (hela året) i Pärlflodsdeltat, cirka 2.7–3.1 i Sydostasien under övertidstak, och 4.2–4.5 för 24×7 i USA och Taiwan estimate. AI-granskning låter oftast en granskare täcka två till fyra linjer, så den realistiska besparingen är **0.5–1.5 heltidspersoner per linje, inte en person per linje per skift**. För det tredje jagar köpare i Taiwan, Vietnam och Malaysia inte bara kostnad: i JETRO:s undersökning sade 61.1% av företagen i Taiwan, 48.2% i Vietnam och 44.9% i Malaysia att rekrytering blivit svårare de senaste två åren, mot 18.6% i Fastlandskina (JETRO 2025) single source.

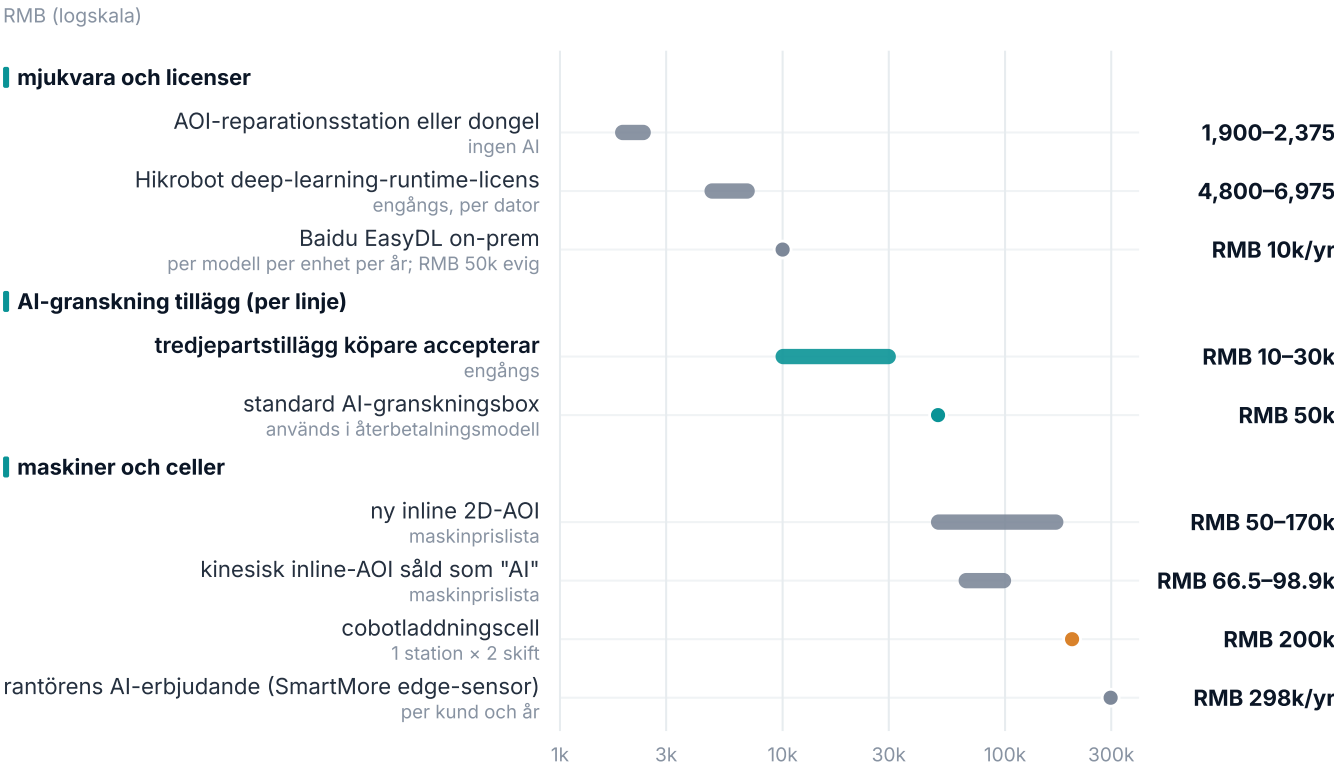
Utrustningspriser: fyra grunder som inte bör medelvärdesbildas

AOI-priser skiljer sig främst för att de mäter olika saker. Onlinelistningar (november 2026, nytt, begagnat och hyra blandat) ligger på cirka RMB 17–80k för offline eller stationär 2D, RMB 50–170k för inline-2D och RMB 80–475k för inline-3D, upp till RMB 700k för koreansk 3D (1688 inline-AOI, 1688 offline-AOI, 1688 3D-AOI) single source, list prices. En upphandling i november 2025 satte ett tak på en 3D-AOI plus en 3D SPI vid RMB 800k (upphandlingsmeddelande). Sinic-Teks prospekt antyder RMB 428–512k per 3D-AOI 2019–2022, cirka RMB 340–470k enligt dess 2025-rapport (Sinic-Tek prospekt, Sinic-Tek 2025 årsredovisning) primary + estimate. En utländsk leverantör uppskattar $3,000–20,000 för stationär 2D, $30–70k inline-2D, $50–90k instegs-3D och $100–200k+ high-end 3D, med samma inline-maskin på $30–50k i Kina (Averroes) vendor estimate från en säljare av konkurrerande AI-mjukvara.

AI-granskning själv har inget offentligt RMB-pris; leverantörer anger pris fall för fall efter en kostnadsfri utvärdering på plats (Sohu om Hengtianyi), och sökningar i e-handelslistningar, upphandlingar, leverantörsartiklar och taiwanesisk media gav inget offentligt pris (1688, 51SMT). Ankarna: cirka €27k per linje antytt av Siemens-fallet estimate; $21,924 listpris för en USA-distributörs modul för fjärrgranskningsnätverk (Manncorp) vendor list price; cirka RMB 298k per kund och år för SmartMores edge-AI-sensorer 2025 (36Kr) single source. En inhemsk inline-2D-AOI som säljs som "AI" listas till RMB 66.5k (1688 inline-AOI), och en NG-buffert för manuell omgranskning cirka RMB 50k (1688 AOI-granskning). Så taket för ett tredjeparts-granskningstillägg är cirka RMB 10–30k per linje, eller en liknande årsavgift; över det skulle köparen hellre ersätta maskinen med en inhemsk AOI som har AI inbyggt estimate.

Ankare	Pris (RMB)	Vad det köper	Källa
AOI-reparationsstation eller mjukvarudongel (äldre TRI, Zhenhuaxing)	≈1,900–2,375	Licensnyckel, ingen AI	1688 list price
Generell djupinlärnings-seende-runtime (Hikrobot VisionMaster-dongel)	≈4,800–6,975 engångs per PC	Plattformen; kunden bygger och underhåller modeller	1688 list price
Baidu EasyDL lokal installation	10,000 per modell per enhet per år, eller 50,000 evig	Paketerad inferenstjänst	Baidu AI-dokumentation [prissida från 2022]
Inhemsk inline-AOI som säljs som "AI"	66,500–98,900 per maskin	Maskinersättning	1688 list price
SmartMore edge-AI-sensorer	≈298,000 per kund per år	Leverantörslösning över flera maskiner	36Kr single source

FIG 18 **Prisstegen** ***Från några tusen RMB mjukvarulicens till RMB 200–300k för en robotcell eller ett AI-erbjudande från en leverantör.***



Källa: 1688-listningar, Baidu AI-dokumentation (prislista 2022), 36Kr, en branschintervju (ej verifierad).

En Baidu-liknande årsavgift, med en inferensserver för en till fyra linjer och en till fyra modeller vardera, antyder en prenumeration på RMB 2.5–40k per linje per år, realistiskt cirka RMB 5–10k; vid 20% av en evig licens når prenumeration och köp break-even år fem, en struktur värd att kopiera estimate. Köpare är vana vid nära-noll mjukvaruavgifter: hela Sinic-Teks "övriga" intäkter (reservdelar, tjänst, hyra, tillbehör) var 7.65% av 2025 års försäljning, cirka RMB 2,300–3,100 per installerad maskin per år (Sinic-Tek 2025 årsredovisning) primary + estimate. En AI-prenumeration på RMB 5–10k per linje per år är två till fem gånger vad köpare nu spenderar på AOI-mjukvara, så den måste motiveras av de granskningstimmar den frigör (cirka RMB 75–90k per heltidsperson per år), inte genom jämförelse med befintliga mjukvaruavgifter.

För robotarmar föll Kinas genomsnittliga cobot-pris till cirka RMB 58,600 2025 (GGII) estimate, och Dobots sexaxliga genomsnitt föll 19.1% det året när billigare modeller tog andelar (Huxiu) verified, trots att listpriserna för samma modeller steg 5–15% 2026 (Sina Finance). I USA listades en FANUC CRX-10iA till $51,434 (Vention) och sammanställningar sätter en UR7e på $30–48k (cobotcost) estimate. Armen är bara 40–50% av en USA-cell (Standard Bots); i Kina går kameror, gripare, säkerhet och ingenjörsarbete lätt till två till fyra gånger en RMB 40k-arm, så budgetera absolut ingenjörskostnad snarare än en multipel estimate.

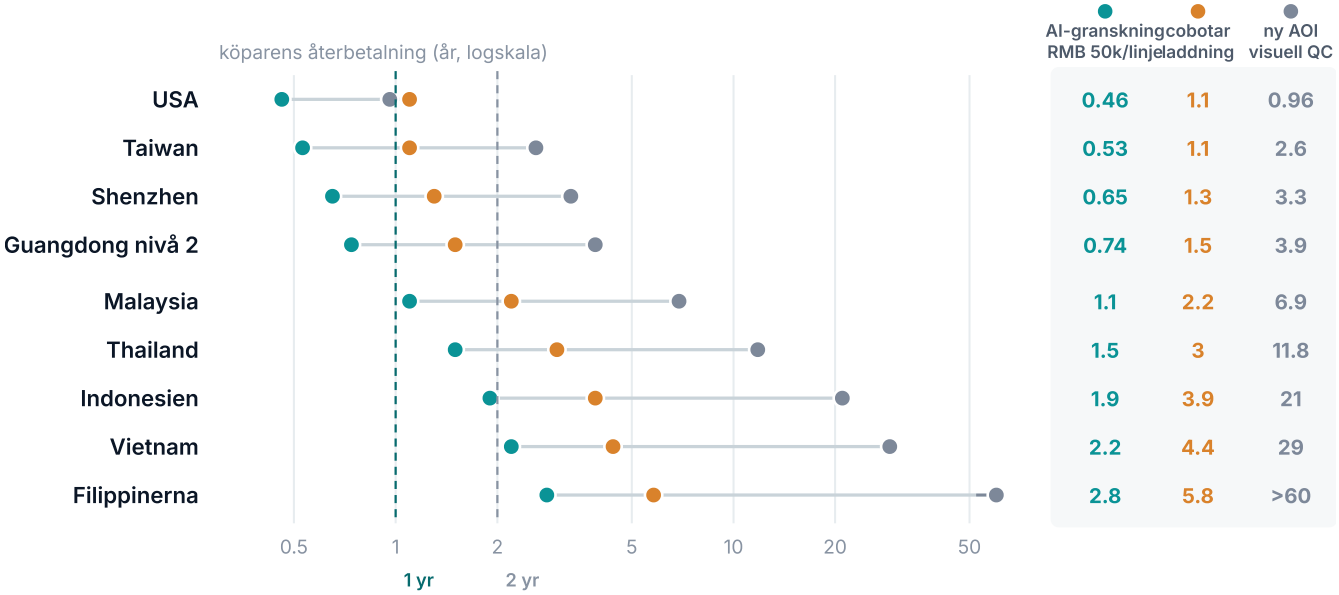
Återbetalningstid: granskning på månader, betjäning på år, utplockning aldrig

Enkel återbetalningstid = investering ÷ (sparade personer × belastad arbetskostnad − årlig driftskostnad), för en tvåskiftslinje utan subventioner eller tullar. Taiwan-utrustning prissätts till 1.3 gånger Kina, Vietnam, Malaysia och Thailand till 1.15 gånger, Indonesien och Filippinerna till 1.2 gånger, och USA till utlandspriser estimate.

