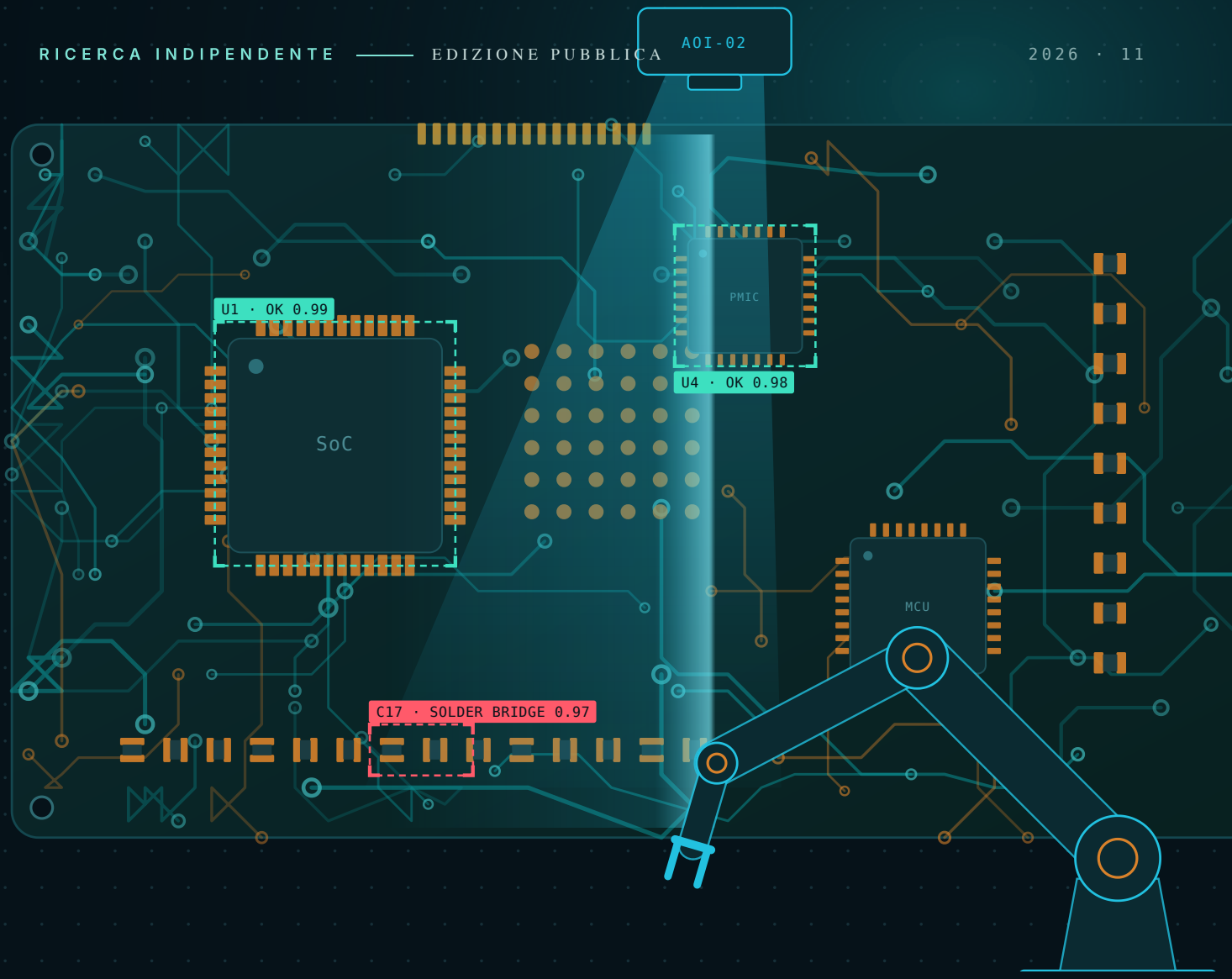


PRODUZIONE ELETTRONICA · ISPEZIONE AI × ROBOTICA

Ispezione visiva AI e bracci robotici nelle fabbriche di elettronica

Valutazione 2026

Configurazioni prevalenti · Fattibilità regionale · Distribuzione on-premises · Modelli insegnante–studente
Ritorno sull'investimento · Concorrenza · Cosa può rispondere solo un progetto pilota

RISULTATO CHIAVE

La revisione AI conviene, in modo mirato. I bracci robotici incarnati possono attendere.

Tan Shuai
tanshuai.comi@tanshuai.com

Dati aggiornati all'11 novembre 2026
Cina continentale · Taiwan · Sud-est asiatico · Stati Uniti

INDICE

Indice

◆ **SINTESI** — Conviene, ma in modo mirato, adesso e con partner — 4

01 **La linea oggi** — Cosa usano oggi le fabbriche di elettronica e cosa è tecnicamente fattibile — 8

02 **On-premises** — Far girare l'AI dentro lo stabilimento: camere bianche, air gap, hardware e quantizzazione — 16

03 **Insegnante e studente** — Insegnante e studente: i modelli frontier possono pre-etichettare i dati di ispezione; per i bracci robotici i veri insegnanti sono ancora le persone — 25

04 **Costi e rientro** — Costi e rientro: la revisione AI rientra in mesi, l'alimentazione macchine in anni, i bracci di prelievo mai — 34

05 **Mercato e concorrenza** — Mercato, concorrenza e come pagano gli acquirenti — 42

06 **AI incarnata** — AI incarnata: prima cambia come si insegnano i robot, poi come si muovono — 51

07 **Questioni aperte** — Cosa non può rispondere il registro pubblico, e come lo risolve un pilota — 57

◆ **L'autore** — Tan Shuai — 60

Tag di confidenza nel testo

- **verified** due o più fonti indipendenti, o due documenti primari indipendenti, concordano
- **single source** una sola fonte, inclusi i documenti primari
- **vendor claim** pubblicato dal venditore del prodotto o del servizio
- **estimate** calcolo o deduzione dell'autore a partire dai dati citati

Come leggere questo rapporto

Le cifre chiave portano un tag di confidenza. **verified** significa che due o più fonti indipendenti concordano, o due documenti primari indipendenti concordano; **single source** significa una sola fonte (inclusi i documenti primari); **vendor claim** significa pubblicato dal venditore del prodotto o del servizio, anche se ampiamente ripetuto; **estimate** significa un mio calcolo o deduzione dai dati citati, con gli input documentati. "Primario" descrive il tipo di fonte (relazione annuale, prospetto, regolamento, norma, statistica ufficiale); non cambia la classe di confidenza. Dove le fonti sono in conflitto, le cifre sono riportate affiancate e non mediate. Le conversioni valutarie usano la parità centrale della People's Bank of China del 9 novembre 2026, $1 = RMB 6.7188 (Xinhua), e per Taiwan il fixing di SMBC dello stesso giorno, $1 = NT$32.12 (SMBC). I dati su lavoro e sussidi della Cina continentale usano il Delta del Fiume delle Perle, dove la produzione elettronica è più concentrata (dati salariali ufficiali di Shenzhen), con il Delta del Fiume Azzurro per confronto. Le affermazioni da interviste di settore sono trattate come affermazioni da verificare, mai come fatti. Le questioni a cui il registro pubblico non può rispondere sono raccolte nel capitolo 7.

FIGURE

Figure



01	Dove l'AI si inserisce oggi in fabbrica	6
02	Una linea SMT in volume: le persone siedono dietro l'AOI	9
03	Circa il 97–99 % degli allarmi AOI sono falsi	10
04	Una fotocamera generica non è un AOI	12
05	Una distribuzione isolata (air gap)	16
06	Le zone pulite ed ESD decidono dove può stare il calcolo	17
07	Limiti di velocità per uso	19
08	Calcolo on-premise: memoria, prezzo e cosa si può acquistare legalmente in Cina continentale	20
09	Una grande workstation non è un assistente condiviso	21
10	Quantizzazione: 8-bit è sicuro, 4-bit dipende dalla ricetta, 3-bit e meno è escluso	22
11	Candidati maestro: i punteggi generici sono raddoppiati in una generazione	26
12	«C'è un difetto?» resta difficile per i modelli generali	27
13	Due pipeline maestro–studente	28
14	Quanto deve essere grande il set di riferimento	29
15	Latenza di politica edge	31
16	Studenti robot: vicini al livello di produzione solo se la postazione è strutturata	32
17	Quanto costa all'anno un addetto a ispezione o movimentazione	35
18	La scala dei prezzi	36
19	Ritorno per regione	38
20	Perché estrarre le schede difettose non ripaga	39
21	Dazio USA su robot, AOI e computer edge	40
22	Le dimensioni di mercato differiscono di oltre dieci volte	43
23	Bottom-up: 40.000 linee → 16.000 servibili → ≈320 M RMB una tantum	44
24	La finestra di retrofit è di circa cinque anni	45
25	Il cuneo di revisione è accerchiato su tre lati	45
26	Il margine sta nell'hardware industrializzato	47
27	Come pagano gli acquirenti	48
28	Giorni di vendite in attesa	49
29	Lo stack di AI incarnata	51
30	Cosa divulgano davvero le distribuzioni robot in stabilimento	53
31	Manodopera netta sostituita per robot	53
32	Cronologia di norme e regolamentazione	55

SINTESI

Conviene, ma in modo mirato, adesso e con partner

RISULTATO CHIAVE

Il verdetto: per le fabbriche di elettronica, ciò che è maturo nel 2026 e per cui le fabbriche pagheranno è la revisione AI e l'ispezione di nicchia su immagini AOI, più l'alimentazione macchine robotica programmata, non "un grande modello che guarda le schede più robot incarnati che le prelevano".