Scenario	Investering (Kina / USA)	Sparade personer	Årlig driftskostnad
(a) AI-granskningsbox eller AI-kamerabox	RMB 50k / $30k per linje	1.0 per linje	15%
(b) Ny inline-2D-AOI som ersätter manuella visuella kontroller (0.5 granskare per skift kvar)	RMB 200k / $60k	1.0	12%
(c) Cobot-lastningscell (en station × två skift)	RMB 200k / $140k	2.0	8%
(d) Integrerad detektera-och-plocka-cell (inspektion 1.0 + plockning 0.07)	RMB 300k / $200k	1.07	8%
(e) En robotarm som bara läggs till för att plocka ut NG-kort	RMB 150k / $100k	0.07	8%

Region	(a) AI-granskning	(b) Ny AOI	(c) Cobot-cell	(d) Detektera-och-plocka	(e) Utplockningsarm
Shenzhen	0.65 år (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	Aldrig
Guangdong tier-2	0.74	3.9	1.5	5.3	Aldrig
Taiwan	0.53	2.6	1.1	3.6	Aldrig
Malaysia	1.1	6.9	2.2	9.2	Aldrig
Thailand	1.5	11.8	3.0	15	Aldrig
Indonesien	1.9	21	3.9	25	Aldrig
Vietnam	2.2 (1.6–4.5)	29	4.4	34	Aldrig
Filippinerna	2.8	>60 (driftskostnad ≈ besparing)	5.8	≈100	Aldrig
USA	0.46	**0.96**	**1.1**	3.4	Aldrig

FIG 19 **Återbetalning per region*AI-granskning betalar sig på månader i USA, Taiwan och Shenzhen och på 1–3 år i Sydostasien; cobotmatning på år.***



en robotarm enbart för att plocka ut NG-kort betalar sig aldrig, någonstans
vid first-pass yield över 90% tar utplockning bara 0.003–0.14 FTE per skift

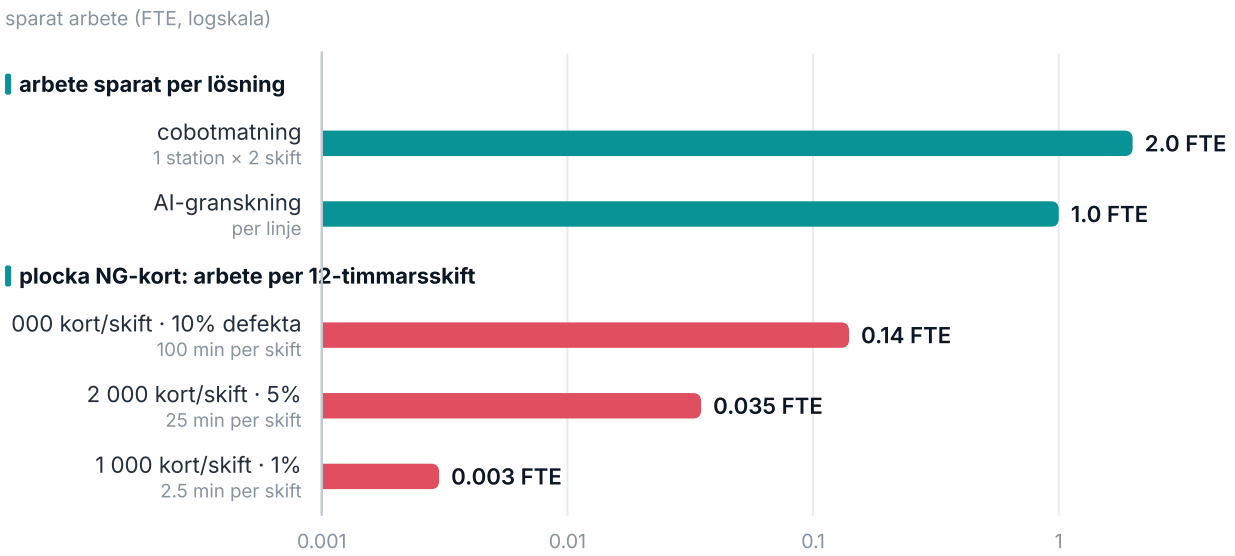
Källa: Författarens modell: RMB 50k AI-granskningsbox per linje; cobotmatningscell på RMB 200k eller $140k, två skift.

Alla värden är uppskattningar från indata ovan. Känsligheter: halvera de sparade personerna och varje återbetalningstid ungefär fördubblas; kör ett skift och (b)–(d) fördubblas; 100% förstaårsavskrivning i USA förkortar återbetalningstiden med cirka 20%, medan tullar på kinesisk, japansk eller koreansk utrustning förlänger den med cirka 10–40%; en 15% lokal subvention i Kina förkortar den med cirka 15%. Jag fann ingen oberoende undersökning av köpares återbetalningströsklar; leverantörer säger att AOI betalar sig "typiskt på 9–18 månader" (Averroes) och Siemens-fallet hävdar åtta månader vendor claims. Jag använder ett och två år som screeningslinjer. Modellen utelämnar kvalitetsvinster (garanti, returer) och personalomsättnings- och utbildningskostnader, så den är konservativ.

Den kontraintuitiva delen av "detektera och plocka" är att arbetskraften ligger i granskningen, inte i att ta de få verkliga NG-korten av linjen. Vid 15 sekunders hantering per NG-kort estimate:

Kort per skift	Defektgrad	Plocktid per 12-timmarsskift	Personer
1,000	1%	2.5 min	0.003
2,000	5%	25 min	0.035
4,000	10%	100 min	0.14

FIG 20 **Varför utplockning av dåliga kort inte betalar sig** ***Över 90% first-pass yield tar utplockning bara 0.003–0.14 av en person per skift.***



Källa: Författarens uppskattningar.

Grundfallet (2,000 kort per skift, 5% defekta, två skift) är 0.07 av en person, värt cirka RMB 6,000 per år i Shenzhen, mindre än armens cirka RMB 12,000 årliga driftskostnad, så det betalar sig aldrig någonstans; även värsta fallet (4,000 kort per skift, 10% defekta) tar cirka 13 år i Shenzhen och 9 i USA estimate. Om oron är att stoppa linjen för att plocka, löser en vanlig NG-avledare eller bufferttransportör det utan en arm. Folk i branschen säger "över 90% förstapassutbyte, stoppa linjen och ta kortet för hand", och **det är rätt för plockning, men inte för inspektion**. Produktutformningen på en SMT-linje är därför "inspektera plus märka eller avleda", med robotarmar reserverade för platser där varje del hanteras (testfixturlastning, tråg- och magasinhantering, förpackning), där inspektion blir ett nästan kostnadsfritt tillägg och scenario (c) gäller.

Subventioner, skatt och tullar flyttar återbetalningstiden med 15–25%, men inte rangordningen

Fastlandskina: stöd sänker mestadels finansieringskostnaden; enskilda maskiner kvalificerar sällan direkt. En nationell lånesubvention betalar 1.5 räntepunkter i upp till två år i 14 prioriterade kedjor (Yicai) verified, cirka RMB 9,000 på en RMB 300k-maskin; Guangdongs uppgraderingsbelöning på upp till 20% kräver ett registrerat projekt med minst RMB 7 miljoner ny utrustning och utesluter Shenzhen (Guangdong-regler), och Dongguan betalade faktiskt 4.42% i juli 2026 (Dongguan-tilldelning) single source, primary. En AI-inspektionsenhet godkänns bara som utrustning "med stödjande mjukvara" inne i köparens registrerade projekt, så grundfallet antar ingen subvention.

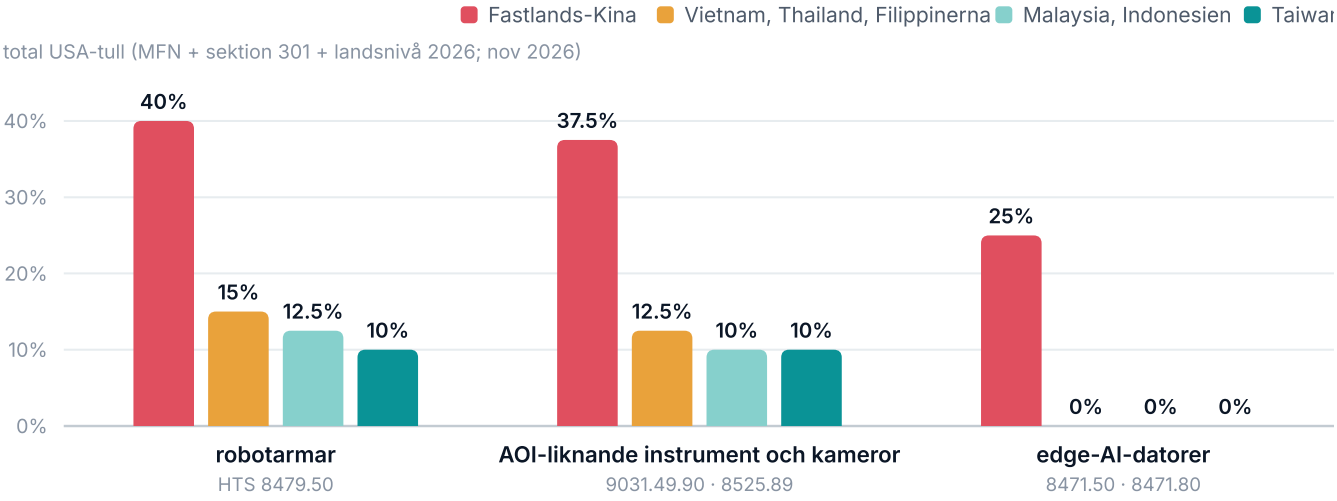
Taiwan: incitamenten går till köparen. Enligt artikel 10-1 i lagen om industriell innovation, ändrad i maj 2025 för att inkludera AI, kan en köpare kreditera 5% av 2025–2029 års utgifter på nya smarta maskiner eller AI (inklusive CNN/DNN-baserad inspektion) mot det årets inkomstskatt, eller 3% per år i tre år (lagen, kreditregler) verified; leverantörssidans bidrag är stängda för fastlandskapital (MOEA) single source.

Sydostasien: bara Malaysia och Thailand belönar att köpa automation. Malaysias 200% automationskapitalavdrag på de första RM10 miljoner är värt cirka 24% av investeringen för en lönsam köpare (MIDA, PwC Malaysia) verified; Thailands BOI ger ett treårigt skatteundantag på upp till 50–100% av en uppgraderingsinvestering, med det högre taket när thailändska automationsleverantörer är inblandade (BOI) single source, primary, så att gå in i Thailand fungerar bäst med en lokal integratör. Vietnam, Indonesien och Filippinerna belönar att bygga fabriker snarare än att köpa automation (Vietnam Law Magazine, PwC Philippines).

USA: avskrivning förkortar återbetalningstiden; tullar förlänger den. Tillgångar köpta efter den 19 januari 2025 får permanent 100% förstaårsavskrivning, cirka 21% federal skattesköld för en lönsam köpare (IRS Notice 2026-11) single source, primary ; Section 232-utredningen av robotar och industrimaskiner avslutades den 28 oktober 2026 med förhandlingar snarare än nya tullar, att omprövas inom 180 dagar (White House, Mayer Brown) verified . Slutmontering i Sydostasien sänker påslaget på armar och AOI-liknande instrument från cirka 37.5–40% till 10–15% (kapitel 1).

Region	Huvudinstrument	Effekt på köparens återbetalningstid estimate
Fastlandskina	1.5-punkts lånesubvention i upp till två år; provinsiella och kommunala uppgraderingsbelöningar (Guangdong behöver ≥RMB 7M registrerade projekt; Dongguan betalade faktiskt 4.42%); Shenzhens 2026–2029-åtgärder (satser inte offentliga)	Cirka 0–5% för en enskild maskin; 4–15% inne i ett stort projekt
Taiwan	Lagen §10-1 kredit, 5% eller 3% × 3 år	Cirka 5% kortare, om köparen betalar tillräcklig skatt
Malaysia	200% automationskapitalavdrag (första RM10M)	Cirka 15–25% kortare för lönsamma köpare
Thailand	BOI treårigt skatteundantag upp till 50–100% av investeringen	Cirka 15–25% kortare för lönsamma köpare utanför en skattesemester
Vietnam, Indonesien, Filippinerna	Inkomstskatteincitament för att bygga fabriker; extra 20% utrustningsavskrivning i Filippinerna	Nästan inget för enskilda automationsköp
USA	100% förstaårsavskrivning; ≈37.5–40% kombinerad tull på fastlandstillverkad utrustning	Avskrivning ≈20% kortare; tullar ≈10–40% längre

FIG 21 **USA-tull på robotar, AOI och edge-datorer** ***Armar tillverkade i fastlands-Kina betalar cirka 40%; skickade från Sydostasien eller Taiwan, 10–15%.***



Källa: USITC Harmonized Tariff Schedule kapitel 99; Federal Register; USTR (november 2026).

Per region: i Fastlandskina betalar en AI-granskningsbox sig på månader och en cobot-cell på 1.3–1.5 år, medan en ny AOI som bara köps för att sänka arbetskraft tar 3–4 år och integrerad detektera-och-plocka 4.5–5.3, båda bortom en tvåårsscreening. Taiwan gör något bättre än Shenzhen (0.53 år för granskning, 1.1 för en cobot-cell), och rekryteringen där är svårare. I Malaysia och Thailand tar en AI-box 1–1.5 år och en cobot-cell 2–3, cirka 15–25% snabbare med lokala skatteincitament för lönsamma köpare. I Indonesien, Vietnam och Filippinerna betalar bara en mycket

billig AI-box sig inom 2–3 år; efterfrågan där kommer mer från kunders kvalitetskrav och arbetskraftsbrist. I USA betalar varje alternativ utom en ren utplockningsarm sig på 0.2–3.4 år, med tullar, Section 232-omprövningen och NRTL-certifiering som huvudvariabler.

05 Marknad och konkurrens
Marknad, konkurrens och hur köpare betalar

Marknadsstorlek: publicerade siffror skiljer sig med multiplar; bara en bottom-up-räkning är användbar

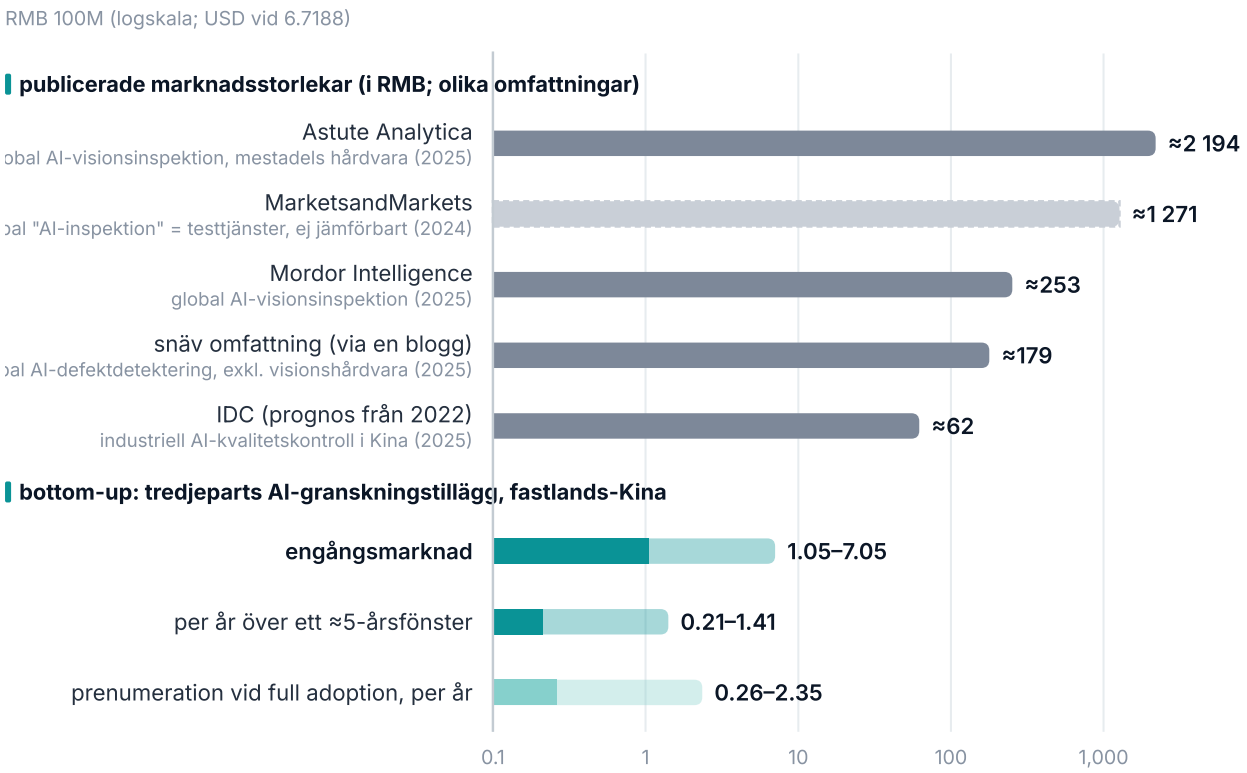
Varje marknadssiffra måste läsas med sitt omfång. Kinas maskinseende-marknad var RMB 18.15 miljarder 2024 (Zhiyan) verified, med 3C-elektronik som största köpare; globala uppskattningar för maskinseende skiljer sig cirka trefaldigt och för AOI cirka 6.5-faldigt (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Bottom-up från intäkterna hos Koh Young, TRI, Jutze, Sinic-Tek och ViTrox ligger global SMT-inspektionsutrustning i storleksordningen $1–2 miljarder estimate.

Det finns ingen fristående marknad för "AI-mjukvara för visuell inspektion". De publicerade siffrorna skiljer sig med mer än tio gånger och alla inkluderar hårdvara eller tjänster:

Omfång	Utgivare	Siffra
Global "AI industriell defektdetektering" (snäv, exklusive seende-hårdvara)	Onamngivet företag, via en iFactory-blogg	$2.66 miljarder 2025, 8.6% CAGR till 2035, elektronik 34%
Global "AI visuell inspektion"	Mordor Intelligence	$3.76 miljarder 2025, $12.8 miljarder 2031, 22.67% CAGR
Global "AI visuell inspektion" (mestadels hårdvara)	Astute Analytica	≈$32.66 miljarder 2025
Global "AI-inspektion" (test- och certifieringstjänster; inte jämförbar)	MarketsandMarkets	$18.92 miljarder 2024
Kina "industriell AI-kvalitetsinspektion"	IDC, prognos gjord 2022	$958 miljoner 2025 (≈RMB 6.2 miljarder), 91.5% inom kommunikation, elektronik och två andra branscher

Källor: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) **var och en single source, mestadels sammanfattningar av betalrapporter**.

FIG 22 **Marknadsstorlekar skiljer sig mer än tiofaldigt** ***Bara den snäva omfattningen och en bottom-up-räkning är användbara för beslut.***



Källa: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; författarens bottom-up-uppskattning.

Bottom-up finns det bara en offentlig räkning av SMT-linjer: New Venture Research uppskattade cirka 73,231 linjer världen över 2020, 47% av dem i Kina (19% år 2000), cirka 70% high-end och 30% low-end (EMSNow, NVR SMT2020-sammanfattning) single source, 2020 ; NVR tillhör nu IPC och nyare data är betald (NVR). Att växa med 3–5% per år med en 40–47% Kinaandel ger cirka 87,000–98,000 linjer världen över 2026 och 35,000–46,000 i Fastlandskina (bas 40,000) estimate .

Indata	Låg	Bas	Hög	Grund
Fastlandskinas SMT-linjer (2026)	35,000	40,000	46,000	NVR 2020-bas × tillväxt × andel
Andel utan AI-granskning	75%	80%	85%	Eftermonteringsfönster, nedan
Andel en tredje part kan nå (inte intern, inte bunden till OEM-AI, kan betala)	40%	50%	60%	Antagande: ≈30% low-end-linjer plus vissa high-end-linjer med blandade märken, exklusive DBG-liknande interna byggen
Adresserbara linjer	**10,500**	**16,000**	**23,500**	Produkt
Engångspris per linje	RMB 10k	RMB 20k	RMB 30k	Prisstege och tak (kapitel 4)
Engångs-adresserbar marknad (installerad bas)	**RMB 105M**	**RMB 320M**	**RMB 705M**	Produkt
Per år över ett ≈5-årsfönster	RMB 21M	RMB 64M	RMB 141M	Fönstret krymper 8–12% per år
Prenumeration (RMB 2.5k / 5k / 10k per linje per år, full adoption)	RMB 26M per år	RMB 80M per år	RMB 235M per år	Baidu-liknande årsavgiftsankare

FIG 23 **Bottom-up: 40 000 linjer → 16 000 servicebara → ≈RMB 320M engångs *Tredjeparts-AI-granskningsnischen i fastlands-Kina är verklig, men liten.***



en nykomling som tar 2–5% av servicebara linjer på 3 år: ≈320–800 linjer, ≈RMB 2–5M per år
som prenumeration (RMB 5k per linje och år): RMB 1.6–4M återkommande, bara 0.3–2.3% av IDC:s prognos för industriell AI-QC i Kina

Källa: NVR (bas 2020); författarens uppskattningar.

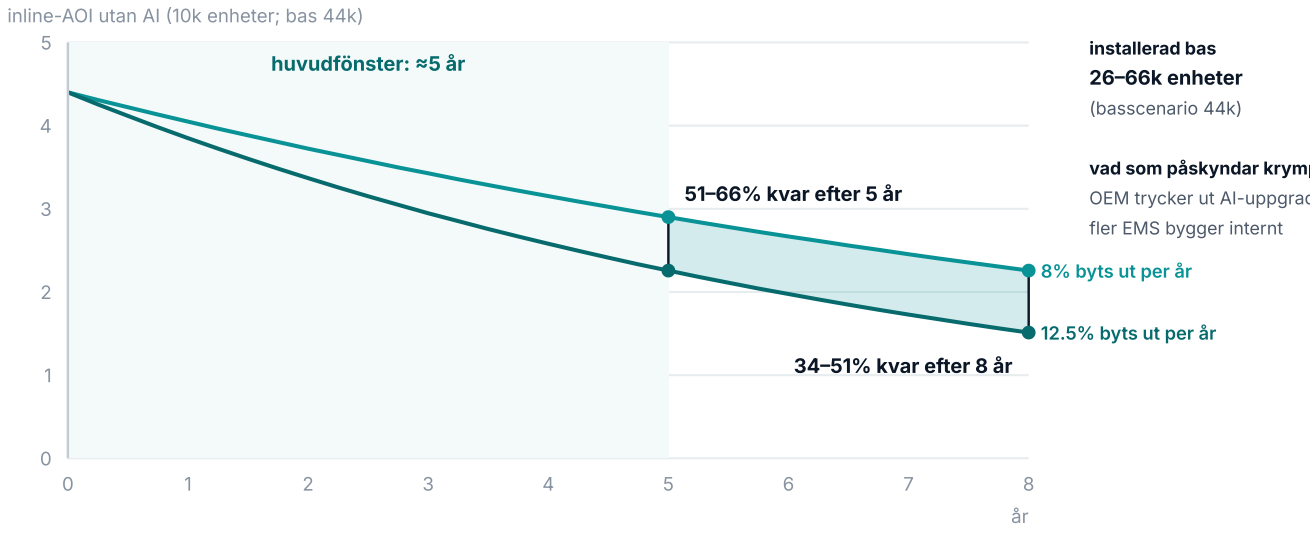
Alla dessa är uppskattningar; varken andelen eller priset har mätts offentligt. Nischen är värd bara 0.3–2.3% per år av IDC:s prognos för all Kinas industriella AI-kvalitetsinspektion 2025 (cirka RMB 6.2 miljarder). En nykomling som vinner 2–5% av basens adresserbara linjer på tre år skulle ha cirka 320–800 linjer: RMB 6.4–16 miljoner totalt vid RMB 20,000 per linje, cirka RMB 2–5 miljoner per år, eller RMB 1.6–4 miljoner per år i återkommande intäkter vid RMB 5,000 per linje estimate. Det räcker för en affärsenhet inne i ett företag, inte för ett fristående företag i venture-skala. Varje 1,000 installerade linjer i Taiwan, Sydostasien eller USA är värd ungefär RMB 3–15 miljoner engångs till fastlandspriser; priserna är högre i Taiwan och USA, men köpare där använder oftare OEM-3D-AOI:er med AI-alternativ estimate.

Eftermonteringsfönstret är cirka fem år

Ingen AOI-tillverkare publicerar maskinlivslängd; EMS-företag skriver av utrustning över 3–10 år (DBG 2025 årsredovisning), och äldre AOI:er förlorar mjukvaruuppdateringar och reservdelar efter end-of-life (I-Connect007) single source. Jag använder 8–12 år (bas 10).