0.46–0.65 anni

Rientro dell'investimento per la revisione AI negli Stati Uniti, a Taiwan e a Shenzhen, a RMB 50k per linea

≥95% · <20 PPM

Copertura ed errori sfuggiti della revisione AI interna di un grande EMS: l'asticella per vendere agli EMS di primo livello

100–700 Mln RMB

Mercato di retrofit accessibile a terzi nella Cina continentale, una tantum, su un arco di circa cinque anni

0 punteggi

Punteggi pubblicati sui benchmark dei difetti industriali per i modelli di punta di novembre 2026: insegnanti, non ispettori

1.1–1.5 anni

Rientro di una cella cobot di alimentazione macchine su due turni negli Stati Uniti, a Taiwan e a Shenzhen

0.003–0.14 FTE

Manodopera risparmiata per turno da un braccio robotico di prelievo su una linea SMT: non rientra mai

Ispezione
ROI

Il deep learning supervisionato applicato alle immagini che un AOI esistente già produce riduce i falsi allarmi o le ore di revisione del 60–90% (vendor claims). DBG, una grande azienda EMS, riporta nella sua relazione annuale che il suo sistema interno copre oltre il 95% dei componenti segnalati, con errori sfuggiti sotto i 20 PPM e oltre l'85% in meno di revisione umana; è questa l'asticella che un prodotto di terze parti deve superare per vendere agli EMS di primo livello. A RMB 50,000 per linea, la revisione AI rientra entro un anno nella Cina continentale, a Taiwan e negli Stati Uniti (Shenzhen 0.65 anni, Taiwan 0.53, Stati Uniti 0.46), le celle cobot di alimentazione macchine in 1.1–1.5 anni e il Sud-est asiatico in 2–6 anni, mentre un braccio robotico aggiunto a una linea SMT solo per estrarre le schede difettose non rientra mai da nessuna parte, perché con una resa al primo passaggio superiore al 90% l'estrazione richiede solo 0.003–0.14 di una persona per turno estimate.

Modelli
ROBOT

I grandi modelli multimodali non possono ancora essere l'ispettore primario. All'11 novembre 2026, nessuno dei modelli di punta (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 e altri) ha un punteggio pubblicato su alcun benchmark di difetti industriali o PCB, e nella sfida PCBA di ACM MM 2026 i primi cinque sistemi, addestrati su 6,600 immagini reali di linea, hanno totalizzato solo 77–82 su "c'è un difetto". I modelli frontier possono pre-etichettare e fare triage. Per i bracci robotici, i veri insegnanti restano le dimostrazioni umane, il reinforcement learning su robot reale e la simulazione: gli studenti raggiungono il 90–99.4% su stazioni fisse e strutturate, ma solo il 30–80% su nastri in movimento, parti mai viste e nuovi prodotti; un agente frontier ha completato 104 di 108 compiti quasi-statici su robot reale, a 280 secondi ciascuno.

On-premises
DISTRIBUZIONE

La distribuzione air-gapped in fabbriche pulite e controllate è pienamente praticabile. Una normale sala SMT è di solito un'area protetta ESD piuttosto che una camera bianca classificata: un box fanless Jetson o NPU domestico va in linea e i server GPU vanno nella sala server. Secondo le classificazioni di export di NVIDIA, Jetson Thor e DGX Spark possono essere venduti a una normale fabbrica della Cina continentale senza licenza statunitense; DGX Station no. La precisione predefinita è a 8 bit; il 4 bit quantization-aware del fornitore è utilizzabile con condizioni; 3 bit e inferiori sono esclusi.

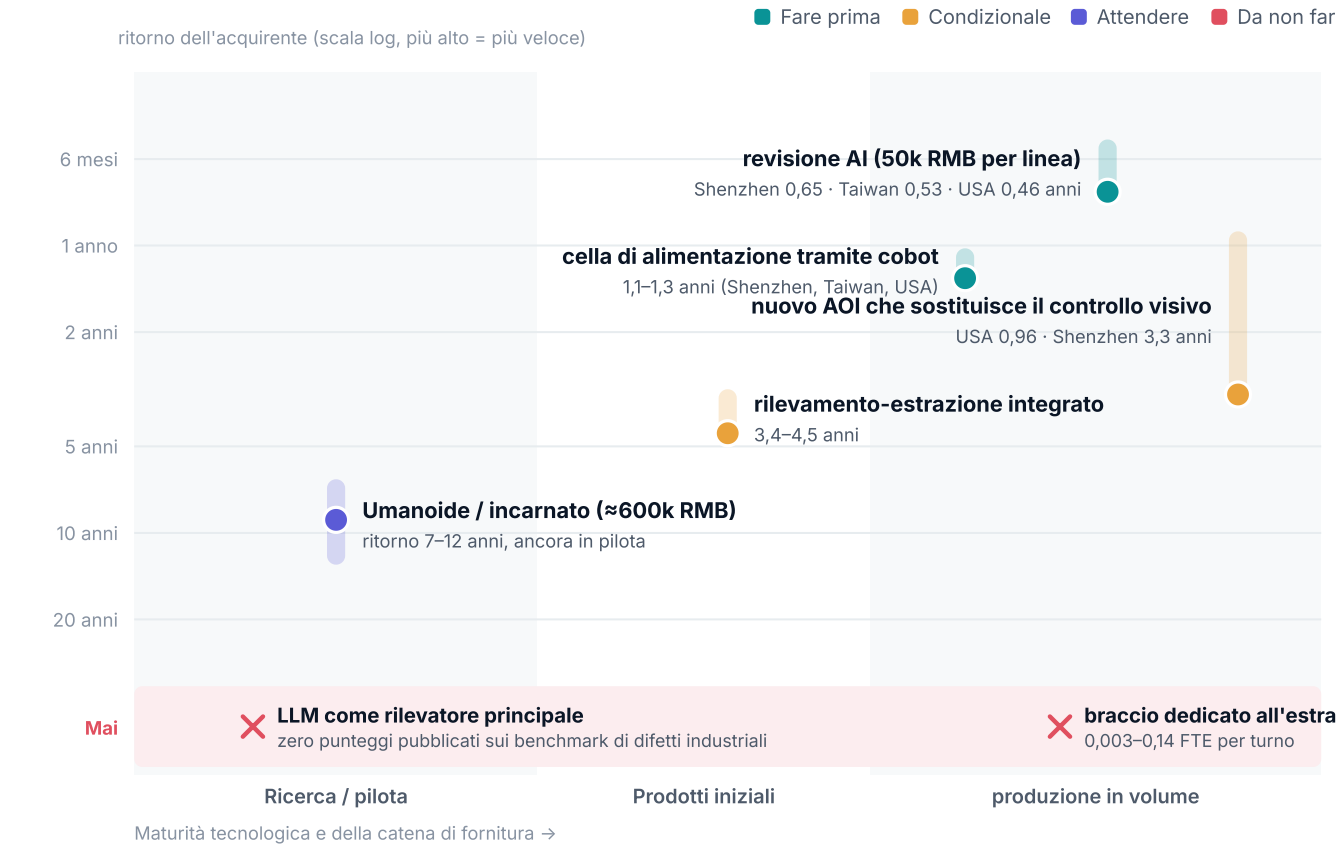
Mercato
CONCORRENZA

Il mercato è reale ma limitato. La Cina continentale ha circa 35,000–46,000 linee SMT e circa 26,000–66,000 AOI in linea che non usano l'AI; il mercato servibile per i retrofit di terze parti è di circa RMB 100–700 milioni una tantum, su un arco di circa cinque anni, e i produttori di AOI hanno iniziato a integrare la revisione nelle nuove macchine estimate. La concorrenza è già forte: il cuneo della revisione è accerchiato dai produttori di AOI, dai team interni degli EMS e dai fornitori di piattaforme, e il margine sta nell'hardware prodotto in serie.

Domanda	Giudizio	Numeri chiave (fonti e livello di confidenza nei capitoli)
1 **Cosa usano oggi le fabbriche di elettronica mainstream?**	Una linea SMT ad alto volume è "SPI 3D + AOI pre- e post-reflow + raggi X campionati + ICT/FCT + riparazione e ispezione visiva", con revisione umana dietro ogni AOI. L'AI entra nella revisione in tre modi: integrata dai produttori di AOI, costruita internamente dai grandi EMS o aggiunta da terze parti. L'assemblaggio e la movimentazione usano bracci, SCARA e AMR con percezione appresa e movimento programmato; l'AI incarnata è ancora in fase pilota	Takt ≈13–22 s per scheda; un ODM ha stanziato RMB 30.6M di AOI per 18 linee, ≈3–5 AOI per linea; ≈97–99% degli allarmi AOI sono falsi allarmi
2 **Quanto lontano arriva la tecnologia, i soldi possono comprare zero persone, e cosa può usare ciascuna regione?**	L'ispezione è di livello industriale come "detector dedicato + revisione AI + poche persone"; nemmeno soldi illimitati raggiungono zero persone. La Cina continentale usa solo modelli domestici o pesi aperti locali; le fabbriche private a Taiwan possono usare quasi tutto, ma fabs e infrastrutture critiche escludono in pratica i marchi della Cina continentale; il Sud-est asiatico ha le minori restrizioni; gli Stati Uniti hanno lo stack completo più dazi e certificazione	La revisione AI riduce i falsi allarmi o le ore di revisione del 60–90% vendor claims ; una fotocamera 4K su un intero pannello ≈78 µm per pixel contro 8–20 µm per un AOI; i bracci prodotti nella Cina continentale pagano ≈40% di dazio statunitense combinato, gli strumenti di tipo AOI ≈37.5%
3 **Può girare on-premises in camere bianche e fabbriche air-gapped, su quale hardware, abbastanza veloce e precisa?**	Sì; i vincoli sono procedurali, non di calcolo. Un box fanless ≤100 W in linea, server GPU nella sala server; Thor per le stazioni, un Mac Studio o un appliance domestico per il Q&A di fabbrica; nessun DGX Station nella Cina continentale. La velocità va bene quando la dimensione del modello corrisponde al compito; mantieni 8 bit, usa il 4 bit quantization-aware del fornitore con condizioni, mai 3 bit	≈0.5 s per ritaglio per un modello di visione 8B su Thor estimate ; kit di sviluppo Thor RMB 59,104 in Cina; DGX Station $123–140k, 139 token/s in totale per 32 utenti su un modello 743B
4 **Il modello di punta più recente può fare da insegnante per decine di round ed essere distillato in studenti locali, anche per la produzione flessibile e i bracci robotici?**	Ispezione: pre-etichettatura e triage; accettare automaticamente le etichette dell'insegnante solo per classi con "≥299 difetti, zero errori" su un golden set. Robot: gli insegnanti sono ancora dimostrazioni umane, RL su robot reale e simulazione, e gli studenti si avvicinano al livello produttivo solo su stazioni strutturate	Nessun modello di punta di novembre ha un punteggio sui difetti industriali; studenti al 97.8% su connettori addestrati, 29–67% su quelli mai visti, 37–59% su nastri in movimento; modelli frontier che comandano direttamente i giunti 0–5.5%

Domanda	Giudizio	Numeri chiave (fonti e livello di confidenza nei capitoli)
⑤ **Quanto costa, e qual è il ROI per regione?**	La revisione AI rientra entro un anno nella Cina continentale, a Taiwan e negli Stati Uniti; l'alimentazione cobot 1.1–1.5 anni; il Sud-est asiatico 2–6 anni; un braccio di prelievo su una linea SMT non rientra mai	Box di revisione da RMB 50k per linea: Shenzhen 0.65 anni, Taiwan 0.53, Stati Uniti 0.46, Vietnam 2.2 estimate ; prezzo accettabile di un add-on di terze parti ≈RMB 10–30k per linea una tantum o RMB 5–10k all'anno
⑥ **I concorrenti sono già abbastanza forti, ed è meglio iniziare adesso o attendere?**	La domanda non è satura, ma l'offerta è affollata; l'AI OEM limita la finestra di retrofit a circa cinque anni. Entra adesso con un prodotto mirato tramite partner di canale, senza scontrarti frontalmente con i fornitori di piattaforme; attendi i dati di produzione 2027–2028 prima di scommettere sull'AI incarnata	≈26–66k AOI in linea senza AI nella Cina continentale, in calo dell'8–12% all'anno estimate ; il principale fornitore di visione AI ha registrato nel 2025 ricavi per RMB 1.086B e una perdita netta di RMB 991M; il ricavo 3C del comparabile quotato più vicino è calato del 79.9% anno su anno in gennaio–maggio 2026

FIG 01 **Dove l'AI si inserisce oggi in fabbrica*La revisione AI dei falsi allarmi AOI e l'alimentazione delle macchine tramite cobot sono mature e ripagano in fretta; un braccio di estrazione dedicato e un LLM come ispettore principale non valgono la pena.***



Fonte: Ritorno dal modello dell'autore (capitolo 4); l'asse di maturità è un giudizio dell'autore.

Cosa costruire e cosa non costruire

	Raccomandazione
Costruire	Revisione AI e classificazione dei difetti su immagini AOI, per EMS di medie dimensioni con AOI più vecchi o di marche miste e senza upgrade AI dal produttore di AOI; ispezione che l'AOI SMT copre male (fori passanti e saldatura manuale, conformal coating, cavi, connettori e FPC, aspetto finale e viti); celle cobot di carico e scarico dove ogni parte viene manipolata (maschere di test, vassoi, imballaggio), con l'ispezione come add-on; un golden set e la modalità ombra su una linea reale come base di ogni affermazione

	Raccomandazione
Non costruire	Una piattaforma generale “AI più braccio robotico”; un LLM multimodale come ispettore primario; un braccio robotico su una linea SMT solo per estrarre le schede NG (un deviatore o un buffer NG lo fa); umanoidi e robot-as-a-service; modelli frontier dentro un loop di controllo o una funzione di sicurezza; montagne di memoria GPU nelle stazioni per far girare modelli “a grandezza piena”; vendite dirette a fabs, infrastrutture critiche taiwanesi o stabilimenti di elettronica per la difesa statunitensi
Con chi collaborare	Produttori e distributori di AOI (interfacce di canale e di immagine; i distributori hanno portato il 46.68% dei ricavi 2025 di Sinic-Tek); produttori di cobot (bracci certificati e funzioni di sicurezza); aziende EMS con 10–40 linee (espansione progressiva e acconti); integratori locali in Thailandia e altrove (per soddisfare le regole BOI di collegamento locale)

La linea oggi

01 Cosa usano oggi le fabbriche di elettronica e cosa è tecnicamente fattibile

La linea ad alto volume standard: diversi AOI, raggi X campionati, revisione centrale, 13–22 secondi per scheda

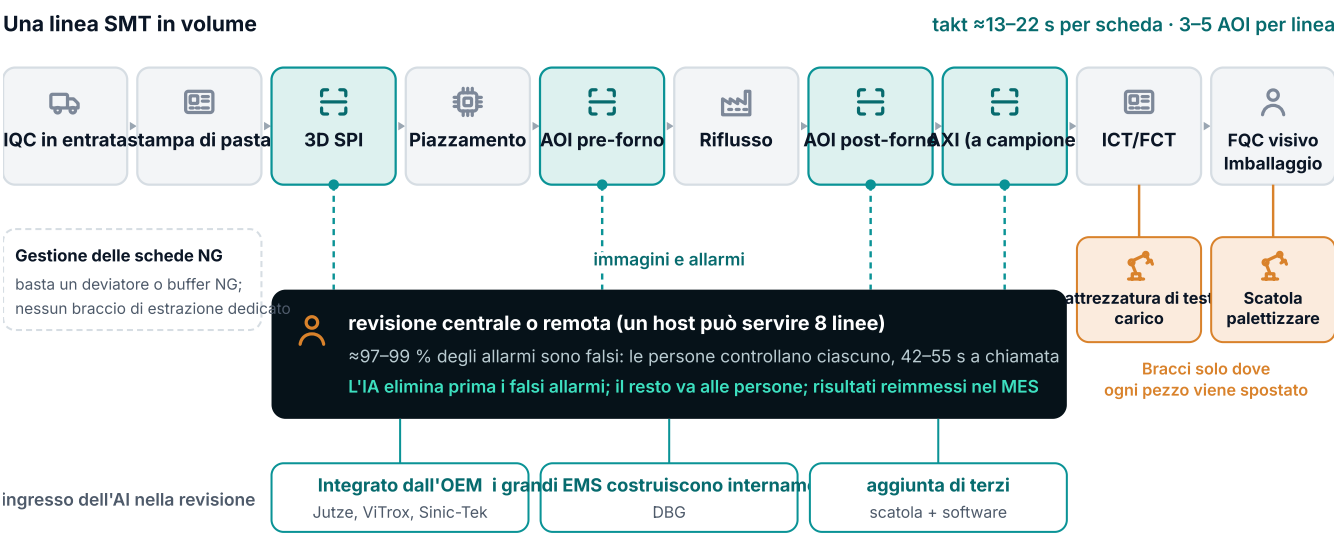
I documenti primari degli ODM quotati mostrano come appare realmente una linea ad alto volume del 2025–2026. Il prospetto di Hong Kong di Longcheer (dati al 30 settembre 2025) elenca i suoi controlli qualità SMT come "ispezione a raggi X, ispezione della pasta saldante (SPI), ispezione ottica automatizzata (AOI), test funzionale PCBA e ispezione visiva"; il suo flusso di processo prevede AOI pre-reflow, AOI post-reflow, raggi X campionati, schede NG a una stazione di riparazione SMT, quindi un controllo visivo finale. I suoi stabilimenti di Huizhou, Nanchang e Vietnam hanno 34, 24 e 8 linee SMT e 67, 60 e 13 linee di assemblaggio finale (prospetto H-share Longcheer) single source, primary filing. Lo stabilimento di laptop 2023 di Huaqin prevedeva 18 linee SMT con un budget AOI di RMB 30.6 milioni, circa RMB 1.7 milioni per linea, che ai prezzi domestici dell'AOI 3D corrisponde a tre-cinque AOI per linea per l'ispezione pre- e post-reflow di schede a doppia faccia (prospetto A-share Huaqin) primary filing + estimate. Una linea ad alto volume ha quindi di solito da due a quattro punti di revisione, non uno.