Vid 1.0–1.65 inline-AOI:er per linje har fastlandet cirka 35,000–78,000 inline-AOI:er i tjänst (bas cirka 55,000), i linje med kända leveranser (Circuits Assembly, Sinic-Tek 2025 årsredovisning). Ingen källa mäter hur många som kör AI-granskning; om man antar att inhemska maskiner har erbjudit det sedan cirka 2023 och 40–60% av de senaste leveranserna använder det, kör **cirka 26,000–66,000 inline-AOI:er i Fastlandskina (bas cirka 44,000) inte AI** estimate. Om varje ny maskin nu levereras med AI krymper den poolen med 8–12.5% per år: 51–66% återstår efter fem år och 34–51% efter åtta, snabbare om AOI-tillverkare driver AI-uppgraderingar till nyare maskiner. **Större delen av möjligheten ligger under de kommande fem åren eller så**, med medelstora EMS som kör äldre inhemska 2D-AOI:er, blandade märken och ingen AI-uppgraderingsväg i kärnan.

FIG 24 **Retrofitfönstret är cirka fem år** ***Inline-AOI utan AI krymper med 8–12.5% per år i takt med att de byts ut eller uppgraderas.***

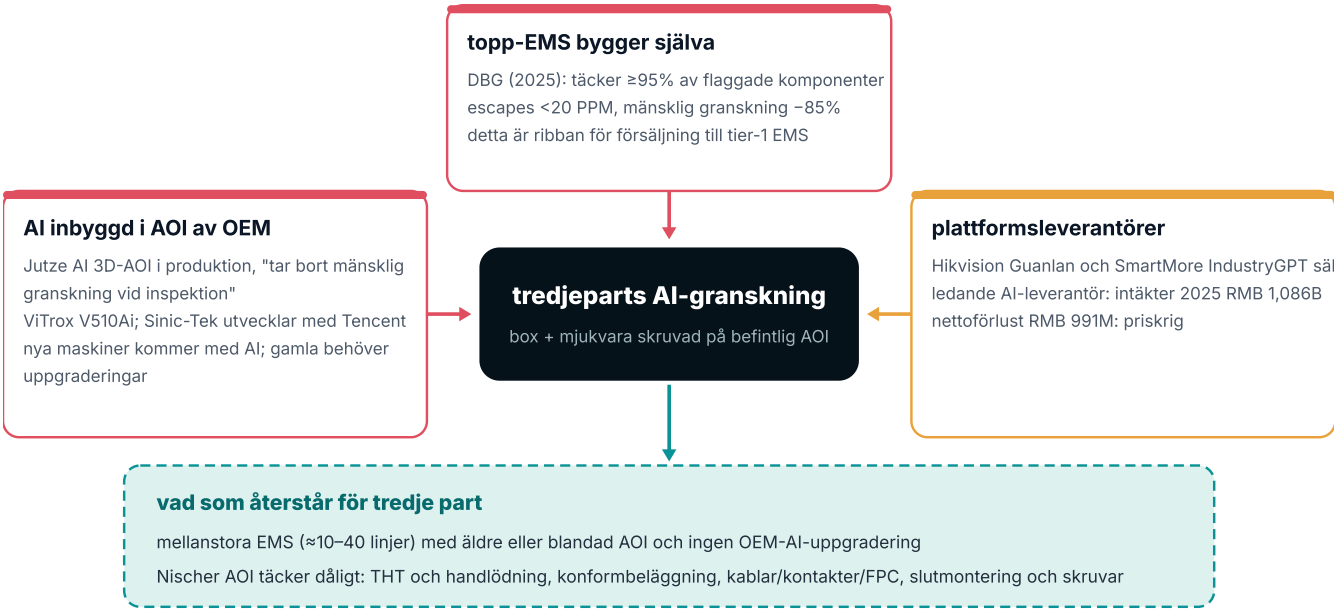


Källa: Författarens uppskattningar; installerad bas och andelar inte oberoende verifierade.

Konkurrenterna är redan starka: granskningskilen är omringad

AI-granskningskilen smalnar från tre håll. AOI- och seende-ledare bygger in AI i nya plattformar: Jutze har "fullt infört AI-algoritmer" (Jutze investor relations, juni 2026), OPT:s industriella AI-intäkter växte 96.77% 2025 (OPT 2025 årsredovisning) single source, primary, och ViTrox hävdar upp till 95% noggrannhet för autonom granskning (ViTrox) vendor claim. Stora tillverkare bygger eget (iThome). Och riskkapitalstödda AI-företag säljer i skala: SmartMores edge-sensorkunder växte från 65 till 251 på två år (36Kr) single source, Hikvision säljer sin Guanlan-modell för inspektion (Xinhua), och "öga, hjärna och hand"-integration erbjuds redan av SmartMore, Hikrobot och Mech-Mind (36Kr, Securities Star).

FIG 25 **Granskningskilen är omgärdad på tre sidor** ***Utrymme för tredje part finns i mellanstora EMS och i inspektionsnischer som AOI täcker dåligt.***



Källa: Jutze, Sinic-Tek och DBG:s årsredovisningar; Hikvision och SmartMore publikt material; sammanställt av författaren.

Utrymmet som lämnas åt en nykomling är smalt: medelstora EMS med äldre 2D/3D-AOI:er och ingen AI-uppgradering från tillverkaren; linjer som blandar flera AOI-märken (en eftermonteringsleverantör stöder redan Omron, VI, Koh Young och Mirtec (Sohu om Hengtianyi)); nischer som SMT-AOI:er täcker dåligt, som hålmonterade och handlödda fogar, konform beläggning, kablar, kontakter och FPC:er, och slutligt utseende och skruvar; och små fabriker som inte har råd med en RMB 230–280k-AOI.

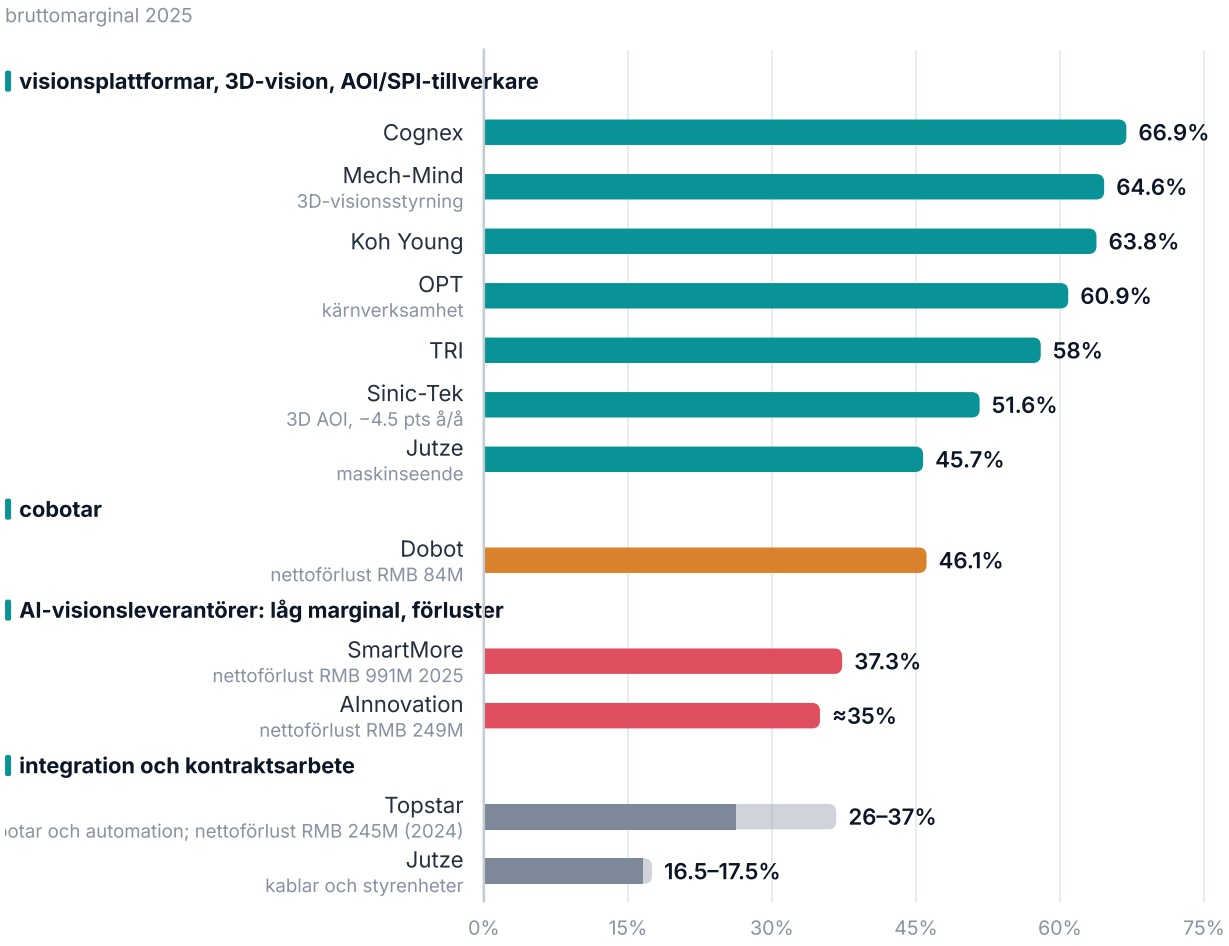
Det närmaste börsnoterade jämförelsebolaget ger en varning. Weiyi Zhizao, som säljer AI-inspektionsutrustning och robotar, såg 3C-intäkterna falla 79.9% år över år i januari–maj 2026, erkände "intensifierad konkurrens och prispress", sköt upp sin Hongkong-notering den 25 oktober 2026, och redovisade 15 avslutade projekt före färdigställande, varav 12 förkroppsligade robotprojekt (Weiyi Zhizao prospekt, uppskjutandemeddelande) single source, primary. AInnovation avslutade 2025 med 29.8% av sina fordringar avsatta (AInnovation 2025 årsredovisning). 3C-inspektionskilen är trång, och prissättningskraften per linje är svag.

Marginalen ligger i produktifierad hårdvara

Position i kedjan	Representativa företag och 2025 bruttomarginal	Vinst och inkassering
Seende-plattformar	Cognex 66.9%	Lönsam; avyttrar ≈$22M lågmarginalintäkter
AOI/SPI-utrustning	Koh Young 63.8%; TRI 58%; OPT kärnverksamhet 60.85%; Sinic-Tek 3D-AOI 51.62% (−4.53 poäng YoY); Jutze maskinseende 45.74%	Lönsamma; fordringsdagar 139 (Jutze) till 247–274 (OPT)
3D-seende-vägledning	Mech-Mind 64.6%	—
Leverantörer av AI-seende-lösningar	SmartMore 37.3%; AInnovation ≈35%	Förlustbringande: SmartMore nettoförlust RMB 991M 2025, AInnovation RMB 249M
Cobotar	Dobot 46.1%	Dobot nettoförlust RMB 84M; UR-intäkter från $326M (2022) till ≈$232M (2025, prognos)
Systemintegration	Topstar robotik och automation 26.2–36.7%	Nettoförlust RMB 245M (2024); fordringsdagar 180–240
Kontraktstillverkning av hårdvara	Jutze kablar och styrenheter 16.5–17.5%	—

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

FIG 26 **Marginalen sitter i produktiserad hårdvara** ***Visionsplattformar och utrustningstillverkare tjänar 46–67% bruttomarginal; AI-leverantörer 35–37% och går med förlust.***



Källa: Årsredovisningar eller prospekt 2025.

En nykomling behåller sin marginal bara genom att göra inspektion till en produkt, utrustning eller mjukvara; projekt-för-projekt-justering pressar marginalerna till cirka 35% med ihållande förluster. Mekaniken verkar också emot små företag: intäkter redovisas bara efter acceptans på plats, förskotten är små (Jutze, OPT), piloter är oftast gratis (A3/Mindtrace), och varumärkesägare specificerar ofta inspektionsutrustningen på sina kontraktstillverkares linjer.

Hur köpare betalar: 30% vid signering, 60% efter acceptans, 10% hållna i ett år

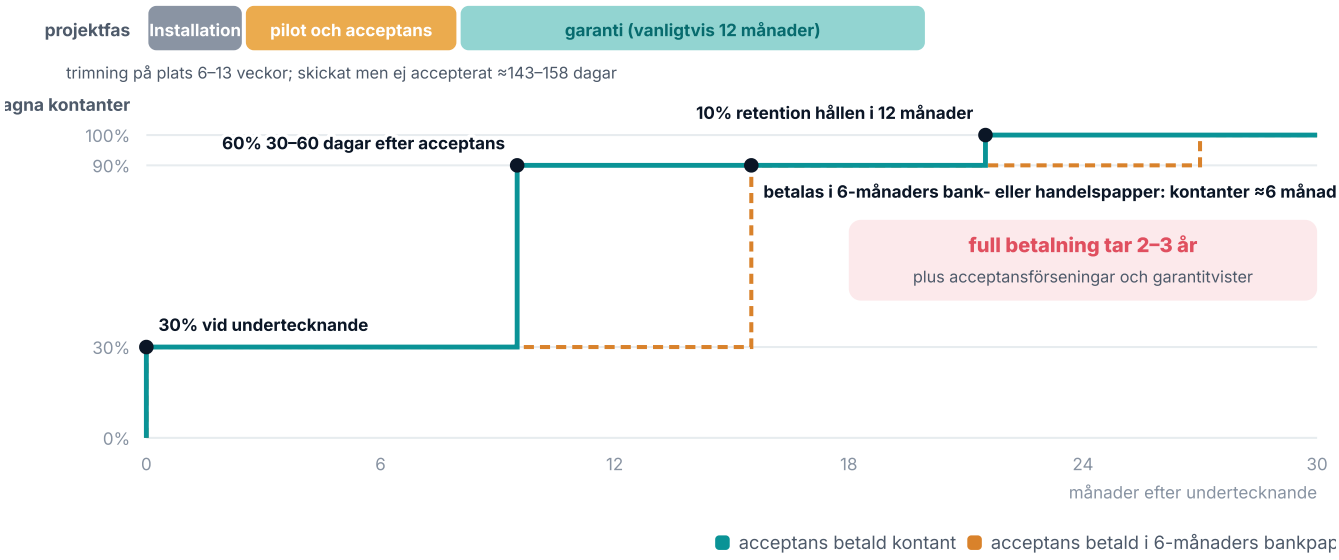
De betalningsvillkor fastlandets elektronikfabriker använder för inspektionsutrustning beskrivs i 13 namngivna köparkontrakt som redovisas i Sinic-Teks noteringsförfrågningssvar (Sinic-Tek förfrågningssvar 8-1-1) single source: en primär inlaga, flera oberoende köpare, 2020–2022 :

Köpare (år)	Signering	Leverans	Acceptans	Retention eller saldo
En Hikvision-enhet (2021)	30% inom 15 dagar	—	65% inom 30 dagar efter slutlig acceptans	5% garantirentention betald 12 månader efter acceptans; del betald i 6-månaders+ bankacceptväxlar
Nanchang Huaqin (2020–2021)	30% inom 10 arbetsdagar	—	60% vid 60 dagar efter acceptans	10% efter ett års garanti
Guangdong Yinuo (2021, 3D-AOI)	30% inom 10 arbetsdagar	—	60% inom 30 dagar	10% efter tre månaders felfri drift

Köpare (år)	Signering	Leverans	Acceptans	Retention eller saldo
Xiamen Hongxin, Jingmen Hongyi (2022)	30% inom 10 arbetsdagar	—	60% inom 30 arbetsdagar efter slutlig acceptans, 6-månaders bankväxlar	10% efter ett års garanti
AKM Panyu, Suzhou (2021–2022)	30% inom 7–10 arbetsdagar	—	60% en månad efter acceptans och fakturering, 180-dagars bankväxlar	10% efter 360 dagar
BYD Precision, Shenzhen, Huizhou (2020)	—	30% vid ankomst	70% efter acceptans och fakturering	Allt i 6-månaders kommersiella växlar
Shenzhen H&T (2020)	—	—	90% efter acceptans och fakturering	10% efter sex månaders felfri användning

Andra primära inlagor visar samma struktur: Mech-Minds stora projekt betalar 30% vid signering, 30% före leverans, 30% efter acceptans och 10% garantirentention över 12–15 månader (Mech-Mind prospekt); GKG säger att retention i utrustningsbranschen oftast är 5–10% av kontraktet, betald ett år efter acceptans (GKG förfrågningssvar) single sources, primary. Normen är alltså cirka 30% vid signering (20–30%, noll vid vissa fabriker som BYD), 60–65% efter acceptans (betald 30–60 dagar senare, ofta i sexmånaders bank- eller kommersiella växlar, vilket fördröjer användbar kassa med ytterligare ett halvår), och 5–10% retention i 3–12 månader (12 vanligast) verified across several contracts.

FIG 27 **Hur köpare betalar *30% vid undertecknande, 60% efter acceptans, 10% hålls i ett år: full betalning tar 2–3 år.***



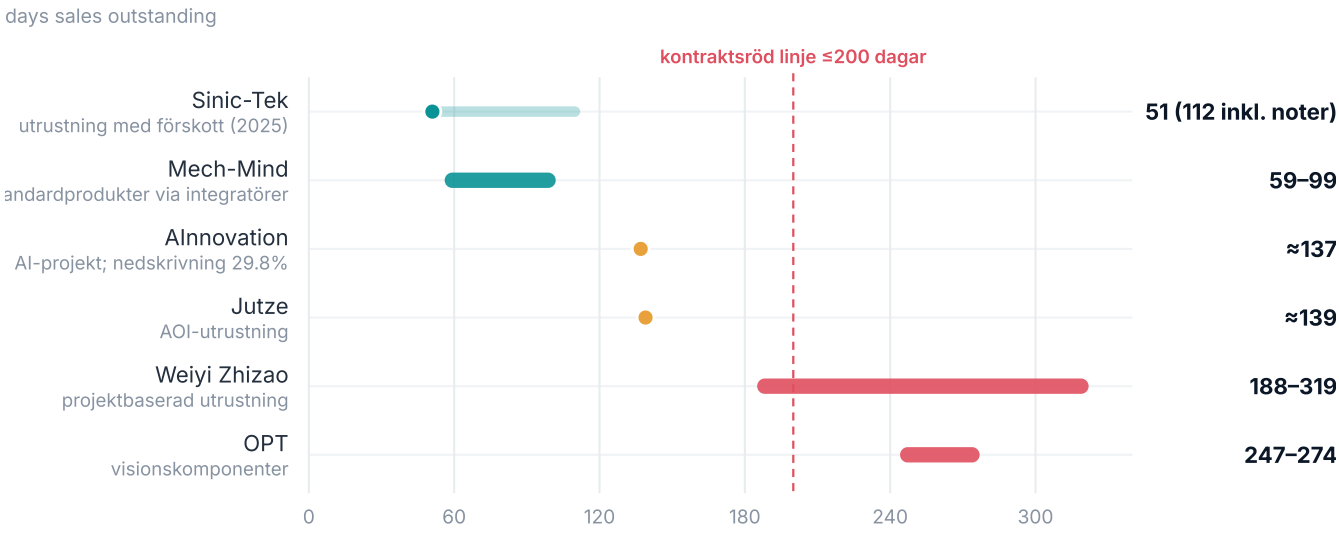
Källa: Sinic-Tek:s svar på förfrågan, Mech-Mind, GKG, JT och andra handlingar; schema på sedvanliga villkor, tidpunkt uppskattad.

Acceptans och försök är långsamma och betalas av säljaren. Sinic-Teks levererade-men-oaccepterade varor tog cirka 143–158 dagar att omsätta, eftersom stora EMS- och varumärkeskunder behöver integrationsprovning av hela linjen och en förlängd provkörning (Sinic-Tek förfrågningssvar 8-1-2); Weiyis AI-projekt tar 4–8 månader från utvärdering till distribution (Weiyi Zhizao prospekt) single sources, primary. Det finns inga konverteringsdata för AI-granskningspiloter; Sinic-Tek konverterade bara cirka 18–19% av sina demomaskiner direkt till försäljning. Jag planerar en av tre till en av fem för gratis piloter och cirka 80% för betalda piloter efter kortlistning estimate.

Företag	Affärsmodell	Dagars försäljning utestående
Sinic-Tek (2025)	Utrustning med förskott	≈51 (≈112 inklusive växlar)
Mech-Mind	Standardprodukter genom integratörer	59–99
Jutze	AOI-utrustning	≈139
AInnovation (2025)	AI-projekt	≈137 vid årsskiftet, med 29.8% nedskrivning
Weiyi Zhizao	Projektbaserad utrustning	188–319
OPT	Seende-komponenter	247–274

Källor: årsredovisningarna och prospekten ovan; fordringar (inklusive kontraktstillgångar) ÷ intäkter × 365 primary + estimate .

FIG 28 **Days sales outstanding** ***Standardprodukter 51–99 dagar; projektbaserad utrustning 188–319 dagar.***



Källa: Företagshandlingar; fordringar (inkl. kontraktstillgångar) ÷ omsättning × 365.

AI-granskningslicenser som säljs genom AOI-tillverkare eller integratörer bör inkassera som Mech-Minds 60–100 dagar; direktprojektförsäljning som Weiyis 190–320 dagar. Sammanställt som en kassatidslinje: gratis pilot en till tre månader, kvalificering hos en stor ODM upp till cirka tolv månader, bygge, leverans och acceptans tre till sju månader; 30% anländer vid beställningen, 60% cirka fyra till nio månader efter den och de sista 10% cirka 16–21 månader efter den. **Från det första pilotprojektet hos ett stort EMS till full betalning kan det ta två till tre år**; medelstora EMS utan formella upphandlingar kan förkorta kvalificeringen till veckor estimate . Acceptanskriterierna i de redovisade kontrakten täcker bara hel-linjeintegration, genomströmning (ett taiwanesiskt projekt krävde 100 enheter per timme) och månader av felfri körning; inget specificerar läckage- eller falsklarmsgrader (Anda förfrågningssvar) single source, primary . Skriftliga acceptanskriterier för AI-granskning måste överenskommas före piloten.

Fyra misslyckandemönster

Dokumenterade misslyckanden faller i fyra mönster: att leverera innan produkten var tillförlitlig (Rethink (The Robot Report)); ett mjukvarulager på en ledares hårdvara som ledarens ekosystem senare ersatte (READY Robotics (The Robot Report)); att få slut på pengar i botten av en cykel (Dextrous Robotics (Robots & Startups)); och modellföretag som slutar i acqui-hires (Covariant (GeekWire)). Lärdomarna: produktifiera smalt i stället för att justera projekt för projekt; håll rörelsekapital för acceptansbaserade intäkter och långa inkasseringar; tillhandahåll lokal service för fabriker som kör 24 timmar; och skala inte innan produkten är bevisad.

Timing: inspektion vinner inget på att vänta; armhårdvara blir billigare

Argumenten för att vänta är svaga. Hårdvara blev inte billigare 2026: robottillverkare höjde priserna 5–15%, vissa produkter 25–50% (Sina Finance) single source, och NVIDIA-kit steg en tredje gång (36Kr). Produktionsklassade VLA:er kommer inte inom fönstret (IFR, Instrumental). Samtidigt tar varje AOI-ersättning bort ett eftermonteringsmål och varje OEM-AI-distribution låser in en linjes data och process; modellkapacitet kommodifieras, och det som är sällsynt är tillgång till fabriker och etiketterad fabriksdata. Lokal support har också fönster: Shenzhens 2026 "AI-pi-onjärstad"-regler, med upp till RMB 2 miljoner för AI-applikationsdemonstrationsprojekt och modellvouchers som täcker 30% av köpen, gäller till **den 30 juni 2027** (Yicai, maj 2026) terms single source.