Le macchine sono già altamente automatizzate: Huaqin riporta il 96% di automazione sulle sue linee SMT (prospetto H-share Huaqin) single source. Le persone restanti sono nella revisione, nella riparazione, nell'ispezione visiva, nel test e nell'assemblaggio finale, e nessuna azienda pubblica quanti revisori ha per linea (capitolo 7).

Tre stime indipendenti concordano su un takt di **13–22 secondi per scheda** (un deposito ambientale 2026 per cinque linee Hebei (EIA progetto SMT Hebei), le linee di laptop di Huaqin e un caso iFactory), ovvero 7–11 secondi per lato per schede a doppia faccia estimate. La parte difficile di una linea ad alta varietà non è il volume ma l'introduzione di nuovi prodotti e i cambi di produzione, e ogni cambio è un altro avvio a freddo per un modello AI.

Sul lato software, le stazioni di revisione dei produttori di AOI sono già centralizzate: il software della stazione di riparazione di Sinic-Tek consente a "un host di controllare l'attrezzatura di otto linee SMT" (prospetto Sinic-Tek) single source, primary, 2022, e dopo che Jabil ha implementato il controllo remoto AOI ViTrox V-ONE nel 2026, un ingegnere poteva gestire 5–8 AOI da remoto contro 1–2 in locale (AtlasPCB) vendor claim. Un prodotto di revisione AI di terze parti si inserisce esattamente in questa stazione centrale, attraverso l'esportazione di immagini e risultati dell'AOI (IPC-CFX, IPC-HERMES-9852 o un formato proprio del fornitore). Gli stessi AOI stanno aggiungendo l'AI rapidamente: Yamaha afferma che il suo giudizio AI richiede circa un secondo per chiamata, aiuta gli ispettori dopo 30 immagini NG e diventa di qualità "molto alta" dopo 1,000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Una linea SMT in volume: le persone siedono dietro l'AOI** ***La manodopera si concentra nelle postazioni di revisione dopo l'ispezione ottica automatica. È lì che entra l'AI.***



Fonte: Documenti pubblici e materiale dei fornitori (Huaqin, DBG, Sinic-Tek e altri); compilato dall'autore.

Fase	Prodotti e tecnologia mainstream	Fornitori rappresentativi	Quanto lontano è arrivata l'AI
Dopo la stampa	SPI 3D (luce strutturata, moiré, alcune triangolazioni laser)	Koh Young, Sinic-Tek, TRI	Giudizio della pasta OEM con AI in produzione; Jutze afferma che alcuni modelli sono "completamente AI"
Dopo il posizionamento e il reflow	AOI 2D/3D con LED multiangolo e multicolore e luce strutturata, 8–20 µm per pixel, fotocamere 8–12 MP	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek e altri	Revisione AI OEM (Jutze in produzione, ViTrox V510Ai, Sinic-Tek con Tencent); i grandi EMS costruiscono la propria (DBG); add-on di terze parti
Giunti a terminazione inferiore e nascosti	Raggi X campionati o in linea (AXI)	—	Hikvision afferma che il suo modello Guanlan sbaglia i vuoti ai raggi X meno dell'1% delle volte vendor claim
Revisione e riparazione	Stazione di revisione centrale (un host, otto linee), revisione remota, stazione di riparazione	Software dei produttori di AOI, ViTrox V-ONE	L'AI chiude i falsi allarmi, le persone fanno il resto
Ispezione visiva, IQC, FQC	Per lo più manuale; alcune macchine di aspetto AI con carico robotico	SmartMore e altri	Macchine di aspetto in produzione; nessun dato pubblico di organico per IQC/FQC
Movimentazione e alimentazione macchine	Bracci a sei assi e cobot, SCARA, AMR, guida visiva 2D/3D	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Percezione appresa più movimento programmato in produzione; VLA e umanoidi ancora in fase pilota
Dati e sistemi	MES, IPC-CFX, IPC-HERMES-9852	—	Risultati riscritti nel MES; il supporto nell'attrezzatura più vecchia è irregolare

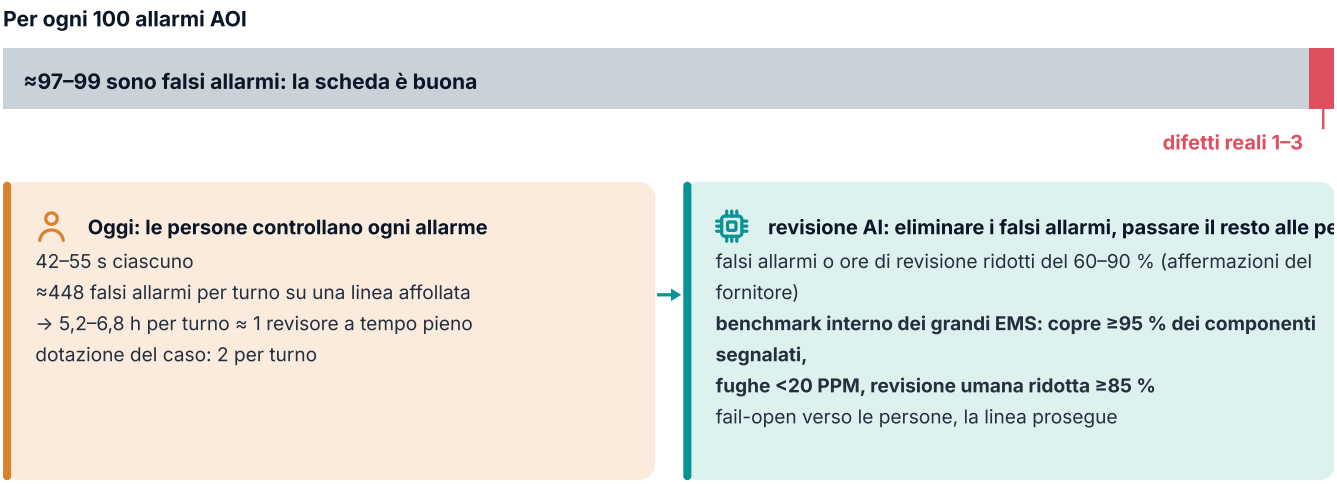
Fonti della tabella: i documenti sopra e (Koh Young, relazione annuale Jutze 2025, relazione annuale Sinic-Tek 2025, ViTrox, Sohu su Hengtianyi, Xinhua su Hikvision).

Le persone siedono dietro l'AOI, perché la maggior parte degli allarmi è falsa

La catena di ispezione su una linea SMT/PCBA mainstream procede: SPI 3D dopo la stampa per pasta mancante, insufficiente o in eccesso, ponti e offset; AOI dopo il posizionamento (opzionale) e dopo il reflow per parti mancanti, spostate, ruotate, invertite o "tombstone" e volume di saldatura; raggi X quando ci sono BGA o altre parti a terminazione inferiore; quindi test in-circuit o a sonde mobili, test funzionale e ispezione visiva manuale. Ogni AOI ha una postazione di revisione e una stazione di riparazione dietro di sé (prospetto Sinic-Tek, JLCPCB) verified. L'AOI post-reflow è il più comune e con il miglior rapporto qualità-prezzo, ma vede solo forma e area; giunti, vuoti e bagnatura sotto un BGA richiedono raggi X o test elettrico (JLCPCB).

La manodopera sta nella revisione perché la maggior parte degli allarmi è falsa. Un EMS britannico ha registrato **54,019 falsi allarmi contro 544 difetti reali su 917 codici articolo, circa il 99% degli allarmi** (arXiv 2505.03576) single source, preprint; nello stabilimento Siemens di Rastatt **il 97.5% delle schede rifiutate dall'AOI era buono** (caso di studio Siemens) vendor claim. Una linea impegnata vede circa 448 falsi allarmi per turno a 42–55 secondi ciascuno, 5.2–6.8 ore per turno, all'incirca un revisore a tempo pieno (iFactory) vendor claim; le ore sono una stima.

FIG 03 **Circa il 97–99 % degli allarmi AOI sono falsi** ***Oggi una persona conferma ogni chiamata; la revisione AI elimina prima i falsi e passa il resto alle persone.***



Fonte: Caso di studio Siemens, iFactory (affermazione del fornitore), rapporto annuale DBG 2025; ore di revisione stimate dall'autore.