Det som fortfarande blir billigare snabbt är humanoider och fingerfärdiga händer (händer föll från cirka RMB 50,000 till under 10,000 på två år (Sina Finance)), och instegscobotar när produktmixen rör sig nedåt; det talar för att skjuta upp den förkroppsligade delen, inte inspektionsdelen. **Sammantaget: inspektion nu, som en smal produkt; robot-celler nu, genom partners och integration; förkroppsligad AI senare, med bara en FoU-station idag.**

Förkroppsligad AI

06 Förkroppsligad AI: först ändrar den hur robotar lärs upp, senare hur de rör sig

Det som förändras är perception och programmering; rörelse förblir deterministisk

Varje robotarmsledare marknadsför nu "fysisk AI", men det som har levererats faller i tre lager: inlärd perception utan undervisning (FANUC:s AI Masterless Vision, som levereras i januari 2027 (FANUC); KUKA:s neurala greppval (KUKA)); stora modeller som skriver program och planerar uppgifter (KUKA:s iiQWorks.Copilot, som varnar för att AI-genererad kod bör verifieras i en digital tvilling (KUKA); Yaskawa med Gemini Robotics-ER, där "generativ AI bestämmer vad som ska göras" och roboten bestämmer "hur den ska röra sig" (Yaskawa)); samt beräknings- och dataplattformar (UR:s AI Accelerator med en Jetson Orin (Universal Robots)) vendor claims. Ingen ledare publicerar produktionsframgångsgrader för sina AI-funktioner, och de japanska ledarna samarbetar med Google, SoftBank, Microsoft och andra snarare än att bygga egna frontier-VLA:er (FANUC news index): det inlärda policylagret kommer från modellföretag, de certifierade armarna och styrenheterna från armtillverkarna.

FIG 29 **Embodied-AI-stacken** ***Stora modeller planerar offline, små policies körs i cellen, rörelse och säkerhet förblir deterministiska.***

Lager	vad den gör och hur långt bevisen räcker	tidsskala · vem äger det
Planering L4	**frontmodeller (offline)** skriva eller ändra robotprogram offline, omställningar, förklara fel skriver om ABB RAPID-program till 77–99.7%, utrullade efter simulering styrloopen bara 0.2–0.4 Hz: håller sig utanför realtid	**sekunder till minuter** modelljättar
färdigheter / policy L3	**ACT, diffusionspolicies, små VLA** strukturerade stationer med fixturer 90–99.4% edge-chips: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**åtgärdschunkar 100–300 ms** chiptillverkare · robottillverkare
rörelse-styrning L2	**robotstyrenhet** banor, servo, kraftreglering, kollisionskontroller deterministisk kod; inlärda policies sätter bara mål	**millisekunder (realtid kräver >83 Hz)** robottillverkare
Säkerhet L1	**säkerhets-PLC, skannrar, kraft- och hastighetsgränser** certifierad enligt ISO 10218-1/-2:2025 och lokala regler ingen ML i säkerhetsfunktioner: EU-bedömning av anmält organ från 20 jan 2027	**hård realtid** certifierad robot + integratör

små företag vinner i integration och sensorhårdvara: sätt inlärd perception och små policies i certifierade robotceller, station för station

Källa: Författarens syntes (kapitel 3 och 6); ISO 10218:2025; EU:s maskinförordning.

Hur långt från ostrukturerat, flexibelt arbete i små serier: strukturerade stationer om 1–3 år, flexibel montering 5+

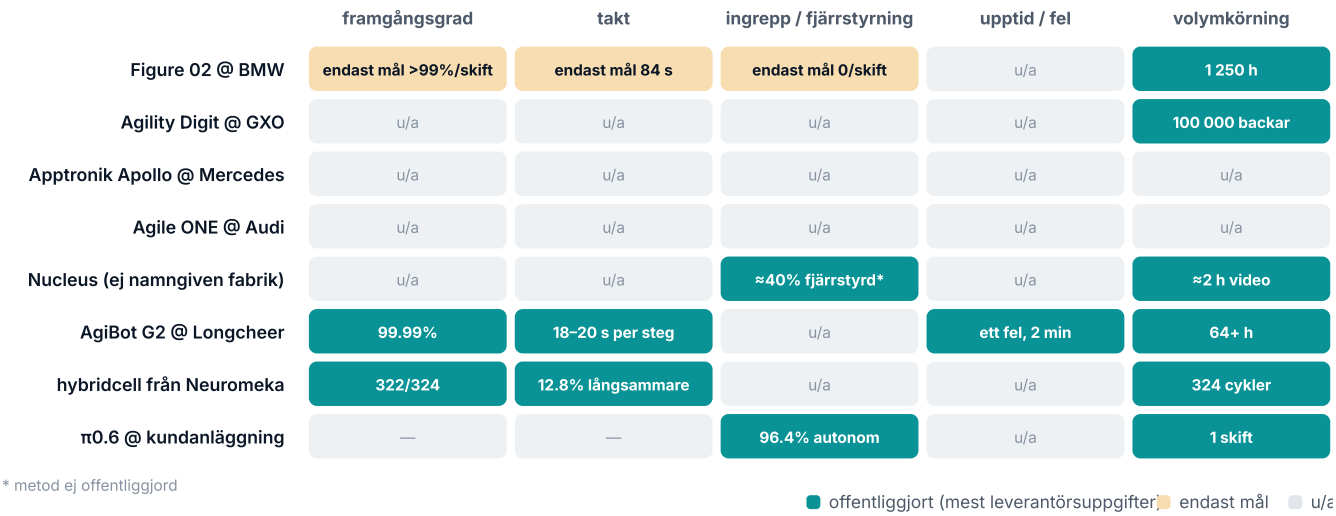
Verkliga fabrikssiffror har börjat dyka upp, men de är leverantörsrapporterade, smala och från strukturerade stationer. AgiBots åtta humanoider på Longcheers sluttestlinje för surfplattor loggade 64,828 uppgifter vid 99.99% över mer än 64 timmar (AgiBot, Yicai Global) vendor and customer, men detta är betjäningsarbete som en konventionell arm med seende kan göra. UBTech säger att dess Walker S2 arbetar på 30–50% av mänsklig effektivitet (Irish Times/FT)

single source. Figure 02:s 11 månader vid BMW Spartanburg producerade volymsiffror (1,250+ timmar, 90,000+ delar), medan dess cykel-, framgångs- och ingreppssiffror är mål, inte resultat (Figure, BMW Group) verified, numera- tors only. Nucleus säger att cirka 40% av dess fabriksvideo är teleopererad (Interesting Engineering) vendor claim, och en Reuters-utredning citerade cirka 20% av en arbetares produktion för vissa kinesiska humanoider (Note- bookcheck) single source. Ingen distribution publicerar framgångsgrad, cykeltid och ingreppsfrekvens tillsammans, och ingen har granskats, ändå är exakt dessa tre vad en återbetalningsberäkning behöver.

Distribution	Framgångsgrad	Takt	Ingrepp eller teleoperation	Drifttid eller felintervall	Volymkörning
Figure 02 @ BMW Spartanburg	Endast mål (>99% per skift)	Endast mål (84 s)	Endast mål (0 per skift)	Ej offentliggjort (underarmens topp felpunkt)	1,250 h
Agility Digit @ GXO	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort	100,000 lådor
Apptronik Apollo @ Mercedes-Benz	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort
Agile ONE @ Audi	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort	Ej offentliggjort
Nucleus (onamngiven fabrik)	Ej offentliggjort	Ej offentliggjort	≈40% teleopererad (metod ej offentliggjord)	Ej offentliggjort	≈2 h video
AgiBot G2 @ Longcheer	99.99% av 64,828 uppgifter	18–20 s per steg	Ej offentliggjort	Ett kommunikationsfel, ≈2 min att återhämta	64+ h
Neuromeka hybridcell	322/324	12.8% långsammare än en person	Ej offentliggjort	Ej offentliggjort	324 cykler
Physical Intelligence π0.6 vid en kundplats	—	—	96.4% autonom över ett skift	Ej offentliggjort	Ett skift

Källor: detta kapitel och kapitel 3 mestadels vendor claims.

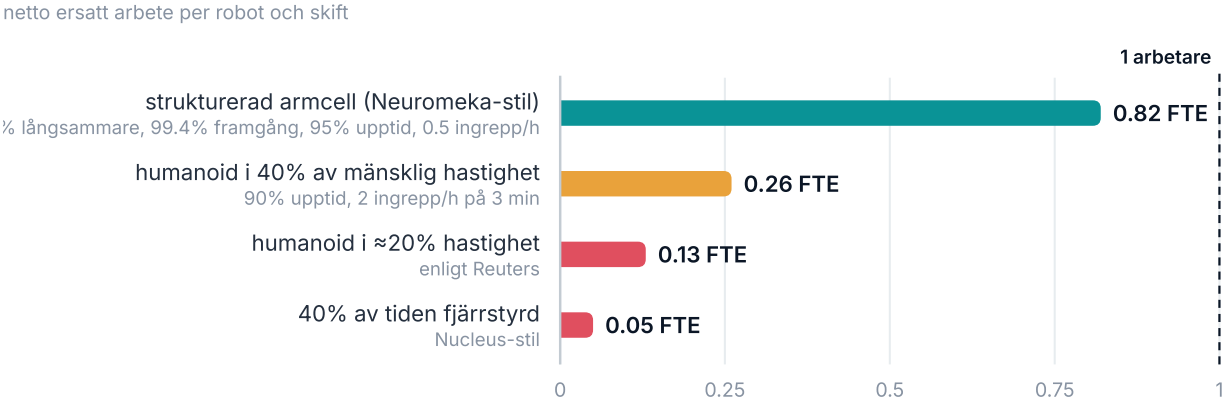
FIG 30 **Vad robotutrullningar i fabrik faktiskt offentliggör** ***De flesta offentliggör bara volym; ingen offentliggör framgångsgrad, takt och interventionsfrekvens tillsammans.***



Källa: Publikt material från BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence med flera.

Teleoperation betyder mer för ekonomin än framgångsgrad. Netto ersatta personer per robot per skift ≈ (mänsklig cykel ÷ robotcykel) × drifttid × framgång, minus ingreppstid och teleoperation estimate, my formula. En Neuromeka-liknande strukturerad station (12.8% långsammare, 99.4% framgång, 95% drifttid, 0.5 ingrepp på 2 minuter i timmen) ersätter cirka 0.82 personer; en humanoid vid 40% av mänsklig hastighet, 90% drifttid och två 3-minutersingrepp i timmen cirka 0.26; vid Reuters 20%, cirka 0.13; och om, som Nucleus, en operatör teleopererar den 40% av tiden, cirka 0.05, nästan inget. Vid cirka RMB 600k per humanoid och RMB 85k per person per år på fastlandet betalar de tre sista sig över decennier; bara den strukturerade armstationen har en realistisk återbetalningstid (kapitel 4) estimate.

FIG 31 **Netto ersatt arbete per robot** ***En strukturerad armcell ersätter cirka 0.82 av en arbetare; en humanoid fjärrstyrd 40% av tiden, cirka 0.05.***



Källa: Författarens formelbaserade uppskattningar.

Bortom strukturerade stationer får generella modeller fortfarande 40–90% på nya uppgifter: Gemini Robotics 2 når 74.2% på lab-pick-and-place, 78.9% på verktygskittning och 44% på att knyta en påse (eesel/DeepMind); på RoboChallenge Table30, 30 uppgifter på riktiga robotar, får π0.5 43.7% och den bästa öppna modellen 50.3% (RoboChallenge, QbitAI).

Externa prognoser skiljer sig, men ingen stöder utbredd användning inom två år. IFR säger att generell fysisk AI är "ännu inte av industrikvalitet" och att de flesta humanoider är "fortfarande prototyper eller piloter" (IFR); Interact Analysis uppskattar att bara cirka 10% av de humanoider som producerades 2025 gick till verklig användning, med den kommersiella vändpunkten "efter 2032" (automation.com); UBS ser ett humanoid-"EV-ögonblick" som osannolikt före 2030 (Yicai Global); kinesiska insiders uppskattar två till tre år till innan användning i fabriksskala (Gasgoo) single sources. Min bedömning estimate: i välresursförsedda elektronik- och bilfabriker med leverantören på plats når **strukturerade stationsuppgifter** (betjäning, OK/NG-sortering, fixerad insättning) ≥99% vid nära mänsklig hastighet om cirka 1–3 år; **kittning av blandade delar** med generella VLA:er om cirka 3–5; **fritt dragna kabelhärvor och flerstegs flexibel montering** om mer än 5. Humanoid-återbetalningstid fungerar inte idag: en Walker S2 vid cirka RMB 600k (10jqka) single source, vid 30–50% av mänsklig effektivitet över två skift, motsvarar cirka 0.6–1.0 arbetare, en enkel återbetalningstid på 7–12 år vid fastlandsarbetskostnader före integration, underhåll och driftstopp estimate.

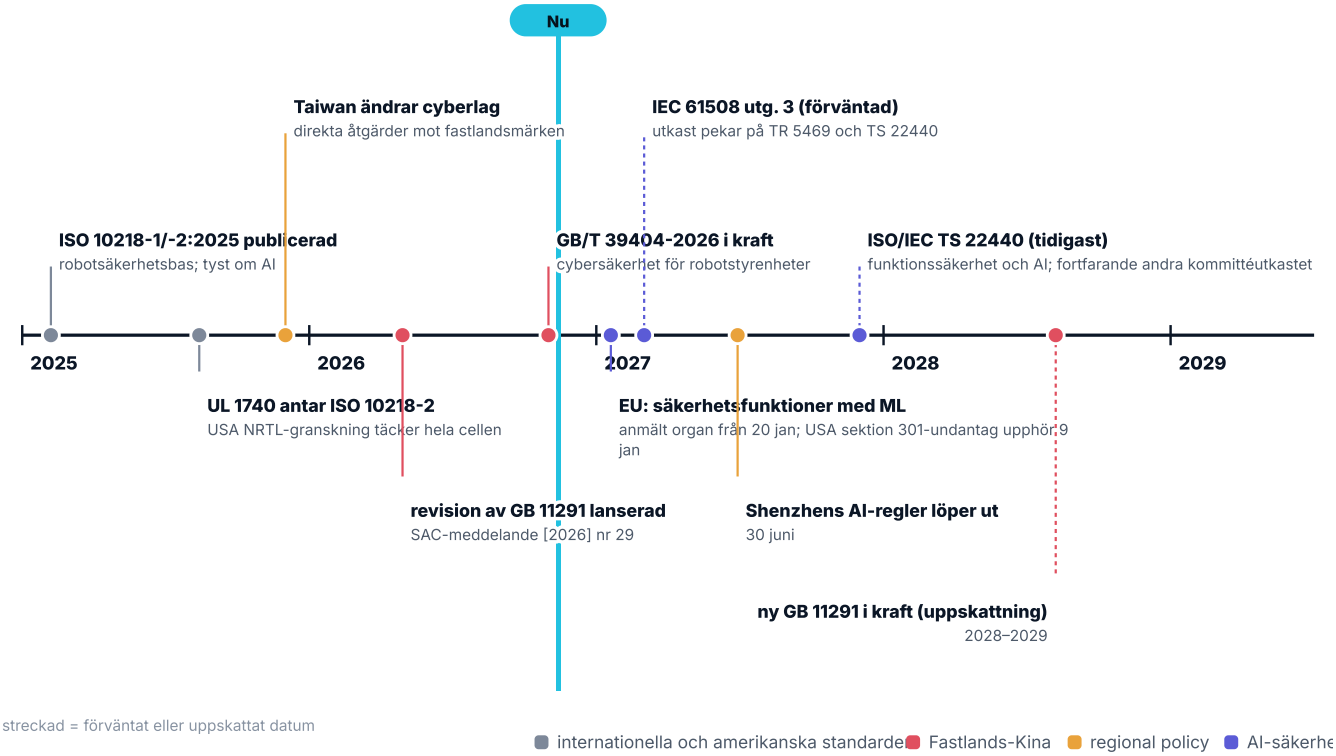
Var det har fastnat

Länk	Bevis 2026
Data	Branschen uppskattar att minst 10 miljoner timmar högkvalitativa verkliga data behövs och att bara cirka 500,000 hade samlats in i början av 2026 (Gasgoo); Reuters rapporterar cirka 500,000 timmar mot cirka 100 miljoner som behövs (Notebookcheck); en nybörjare inom teleoperation behöver cirka 300 försök per användbar rörelse; fabriker delar inte processdata; simulerings-tränade policyer faller kraftigt på riktiga robotar

Länk	Bevis 2026
Tillförlitlighet	Strukturerade stationer 98–99.99% (leverantörspåståenden, fel odefinierat); generella modeller 40–90% på nya uppgifter; osedda kontakter 29–67%
Hastighet	Humanoider på 30–50% av mänsklig hastighet (UBTech), vissa på cirka 20% (Reuters); frontier-agenter cirka 280 s per kvasistatisk uppgift på riktig robot; LLM-styrkretsar på 0.2–0.4 Hz mot cirka 83 Hz som behövs (Anthropic)
Kostnad	Walker S2 i genomsnitt cirka RMB 600k; 7–12 års återbetalningstid estimate
Säkerhetscertifiering	ISO 10218-1/-2:2025 gör funktionssäkerhet mer explicit och absorberar samarbetskrav, men säger inget om AI (The Robot Report); ISO/IEC TS 22440 (funktionssäkerhet och AI) var fortfarande ett andra kommittéutkast i november 2026 (ISS, ISO Open Data); IEC 61508-3-vägledning avråder från AI över SIL 1 (Analog Devices); enligt EU:s maskinförordning behöver säkerhetsfunktioner med maskininlärning bedömning av anmält organ från den 20 januari 2027 (EUR-Lex)
Händer och verktyg	Fingerfärdiga händer hanterar cirka 30 N (≈3 kg) per finger mot cirka 10 kg för en mänsklig hand (Gasgoo); verktygsbyten är fortfarande manuella
Integration	Sensorfelavläsningar, protokollfel och MES-länkar på plats; ingreppsfrekvenser nästan aldrig offentliggjorda

Säkerhetsimplikationen är direkt: standarderna varken förbjuder inlärda policyer eller erkänner dem. ISO 10218-1/-2:2025 säger inget om AI; ISO/IEC TS 22440 (funktionssäkerhet och AI) är fortfarande på kommittéutkast och osannolik före slutet av 2027 eller 2028 estimate; ISO/IEC TR 5469:2024 är bara vägledning, som rekommenderar icke-AI-fallbacks, övervakande begränsningar och redundant röstning; IEC 61508 utgåva 3 väntas i början av 2027 (ISO Open Data, BSI, 61508 Association) verified. Ingen publicerad standard erbjuder ännu en accepterad väg för att certifiera maskininlärningsperception som en säkerhetsfunktion. Regionala regler rör sig också: Kina reviderar GB 11291 (i kraft cirka 2028–2029) (revisionsplan 20262543); UL 1740 har fört in ISO 10218-2 i USA:s NRTL-granskning sedan augusti 2025; EU kräver bedömning av anmält organ för säkerhetsfunktioner med maskininlärning från den 20 januari 2027; och ingen humanoid är certifierad för inhägnadslös fabriksanvändning medan ISO 25785-1 fortfarande är i utkast (i-SCOOP). Den försvarbara arkitekturen: maskininlärning fattar bara processbeslut (“vilken del”, “är detta kort bra”), medan ett certifierat deterministiskt säkerhetslager (typiskt PL d / Kategori 3) och en säkerhets-PLC bestämmer hastighet, kraft och stopp. Behandla varje påstående om “kamera-AI för inhägnadslös säkerhet” som obevisat tills ett PL d- eller SIL 2-certifikat namnger maskininlärningskomponenten.

FIG 32 **Tidslinje för standarder och reglering** ***En certifieringsväg för AI inom säkerhetsfunktioner klarnar inte före 2027–2029.***



Källa: ISO, IEC, UL, EUR-Lex, SAC och nationella regleringsdatabaser.

Vilket lager är möjligheten: modeller tillhör jättarna, edge-boxar till chiptillverkarna, små företag till integration och avkänningshårdvara

Modellagret behöver kapital och data och tillhör jättarna (AgiBot tränade GO-2 på "tiotusentals timmar" (iFeng Tech); de starkaste modellerna är stängda eller molnbaserade). Edge-box-lagret tillhör chiptillverkarna: Jetson Thor är de facto-standard, och på fastlandet konkurrerar D-Robotics RDK-kort (D-Robotics) och Rockchip-samprocessorer (CSDN), mestadels utan publicerade VLA-resultat.

Små företag har två öppningar. En är avkänningshårdvara som direkt köper tillförlitlighet: beröring lyfter Contact-DP:s insättningsframgång från 58.7% till 97.8%, och kraft/moment-sensorer och handledskameror gör liknande arbete; datainsamlingskit säljer också (10jqka), även om händer, sensorer och armar alla faller i pris. Den andra är integrationslagret mellan armtillverkare och modellföretag, som äger varje stations uppgiftsdata, säkerhetshölje och MES-länk, vilket armtillverkare inte säljer som en produkt idag. Konkurrenter faller i fyra grupper: armledare (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot med flera), seende-plus-arm-företag (Mech-Mind, Hikrobot, SmartMore), förkroppsligade AI-företag (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) och bilchip-entranter.

Köparens acceptans: ingen siffra utan en nämnare

Eftersom ingen leverantör publicerar ingreppsfrekvenser, felintervall eller långsiktiga framgångsräkningar hör dessa hemma i kontraktets acceptansvillkor. Statistiskt bevisar 3,000 på varandra följande produktionscykler med noll fel en felfrekvens på högst 0.1% vid 95% konfidens; 10 fel i 3,000 bevisar bara ≤0.57%; att bevisa ≤1%, ≤0.5%, ≤0.1% och ≤0.01% med noll fel tar 299, 598, 2,995 respektive 29,956 cykler estimate, exact binomial. Enligt den standarden har Neuromekas 322/324 ett tvåsidigt 95%-intervall på 97.8–99.9% och bevisar inte ≥99%; AgiBots ungefär 6 fel i 64,828 uppgifter antyder 99.98–100%, men bara om "fel" definieras som köparen definierar det estimate. Dessa är

de acceptanskrav jag skulle ställa på varje robotarms- eller förkroppsligad cell; trösklarna bygger på Figures egna mål vid BMW och på Neuromeka- och AgiBot-fallen, och de är bedömning, inte en standard estimate :

Punkt	Kräv	Föreslagen godkännandetröskel
Definitioner	Undertecknade definitioner av cykel, framgång, fel (delframgång, omgrepp, tapp, kvalitetsläckage), ingrepp (beröring, fjärrteleoperation, återställning, nödstopp) och hur autonomi beräknas	Definitioner kommer före varje mått; ingen procent utan en formel
Ingrepp och teleoperation	Per robottimme, per typ och varaktighet; medeltid mellan ingrepp; teleoperationsminuter som andel av drifttid; operatörer per robot	Strukturerade stationer ≤0.5 ingrepp per robottimme (≥2 h mellan); noll teleoperation under acceptans
Långsiktig framgång	≥3,000 på varandra följande produktionscykler över ≥3 skift, ≥2 artikelnummer och normal belysnings- och leveransvariation; fel klassificerade; exakt undre gräns	Undre gräns ≥99.5% för linjestoppande inline-stationer, ≥99.0% offline; kvalitetsläckage rapporteras separat
Cykeltid	P50/P90/P99 per del inklusive återhämtning, mot ≥30 tidsmätta mänskliga cykler vid samma station	≤1.15 × mänsklig P50, eller inom linjetakt
Drifttid och fel	Körtimmar, driftstoppslogg, fel per komponent, delbytestid	≥95% drifttid över ≥2 veckor; MTTR ≤20 min med reservdelar på plats
Återhämtning	Skriptad felinjektion: tapp, ocklusion, belysningsförändring, transportbandsstopp, MES- eller nätverksbortfall, strömavbrott, nödstopp	≥90% autonom återhämtning, tillbaka i produktion inom 2 min
Omställning	Timmar, demonstrationer, operatörsskicklighet och datainsamlingskostnad per nytt artikelnummer eller fixtur	Strukturerade stationer: ≤1 skift och ≤100 demos per nytt artikelnummer
Säkerhet	Arm och cell enligt ISO 10218-1/-2:2025, cellriskbedömning, lokala regler (GB 11291 på fastlandet; CNS 14490 plus skriftlig bedömning i Taiwan; ANSI/A3 R15.06-2025 och UL 1740 i USA)	Ingen säkerhetsfunktion implementerad av maskininlärning; ett certifierat säkerhetslager (typiskt PL d / Kat 3) begränsar hastighet och kraft; inga inhägnadslösa gående robotar tills ISO 25785-1 är publicerad och bedömd
Beräkningskraft och offline-beteende	Inferenshårdvara, P99-latens, beteende utan nätverk, modellversionering, omvalidering efter varje uppdatering	Taktkritisk inferens beror inte på molnet; full om-acceptans efter varje modell-, kvantiserings- eller motorbyte
Bevis	Oredigerad video och råloggar för hela acceptansfönstret; bevittnat av kundens kvalitetsteam eller en tredje part	Inga highlight-rullar eller leverantörs-sammanfattningar
Kommersiella villkor	Betalningsmilstolpar knutna till acceptansmått; tillgänglighetsviten när det faktureras per robottimme; dataägande och fjärråtkomstregler	Slutbetalning och retention frigörs bara efter 3,000-cykelacceptansen och ≥4 veckors stabil drift
Ekonomi	Netto ersatta personer = hastighetskvot × drifttid × framgång − ingreppstid − teleoperationsandel	Avvisa varje förslag vars nettoersättning inte kan stödja dess påstådda återbetalningstid

Källor: (Figure, Neuromeka, Yicai Global, i-SCOOP); trösklarna är mina förslag.