Da "ore di revisione per turno = schede × falsi allarmi per scheda × secondi per chiamata", un revisore può coprire circa 0.7–1 linea ad alto volume con un AOI basato su regole. Una volta che l'AI rimuove il 60–90% dei falsi allarmi, il solo tempo di giudizio consentirebbe 2–14 linee per persona, ma gli spostamenti, la gestione del buffer NG e i passaggi di riparazione limitano la copertura locale a due-quattro linee adiacenti; solo la revisione centrale o remota si avvicina al tetto teorico estimate; i secondi per chiamata provengono da una sola fonte. Poiché gli AOI mainstream scansionano 60–80 cm² al secondo (guida AOI Yunheng 2026), un lato 20×15 cm richiede 4–5 secondi; l'AI deve solo elaborare le poche regioni segnalate entro un takt, non riscansionare la scheda.

Cosa funziona già in produzione: classificazione supervisionata sulle immagini stesse dell'AOI

Il modello collaudato è la classificazione supervisionata con deep learning sulle immagini dell'AOI che una fabbrica già possiede. L'add-on di Fujitsu sull'AOI di ADLINK a Taiwan ha ridotto i falsi allarmi di **oltre il 90%** (caso Fujitsu); OKI usa l'AI da luglio 2026 per ridurre la conferma manuale dopo l'AOI di circa l'**80%** nella propria attività EMS (Circuits Assembly); Siemens riporta il 60% in meno di analisi manuale, €160,287 risparmiati nel primo anno e un rientro in otto mesi (caso di studio Siemens) tutti vendor claims. Due cose sono comuni a questi casi: quasi nessuno pubblica un tasso di errori sfuggiti, e tutti mantengono persone sui restanti allarmi. L'unico caso che fornisce una cifra di er-

rori sfuggiti (dallo 0.8% allo 0.3%, contro uno 0.5% contrattuale) è internamente incoerente (iFactory). L'AI riduce le ore di revisione, non il lavoro di revisione in sé.

La revisione AI è passata dagli add-on di terze parti alle funzioni OEM e ai sistemi interni, il che cambia il metro con cui si misura un nuovo entrante. La relazione annuale 2025 di Jutze elenca il suo "upgrade AOI 3D" come in produzione di volume e afferma che "usa algoritmi AI per eliminare la revisione manuale al punto di ispezione" (relazione annuale Jutze 2025); Sinic-Tek ha sviluppato un "sistema di revisione manuale assistita da AI multimodale" dal 2019 e sta co-sviluppando un AOI 3D superiore-inferiore con Tencent, puntando a oltre il 99.5% di copertura dei difetti e al 98% di accuratezza di riconoscimento (relazione annuale Sinic-Tek 2025) vendor claims in primary filings. Le grandi aziende EMS costruiscono la propria. La relazione annuale 2025 di DBG elenca il "giudizio AI per l'ispezione visiva" come progetto completato che "copre oltre il 95% dei componenti segnalati dall'AOI", con un "tasso di errori sfuggiti sotto i 20 PPM" e "oltre l'85% in meno di revisione manuale", pienamente tracciabile (relazione annuale DBG 2025) single source, primary. È l'unica cifra di stabilimento elettronico che ho trovato che fornisce insieme copertura, errori sfuggiti e riduzione. **Un prodotto di terze parti venduto agli EMS di primo livello deve eguagliare il 95% di copertura, i 20 PPM di errori sfuggiti e l'85% di riduzione di DBG**; sotto il primo livello, il riferimento è la riduzione del 50–85% pubblicizzata dai fornitori.

Il rilevamento di anomalie addestrato solo su parti buone è veloce (EfficientAD-S impiega 2.2 ms per immagine (EfficientAD, WACV 2024)) ma debole sulle anomalie "logiche" come parti mancanti o sbagliate, e i metodi migliori hanno una media sotto il 60% di AU-PRO sul più difficile MVTec AD 2 (MVTec AD 2) single source. Le parti mancanti e sbagliate si verificano meglio contro CAD e BOM.

Il tetto a qualsiasi costo: strati di macchine, pochissime persone, ma non zero

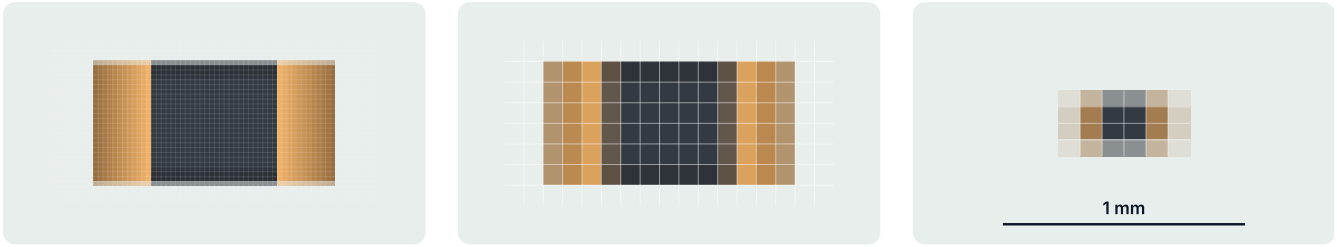
Non ho trovato alcuno stabilimento di elettronica che abbia eliminato del tutto la revisione manuale raggiungendo zero errori sfuggiti. La prassi per i prodotti IPC-A-610 Classe 3 (alta affidabilità) prevede ancora che ispettori addestrati decidano le chiamate borderline che le soglie automatiche non possono classificare, e un assemblatore di PCB definisce "l'ispezione automatizzata può sostituire la firma umana" un "fraintendimento comune" (PCBCart) single source. Anche la baseline umana va corretta: citando uno studio Sandia del 2012, un singolo ispettore visivo coglie al massimo circa l'**80%** dei difetti, due in serie circa il 96% (Instrumental) single source, citato da un fornitore. L'affermazione difendibile è "l'AI più un piccolo ciclo di revisione umana batte l'ispezione tutta manuale", non "l'AI sostituisce l'ispezione".

Senza limite di budget, il tetto è SPI 3D, AOI 3D, raggi X, revisione AI e stazioni di aspetto AI impilati insieme, con persone lasciate solo per le chiamate borderline, i nuovi prodotti e i cambi, i nuovi tipi di difetto e la firma del cliente. Gli stabilimenti di punta riportano l'85–92% in meno di ispezione manuale (I-Connect007, 2021, WiAdvance) vendor claims; la mia stima del tetto senza limiti di budget è una manodopera di revisione all'8–15% di quella attuale estimate. L'assemblaggio finale resta ad alta intensità di lavoro: l'automazione dei pulsanti dell'iPhone 16 è stata abbandonata per gli alti tassi di difetto (9to5Mac) single source.

Una fotocamera generica ad alta risoluzione non è un AOI

Gli AOI di produzione acquisiscono piastrella per piastrella a 8–20 µm per pixel, con fotocamere 8–12 MP, illuminazione programmabile multiangolo multicolore e 3D a luce strutturata (Koh Young, Saki) verified. Il settore ha notato da tempo che una parte 01005 copre solo circa 8×16 pixel a 25 µm per pixel, "forse non abbastanza" per un giudizio affidabile (Circuits Assembly, 2007). Con la stessa aritmetica estimate, una fotocamera 4K su un pannello da 300 mm ottiene circa **78 µm per pixel**: una parte imperiale 0402 è circa 13×6 pixel, sufficiente per presenza e grossi offset ma non per la forma del giunto, e un 01005 è circa 5×3 pixel, di fatto invisibile. Raggiungere 15 µm per pixel su un intero pannello richiederebbe circa 330 megapixel, motivo per cui gli AOI muovono una fotocamera sulla scheda. Quando qualcuno dice "0402", verifica quale intende: il 0402 metrico è il 01005 imperiale, ben oltre qualsiasi fotocamera grandangolare.

FIG 04 **Una fotocamera generica non è un AOI** ***Una fotocamera 4K su un pannello da 300 mm dà ≈78 µm per pixel: un 0402 si riduce a ≈13×6 pixel e un 01005 scompare.***



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

Fotocamera 4K su un pannello da 300 mm: ≈78 µm/px
Il 0402 si riduce a ≈13×6 px: appena sufficiente per presenza e scostamento grossolano

Stessa fotocamera su un 01005
≈5×3 px: di fatto invisibile

Fonte: Circuits Assembly; calcolo dell'autore per un pannello da 300 mm e un sensore 4K. Le griglie di pixel sono illustrative.

Sulle linee che hanno già gli AOI, il posto migliore per l'AI è sopra le immagini dell'AOI. Solo le linee senza AOI, o le stazioni di fine linea, giustificano una stazione fotografica chiusa con illuminazione controllata; una fotocamera aperta ad alta risoluzione sopra la linea è la più esposta a luce ambientale e ombre (CSDN).

Assemblaggio e movimentazione: percezione appresa più movimento programmato in produzione

Il rapporto 2026 dell'IFR afferma che "la visione artificiale per ispezione e guida dei robot è matura e di livello industriale", mentre "l'AI fisica general-purpose e l'apprendimento da simulazione ed esperienza non sono ancora di livello industriale" (sintesi IFR World Robotics 2026) single source, primary. L'elettronica è il maggior acquirente di robot (142,235 unità installate nel mondo nel 2025, il 59% di tutte le installazioni in Cina (IFR, comunicato IFR Cina)), eppure la Cina ha solo 166 robot ogni 10,000 addetti alla produzione contro 307 negli Stati Uniti (IFR, aprile 2026). Il modello in produzione è percezione appresa più movimento deterministico; il movimento appreso end-to-end è vicino al livello produttivo solo su stazioni strutturate (capitoli 3 e 6).

Cosa può usare ciascuna regione

La Cina continentale si affida a modelli domestici e pesi locali; a Taiwan decide l'identità del cliente; negli Stati Uniti, dazi e certificazione.

Elemento	Cina continentale	Taiwan (fabbriche private)	Sud-est asiatico	Stati Uniti
Claude (qualsiasi canale)	🔴 Non disponibile	🟠 Condizionato: il fornitore non deve essere a maggioranza di proprietà o controllato da persone o entità della Cina continentale, Hong Kong o Macao	🟠 Idem	🟠 Idem
Cloud GPT e Gemini	🔴 Non disponibile	🟢 Disponibile	🟢 Disponibile (liste non verificate paese per paese)	🟢 Disponibile
Modelli cloud cinesi (Qwen, DeepSeek, Doubao, GLM, Kimi)	🟢 Disponibile (modelli registrati; offrire un servizio richiede la registrazione)	🟠 Legale, ma l'hosting nella Cina continentale raramente supera la revisione di sicurezza del cliente; vietato nel settore pubblico, nelle imprese statali e nelle infrastrutture critiche	🟠 Legale (nodi internazionali)	🟠 Legale, revisione rigorosa

Elemento	Cina continentale	Taiwan (fabbriche private)	Sud-est asiatico	Stati Uniti
Pesi aperti locali di origine cinese (Qwen, GLM, DeepSeek)	🟢 Prima scelta; la distribuzione privata in fabbrica non richiede registrazione di modello o algoritmo (lettura di uno studio legale)	🟢 Disponibile (settore pubblico limitato)	🟢 Disponibile	🟡 Condizionato: revisione dell'origine
Pesi aperti locali di origine non cinese (es. Gemma 4)	🟢 Disponibile	🟢 Disponibile	🟢 Disponibile	🟢 Disponibile
Modelli di visione classici non generativi	🟢 Disponibile (fuori dalle regole sull'AI generativa)	🟢 Disponibile	🟢 Disponibile	🟢 Disponibile
Jetson Thor/T5000/T4000, DGX Spark	🟢 Disponibile: NVIDIA li classifica 5A992.c, nessuna licenza statunitense necessaria per una normale fabbrica civile; Spark ha una SKU Cina; prezzi di canale circa il 40–60% sopra il listino statunitense	🟢 Disponibile	🟢 Disponibile	🟢 Disponibile
RTX 5090, RTX PRO 6000, DGX Station	🔴 Non disponibile: i primi due sono 4A090.a, DGX Station 5A992.z; alternative legali sono 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	🟢 Disponibile	🟢 Disponibile, ma gli acquirenti con sede o in ultima istanza di proprietà cinese contano come cinesi; la Malaysia richiede permessi per export e trasbordo	🟢 Disponibile
Mac Studio M5 Ultra 512 GB	🟢 Disponibile (Apple Store Cina)	🟢 Disponibile	🟡 Presunto disponibile	🟢 Disponibile
Acceleratori Huawei Ascend e altri cinesi	🟢 Disponibile	🔴 Rischio elevato	🔴 Rischio elevato	🔴 Non utilizzabile in pratica
Bracci fissi, AOI, fotocamere industriali prodotti nella Cina continentale	🟢 Disponibile	🟡 Importabile; dal 1° dicembre 2025 le autorità possono agire direttamente sui marchi della Cina continentale, e agenzie, imprese statali, alcune fondazioni e fornitori di infrastrutture critiche possono essere limitati o esclusi dal loro uso	🟢 Disponibile	🟡 Legale, ma il dazio combinato è circa il 40% sui bracci, il 37.5% sugli strumenti di tipo AOI e le fotocamere, il 25% sui computer edge
AMR e manipolatori mobili di produzione estera	🟢 Disponibile	🟢 Disponibile	🟢 Disponibile	🟡 I nuovi modelli richiedono l'approvazione condizionata del Dipartimento della Guerra statunitense
Fotocamere Hikvision e Dahua	🟢 Disponibile	🟡 Uso privato consentito, settore pubblico vietato	🟢 Disponibile	🔴 Nuove autorizzazioni FCC bloccate; vietate per i fornitori federali

Elemento	Cina continentale	Taiwan (fabbriche private)	Sud-est asiatico	Stati Uniti
Sicurezza e certificazione dei robot	● GB 11291.1-2011 e GB 11291.2-2013 obbligatorie, in revisione da aprile 2026; certificazione CR volontaria; GB/T 39404-2026 sulla cybersecurity dei controller	● Recinzione di default; la collaborazione senza recinzione deve rispettare CNS 14490 o ISO 10218 con una valutazione scritta	● Per cliente e destinazione di export, confermato per iscritto prima della spedizione	● L'attrezzatura elettrica deve essere "approvata" (in pratica un marchio NRTL o una valutazione sul campo); UL 1740 richiede che le celle robot soddisfino ISO 10218-2 dall'11 agosto 2025
Fabs e stabilimenti a prodotti controllati	—	● I contratti di acquisto di TSMC richiedono SEMI E187; l'attrezzatura della Cina continentale viene rimossa dalle fabs a 2 nm	—	● I dati tecnici ITAR includono le "fotografie"; politica di diniego per la Cina; i fornitori controllati dalla Cina continentale non possono in pratica servire gli stabilimenti di elettronica per la difesa

Fonti della tabella: (regioni supportate Anthropic, paesi supportati OpenAI, disponibilità Gemini, misure provvisorie CAC sull'AI generativa, dati di classificazione export NVIDIA, eCFR 15 CFR 774, NVIDIA Cina DGX Spark, Apple Cina, Covington sulle nuove regole BIS, Greenberg Traurig, giugno 2026, Fulcrum, Tom's Hardware sulle indicazioni Ascend, lista di importazione BOFT Taiwan, storia del Cyber Security Management Act di Taiwan, regole di revisione di Taiwan per i prodotti che mettono in pericolo la cybersecurity nazionale, Taipei Times, USITC HTS capitolo 99, Latham & Watkins, IPVM, piattaforma nazionale norme cinesi, GB 11291.2, aggiornamento Intertek UL 1740, eCFR 22 CFR 126.1).

Cina continentale: uno stack completo, ma né i modelli frontier chiusi né le singole macchine più forti possono entrare. Modelli cloud domestici registrati, pesi aperti locali (Qwen sotto Apache-2.0, GLM e DeepSeek sotto MIT), modelli di visione classici e AOI, fotocamere e cobot domestici sono tutti utilizzabili. Le stesse classificazioni di export di NVIDIA elencano il kit Jetson AGX Thor, i moduli T5000/T4000 e DGX Spark come 5A992.c, quindi una normale fabbrica EMS o PCBA non ha bisogno di licenza statunitense, mentre RTX 5090 e RTX PRO 6000 (4A090.a) e DGX Station (5A992.z, TPP 58,484) non si possono acquistare nella Cina continentale (classificazioni export NVIDIA, eCFR) primary; la lettura giuridica è mia e non è consulenza legale. RS Cina elenca il kit Thor a RMB 59,104 tasse incluse, circa 1.4 volte il prezzo di listino statunitense (RS Cina).

Quattro studi legali concordano che una distribuzione privata in fabbrica per uso proprio non richiede registrazione di modello o algoritmo, mentre un servizio multi-tenant ospitato sta in una zona grigia (Covington, FPF, JunHe) verified: letture coerenti, nessuna pronuncia ufficiale. GB 11291.1-2011 e GB 11291.2-2013 restano la baseline obbligatoria di sicurezza dei robot e sono in revisione dal 28 aprile 2026; GB/T 39404-2026 sulla cybersecurity dei controller è entrata in vigore il 1° novembre 2026 (piano di revisione 20262542, GB/T 39404-2026) single source, primary. Finché la nuova GB 11291 non arriva (2028–2029 estimate), progetta secondo ISO 10218:2025 e mantieni i file di conformità GB 11291.

Taiwan: le fabbriche private possono usare quasi tutto; decide l'identità del cliente. API frontier, pesi aperti e attrezzatura della Cina continentale sono tutti legali (data.gov.tw). Ma da quando il Cyber Security Management Act modificato è entrato in vigore il **1° dicembre 2025**, le autorità possono agire direttamente sui prodotti che sono "marchi della Cina continentale" e limitarli presso enti governativi, imprese statali e fornitori di infrastrutture critiche (Database leggi e regolamenti, regole di revisione) verified; l'elenco dei fornitori di infrastrutture critiche è riservato; e TSMC ha scritto SEMI E187 nei suoi contratti di acquisto (CNA) e starebbe rimuovendo l'attrezzatura della Cina continentale dalle sue fabs a 2 nm (TrendForce) single source. Gli EMS privati, l'assemblaggio e i piccoli stabilimenti PCBA sono aperti; fabs, utility di parco e clienti legati al governo sono chiusi in pratica. Un fornitore controllato dalla Cina continentale dovrebbe progettare per E187, fornire una distinta base software, offrire aggiornamenti offline e non collegarsi mai da remoto dalla Cina continentale.