Öppna frågor

07 Vad det offentliga källläget inte kan besvara, och hur ett pilotprojekt avgör det

Frågorna som offentliga källor inte kan besvara faller i fem grupper: vad AI-granskning faktiskt säljs för och vad köpare betalar; hur AOI-tillverkare paketerar AI och uppgraderar gamla maskiner; hur gamla de installerade AOI:erna är och hur många som redan kör AI; hur många personer och timmar granskning verkligen tar per linje; och hur piloter konverterar och hur köpare betalar. Ovanpå detta kommer hur de nyaste modellerna, kvantiserade byggen och robotarmarna faktiskt presterar under en given fabriks förhållanden. Var och en kan omvandlas till ett test med en tydlig tröskel. Trösklarna nedan är beslutsregler jag härlett från bevisen i denna rapport estimate ; räkna om dem när den första fabriksdatan kommer in.

Öppen fråga	Hur man avgör den	Hur ett bra resultat ser ut
Pris och betalningsvilja för AI-granskning	En standardoffertförfrågan (till exempel fyra linjer med blandade äldre 2D- och nyare 3D-AOI:er, två skift, cirka 2,000 kort per skift, 12 defektklasser, endast lokalt, MES- eller IPC-CFX-integration) skickad till flera tredjepartsleverantörer; en priskänslighetsundersökning med fyra frågor till 20–30 EMS-fabriker	De flesta offerter inom RMB 40–120k engångs eller 20–40k per år för fyra linjer; minst hälften av fabrikerna accepterar RMB 10–30k per linje med återbetalningstid inom 12 månader
AOI-tillverkares AI-paketering och uppgraderingar för äldre maskiner	Korta strukturerade intervjuer med cirka ett dussin AOI-tillverkare och distributörer	Om AI-granskning är standard, tillval eller frånvarande på 2025–26-modeller; om maskiner före 2020 kan uppgraderas och till vilket pris
Ålder på installerade AOI:er och andel som kör AI	Platsundersökning av cirka 30 EMS-fabriker, stratifierad efter storlek och region, med utrustningslistor granskade under NDA	Median-AOI-ålder ≥5 år och ≥60% utan AI innebär att eftermonteringspoolen är reell
Granskningsarbete och timmar	Intervjuer och 2–4 veckors AOI-larm-, granskning-, reparations- och skiftloggar vid två eller fler fabriker, med arbetsprovtagning	≥1 dedikerad granskare per linje per skift och ≥3 granskningstimmar per linje per skift, med minst 0.5 FTE per linje fabriken är villig att omfördela
Prestanda för AI-granskning i sig	Guldset plus skuggläge på en riktig linje (kapitel 3): dataavtal i vecka 0, baslinje och etikettering i veckorna 1–2, skuggkörning i veckorna 2–4, valfri assisterad körning i veckorna 4–8	Falsklarm eller granskningstimmar sänkta ≥50–60%; per klass läckagegränser inte sämre än den mänskliga baslinjen; noll missar på kritiska klasser; p95-latens ≤½ takt; automatiskt fail-open till människor
Lärar-etikettkvalitet hos frontiermodeller	Kör två eller tre kandidater på samma guldset, tre promptformuleringar vardera, en bra referensbild, enkomponentsutklipp ≥42–48 px	Automatisk acceptans endast för klasser med noll missar på ≥299 defekter och mänsklig-nivå-överensstämmelse; annars endast förmärkning och triage
Kvantiserade byggen	Ändra bara precisionen; fastnåla vikthashar, mall, förbearbetning, avkodning, motor och batchstorlek; inaktivera eller återställ prefix-cache; kontrollera att referensbygget reproducerar leverantörsresultat; poängsätt par på guldsetet och drift-slicen	Inga nya missar på kritiska klasser; ≤1 ny miss i 1,000 defekter; rensning ner ≤2 poäng; ≥99.5% identiska beslut över tre körningar; annars fall tillbaka till FP8/INT8

Öppen fråga	Hur man avgör den	Hur ett bra resultat ser ut
Linjesidig hårdvara, renhet och ESD	≥2 enheter av varje kandidatbox stresstestad vid två gånger den förväntade larmfrekvensen; 72-timmars soak vid verkligt kapslingsluftflöde; ESD-kontroll av kundens samordnare; ISO 14644-14-test för ISO 7- och renare kunder	p95 ≤½ takt; inga kraschar eller throttling över 72 timmar; ESD-sign-off; partikelrapport vid rätt klass
Betalningsvillkor, pilotregler och konvertering	Betalningsklausuler från senaste inspektionsutrustningskontrakt vid de undersökta fabrikerna; intervjuer om pilotregler, acceptanskriterier och kvalificeringstid	Cirka 30/60/10 med kontanter inom 60 dagar efter acceptans; piloter ≤8–12 veckor med en namngiven sign-off-ägare; ≥1 av 3 försök konverterar; ett ramverk som täcker ≥5–10 linjer år ett
Data, IT och ansvar	Intervjuer och kontraktsgranskning med jurist	En lokal box får ansluta och komma åt data; ansvar begränsat till mjukvarudefekter
Regional efterlevnad innan utlandsetablering	Taiwan: kundens skriftliga status enligt Cyber Security Management Act, E187-självbedömning; USA: tullutslag om hur en cell klassificeras, NRTL-offerter, ITAR-kontroll; Sydostasien: BOI lokal-koppling och malaysiskt avdrag, kundens moderbolag; fastlandet: CCC-omfattning för paketerade PC:ar och strömförsörjningar	Kunden är inte ett kontrollerat organ, och återbetalningstiden förblir ≤2 år med tull och certifiering inräknade
Acceptans av en robot- eller förkroppsligad cell	12-punkters acceptanstabellen i kapitel 6: undertecknade definitioner, ≥3,000 på varandra följande produktionscykler över ≥3 skift och ≥2 artikelnummer, skriptad felinjektion, ≥30 tidsmätta mänskliga cykler, bevittnat	Undre gräns ≥99.5% inline (≥99.0% offline); cykel ≤1.15 × mänsklig; ≤0.5 ingrepp per timme och noll teleoperation; ≥95% drifttid; MTTR ≤20 min; ≥90% återhämtning
ACT- och VLA-policyer på kinesiska edge-chip	Fyra veckor: portning och mikro-benchmark (≥1,000 inferenser, latensdistribution, effekt, 60-minuters soak); sluten-loop-simulering med uppmätt latens injicerad; A/B på riktig robot mot en GPU; 8–24 timmars uthållighet (≥3,000 cykler)	ACT-block p95 ≤100 ms, VLA ≤300 ms; framgång på riktig robot inom 3 poäng av GPU:n och cykel ≤1.2 × mänsklig; annars behåll policyn på Jetson eller en GPU
Cobot-, AMR- och integrationspriser	Modell-för-modell-offerter från flera cobot-distributörer, och helcells-offerter från två eller tre integratörer för testfixturlastning (en station, två skift)	Helcell ≤RMB 200k och återbetalningstid ≤1.5 år över två skift i Shenzhen
SMT-linjeantal i Taiwan, Sydostasien och USA	IPC (tidigare New Venture Research) data eller lokala utrustningsdistributörers installerade bas	Tillräckligt många adresserbara linjer per region för att motivera lokal applikationsingenjörskonst

Av alla dessa avgör fyra de mest känsliga siffrorna i denna rapport: priset köpare betalar, det arbete AI-granskning faktiskt frigör, om AI-granskning verkligen inte missar mer än människor gör, och hur snabbt kontanter kommer tillbaka. De är värda att avgöra före varje betydande åtagande, av en fabrik som överväger ett köp eller av en leverantör som överväger en produkt.

En pilotsekvens som fungerar för båda sidor

För en fabrik, och för en leverantör som vill ha en trovärdig referens, är ordningen densamma:

1. **Vecka 0–2: enas om data och definitioner.** Skriv under dataavtalet, definiera defektklasserna och vad som räknas som ett läckage, och registrera en baslinje: falsklarm per kort, granskningssekunder per larm, läckage som hittas nedströms.
2. **Vecka 1–4: bygg guldsetet** från AOI-arkivet och sparade NG-kort, etiketterat blint och avgjort (kapitel 3).
3. **Vecka 2–4 (eller längre): skuggläge.** Modellen loggar sina beslut medan människor bestämmer som tidigare; jämför per klass.

4. **Vecka 4–8: assisterad körning på slumpmässiga skift,** med modellen som bara rensar klasser som passerade grinden och allt annat går till människor.
5. **Acceptans, skriftligt, före skalning:** resultaten per klass, latens, fail-open-beteende, och vem som äger modeller och etiketter. Skala linje för linje, och omvalidera på varje ny produkt, modellversion, kvantiserings- eller motorbyte.

Begränsningar i denna rapport

Detta är skrivbordsforskning. Jag genomförde inga primärintervjuer med fabriker för denna utgåva; mycket linjedata kommer från leverantörer och har inte granskats oberoende. Flera myndighetswebbplatser, upphandlingsdatabaser och rekryteringsplattformar kunde inte nås, så Shenzhens uppgraderingssubventionssatser för 2026, lönevägledningen för Yangtzeflodsdeltat, transaktionspriser för AI-granskning, granskare per linje och pilotkonverteringsgrader är uppskattningar eller analogier; var och en framträder ovan som en öppen fråga. Växelkurser är People's Bank of Chinas centralparitet den 9 november 2026 ($1 = RMB 6.7188) och, för Taiwan, SMBC:s fixing samma dag ($1 = NT$32.12); den vietnamesiska dongen använder data från maj 2026. Alla trösklar är beslutsregler härledda från de citerade bevisen, inte standarder.

Om författaren

Tan Shuai

Tan Shuai är teknolog och grundare med mer än tjugo års arbete inom chip, hårdvara och systemmjukvara, och 20 beviljade patent. Hans arbete har fört honom genom Silicon Valley, New York, Chicago, Taiwan, Greater Bay Area (Shenzhen och Hongkong), Vietnam och Malaysia, med ett långt fokus på avancerad tillverkning och leveranskedjor. Idag arbetar han som AI-forskare och -utvecklare, med fokus på hur AI faktiskt används på fabriksgolvet.

Han välkomnar samtal med personer som arbetar med AI inom tillverkning, oavsett om man bygger ett team, söker en partner eller jämför anteckningar: i@tanshuai.com · tanshuai.com

Anteckningar och ansvarsfriskrivning

Denna rapport är oberoende skrivbordsforskning baserad på offentliga dokument tillgängliga fram till den 11 november 2026. Den tillhandahålls endast i informationssyfte och är inte investerings-, juridisk, skatte- eller ingenjörsrådgivning. Produktnamn, priser, modellsläpp, exportklassificeringar, tullar och regelverk förändras snabbt; verifiera allt du förlitar dig på mot primärkällan. Siffror markerade vendor claim eller estimate ska läsas som sådana. Företags- och produktnamn tillhör sina ägare och används endast för identifiering; inget företag har granskat eller sponsrat denna rapport.

© 2026 Tan Shuai. Du får dela denna rapport oförändrad med angivelse av källa.

ELEKTRONIKTILLVERKNING · AI-INSPEKTION × ROBOTIK

AI-visuell inspektion och robotarmar i elektronikfabriker: 2026 års bedömning

Varje nyckelsiffra bär sin källa och en tillförlitlighetsmarkör: verified, single source, vendor claim eller estimate. Det som det offentliga källäget inte kan besvara är samlat i kapitel 7, tillsammans med hur ett pilotprojekt avgör varje fråga.

Tan Shuai
tanshuai.comi@tanshuai.com

OFFENTLIG UTGÅVA · DATA PER DEN 11 NOVEMBER 2026