Sud-est asiatico: le minori restrizioni; i problemi sono il rientro lento e le case madri cinesi. Tutto è utilizzabile, ma le indicazioni statunitensi del 31 maggio 2026 richiedono una licenza per esportare articoli di calcolo avanzato a entità con sede, o in ultima istanza di proprietà, cinese anche quando il destinatario è in un paese terzo (Greenberg Traurig) single source, e la Malaysia ha permessi propri per i chip avanzati (Fulcrum). La domanda si sta spostando

qui: TrendForce prevedeva che Thailandia, Vietnam e Malaysia si avvicinassero al 50% della capacità SMT degli ODM di server nel 2026 (TrendForce) single source, una previsione del 2023 .

Stati Uniti: lo stack completo, al prezzo di dazi, certificazione e dati controllati. Sotto il capitolo 99 della tariffa doganale statunitense, un braccio robotico prodotto nella Cina continentale paga circa il 40% di dazio combinato, gli strumenti di tipo AOI e le fotocamere circa il 37.5% e i computer edge il 25%; prodotto in Malaysia o Indonesia circa il 10–12.5%, in Vietnam, Thailandia o Filippine il 12.5–15%, a Taiwan il 10%, con i computer allo 0% fuori dalla Cina continentale (USITC HTS capitolo 99, Great Lakes Customs Law) primary schedule + estimate . L'avviso USTR del 10 novembre ha esteso 178 vecchie esclusioni della Sezione 301 al 9 gennaio 2027, nessuna delle quali riguarda robot, strumenti di ispezione, fotocamere o computer (Federal Register 2026-22925) verified .

OSHA richiede che l'attrezzatura elettrica sul posto di lavoro sia "approvata" (in pratica un marchio NRTL, o una valutazione sul campo di circa $3,000–$20,000+ per cella), e dall'11 agosto 2025 UL 1740 porta le intere celle robot sotto ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). L'elettronica per la difesa è chiusa in pratica ai fornitori controllati dalla Cina continentale: l'ITAR tratta le fotografie di produzione come dati tecnici e applica una politica di diniego alla Cina (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Una cella destinata agli Stati Uniti dovrebbe usare un braccio fisso, fotocamere senza restrizioni e modelli locali o cloud statunitensi, con assemblaggio finale nel Sud-est asiatico dove questo cambia davvero l'origine.

A qualsiasi budget, il tetto differisce per regione ma da nessuna parte raggiunge zero persone estimate . La Cina continentale può impilare SPI 3D, AOI 3D, raggi X, revisione AI e un team AI interno per portare la manodopera di revisione all'8–15%, ma senza modelli frontier chiusi o DGX Station, il Q&A LLM di fabbrica deve girare su pesi aperti locali e calcolo domestico (un modello a grandezza piena per molti utenti richiede almeno due server Ascend a otto schede). Taiwan e il Sud-est asiatico possono tecnicamente usare il meglio di tutto, incluse API frontier e DGX Station, limitati dall'identità del cliente, dai marchi della Cina continentale e dalle licenze delle case madri cinesi. Gli Stati Uniti non hanno limiti tecnici, al prezzo di dazi, certificazione NRTL e separazione del personale per i dati controllati. Ovunque, le persone restano per le chiamate borderline, i nuovi prodotti e i cambi, i nuovi tipi di difetto e la firma del cliente.

Un prodotto venduto tra più regioni ha quindi bisogno di due stack di modelli: uno stack della Cina continentale su modelli domestici registrati o pesi aperti locali, e uno stack estero su API frontier più un modello locale di origine non cinese. La policy di Anthropic in vigore dal 12 novembre 2026 giudica per proprietà e controllo, quindi un fornitore controllato dalla Cina continentale non può costruire su Claude nemmeno tramite una sussidiaria estera, una piattaforma cloud o un rivenditore (Anthropic) single source, primary .

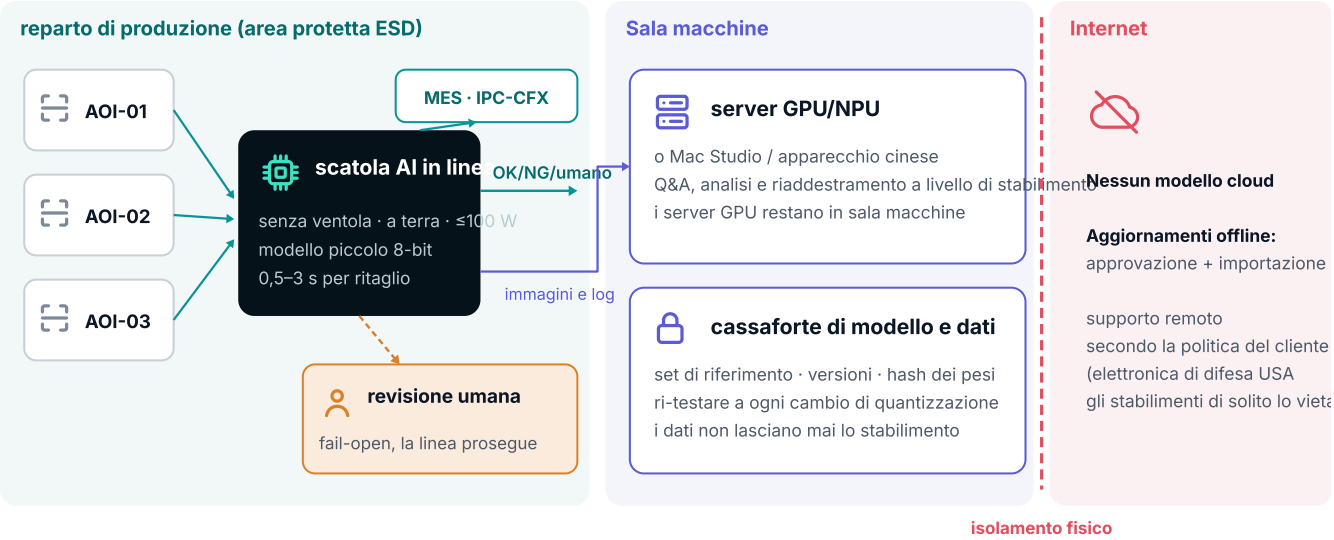
On-premises

02 Far girare l'AI dentro lo stabilimento: camere bianche, air gap, hardware e quantizzazione

In uno stabilimento air-gapped il vincolo è il processo, non il calcolo

L'ispezione visiva completamente offline è routine; i vincoli sono procedurali. La guida MIIT cinese del 2024 sulla sicurezza dei sistemi di controllo industriale chiede zonizzazione, firewall o diodi dati, manutenzione remota a tempo limitato e verificata, supporti scansionati e patch testate (testo della guida) single source ; i clienti dei semiconduttori aggiungono SEMI E187 ed E188, che TSMC ha scritto nei suoi contratti di acquisto nel 2023 (SEMI, TXOne) verified . I risultati tornano al MES su IPC-CFX e l'identità della scheda viaggia su IPC-HERMES-9852 (standard Hermes). La trappola comune è la licenza che presuppone una rete: la modalità offline di Roboflow mette in cache i pesi su un lease di 30 giorni (Roboflow). Un prodotto offline ha bisogno di attivazione offline, aggiornamenti firmati su supporti scansionati, una distinta base software, controllo USB, log di audit locali e un adattatore CFX.

FIG 05 **Una distribuzione isolata (air gap)** ***Una scatoletta in linea decide; la sala macchine risponde alle domande; i guasti ricadono sulle persone e i dati non lasciano mai lo stabilimento.***



Fonte: Sintesi dell'autore (capitolo 2).

I modelli frontier chiusi, in pratica, non entrano nella sala server di uno stabilimento. Gemini è offerto air-gapped solo tramite Google Distributed Cloud per governi e grandi imprese (Google Cloud); Foundry Local di Microsoft apre i grandi modelli solo a "clienti qualificati" e non nomina alcun modello GPT (Microsoft); Claude è offerto solo nelle regioni cloud governative, senza distribuzione on-premises (Claude for government); nessuno dei tre è offerto nella Cina continentale. L'attrezzatura di stabilimento deve essere progettata attorno ai pesi aperti; i modelli frontier possono al massimo etichettare o spiegare le immagini in sede, offline, quando il cliente consente che le immagini anonimizzate escano.

L'hardware della sala server va trattato come di grado ufficio: il Mac Studio è omologato per 10–35 °C e DGX Spark per 5–30 °C, con segnalazioni di spegnimenti termici sotto inferenza sostenuta (Apple, forum NVIDIA) verified . La linea riceve box industriali fanless, come un PC industriale Jetson T4000 omologato per −25 a 60 °C (AAEON) vendor

claim . Poiché un'ora di fermo costa a un grande stabilimento da $36,000 a $2.3 milioni (Siemens True Cost of Downtime 2024) vendor survey , ogni passo AI in linea deve **fallire in apertura (fail open)**: su qualsiasi guasto ricade sul processo esistente AOI-più-persone.

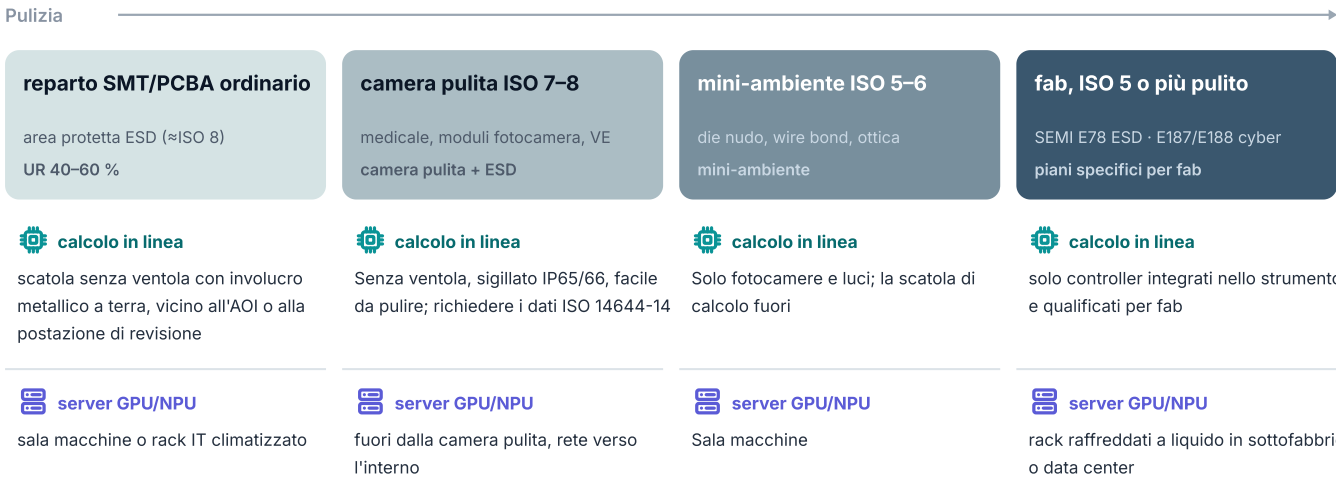
Zone pulite ed ESD: una normale sala SMT è un'area protetta ESD, non una camera bianca classificata

Ciò che limita dove può stare il calcolo è l'elettrostatica e il flusso d'aria, non la classe di pulizia. La maggior parte delle aree di assemblaggio SMT "non è formalmente a controllo particellare", e gli stabilimenti di classe mondiale "possono raggiungere o avvicinarsi alla ISO 8" (panel di esperti Circuitnet) single source . I PCBA medicali e i moduli fotocamera stanno tipicamente in ISO 7, i die nudi e l'allineamento ottico in ISO 5–6 locale (PCBSync) single source ; i costruttori di camere bianche quotano ISO 7–8 per SMT in generale (Angstrom Technology) vendor claim .

Il requisito universale è il controllo ESD: ANSI/ESD S20.20-2021 richiede conduttori isolati sotto i 35 V e mantiene gli isolanti a forte campo ad almeno 30 cm dai dispositivi sensibili (ANSI, ESDA, Desco) verified . Le fabs mettono il calcolo nel sub-fab o nel data center, in rack a liquido che soddisfano SEMI S2/S8/F47 (Prodrive) vendor claim .

Se l'attrezzatura può entrare in un'area pulita è testato per l'emissione di particelle sotto ISO 14644-14; i robot possiedono tali certificati (Fraunhofer IPA), ma non ho trovato alcun PC industriale, box edge AI o server GPU che ne abbia uno. CE, FCC e IEC 60068 mostrano che un box tollera il suo ambiente, non che possa entrare in un'area protetta ESD; è il coordinatore ESD del cliente a verificarlo. I server GPU non vanno mai in linea: un DGX Station ha assorbito 1,292 W alla GPU nei test (StorageReview).

FIG 06 **Le zone pulite ed ESD decidono dove può stare il calcolo** ***Un reparto SMT ordinario è un'area protetta ESD, non una camera pulita. I server GPU e NPU restano in sala macchine.***



Fonte: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, norme SEMI e materiale dei fornitori; compilato dall'autore.

Zona	Ambiente tipico	Calcolo in linea	Server GPU/NPU
Sala SMT/PCBA ordinaria (consumer, industriale)	Area protetta ESD sotto S20.20 o IEC 61340-5-1, di solito non una camera bianca classificata (≈equivalente ISO 8); 40–60% RH	Box fanless in un involucro metallico collegato a terra (Jetson T4000/Thor, box NPU cinese) accanto all'AOI o alla postazione di revisione; monitor e involucri in plastica fuori dall'area o a ≥30 cm dalle schede	Sala server o rack IT climatizzato

Zona	Ambiente tipico	Calcolo in linea	Server GPU/NPU
Medicale, moduli fotocamera, elettronica EV (camera bianca ISO 7–8)	Camera bianca più protezione ESD	Fanless, IP65/66 sigillato, liscio e facile da pulire; chiedi dati ISO 14644-14 o un certificato IPA	Fuori dalla camera bianca, collegato in rete
Die nudi, wire bond, stazioni ottiche (ISO 5–6 locale)	Mini-ambiente	Solo fotocamere e luci dentro; box di calcolo fuori	Sala server
Camera bianca fab (ISO 5 e più pulita)	Piani ESD e di attrazione elettrostatica specifici per fab, attrezzatura secondo SEMI E78; cybersecurity secondo SEMI E187/E188	Solo controller integrati nell'attrezzo e qualificati per fab; nessun PC add-on con ventole	Rack a liquido sub-fab o data center di fab

La tabella è la mia sintesi delle fonti sopra estimate. Per un prodotto, ciò significa un box fanless da linea sotto circa 100 W, un involucro metallico collegato a terra di protezione, ingresso 12–24 V DC, funzionamento a temperatura estesa al flusso d'aria reale dell'armadio, e un rapporto di emissione particellare ISO 14644-14 pronto per i clienti a ISO 7 e più pulita. Qualsiasi progetto in cui "un server GPU serve quattro linee" mette il server nella sala server e lascia in linea solo un piccolo box o una fotocamera.

Limiti di velocità per uso: l'accuratezza è più difficile della velocità

Fabbriche e fornitori non pubblicano quasi mai obiettivi di latenza o disponibilità dell'AI. Le soglie sotto derivano da takt, tempo di revisione umana, tassi di controllo dei robot e ricerca sull'interazione estimate:

Uso	Limite di velocità	Limite di accuratezza	Disponibilità
Ispezione in linea di ogni parte	Entro ½–1 takt: 240–1,000 ritagli in un takt di 10–30 s, 8–100 ritagli al secondo	Errori sfuggiti non peggiori dell'attuale "AOI più persone" (esempio contrattuale <0.5%)	Fail open, mai fermare la linea; obiettivo ≥99.9% (nessuno SLA pubblico)
Secondo parere in linea solo sui ritagli segnalati dall'AOI	Segnalazioni (di solito ≤5–20 per scheda) entro ½ takt: 0.5–3 s per ritaglio	Idem; addestrato sulle immagini della linea stessa	Ricade sulla revisione umana
Revisione AI dei falsi allarmi	2–5 s per chiamata, oltre dieci volte più veloce dei 42–55 s di una persona	Errori sfuggiti non peggiori di un revisore umano	Il degrado aggiunge persone, mai ferma la linea
Controllo robot (VLA)	100–300 ms per chunk di azione con chunking in tempo reale, output 10–50 Hz	≥99% di successo del compito per battere il movimento programmato	Sistema di sicurezza indipendente dal modello
Pianificazione robot e scelta delle skill	1–3 s quando serve a ogni ciclo; ≤1 min solo su eccezioni o cambi	L'output deve essere codice verificabile o un ID di skill	Ricade sul programma fisso
Q&A (SOP, query MES)	Primo token idealmente ≤1–2 s, al massimo 10 s; ≥15–20 token/s per utente; 5–20 utenti concorrenti	Citazioni accurate; fedele su contesto lungo	Grado IT d'ufficio (≈99–99.5%)
Analisi e rapporti di root-cause	I minuti vanno bene (batch)	Massima; fedele su contesto lungo	Grado IT d'ufficio

Physical Intelligence ha misurato 97 ms di inferenza per un VLA di classe 3B (circa 108–139 ms con rete e preprocessing), ha scoperto che il chunking in tempo reale mantiene invariato il throughput fino a 200 ms di ritardo ag-

giuntivo, e ha completato compiti fini come accendere un fiammifero o collegare un cavo di rete a oltre 300 ms (Physical Intelligence) vendor research; la ricerca sull'interazione pone il limite per mantenere il filo del pensiero a un secondo e per l'attenzione a dieci (NN/g). L'accuratezza è la parte difficile: Qwen3-VL-8B aperto ha una media di soli 56.8 zero-shot su cinque dataset di anomalie industriali (IndusAgent) single source. Nessun hardware salva un modello visione-linguaggio che non è stato messo a punto sulle immagini della linea stessa.

FIG 07 **Limiti di velocità per uso*L'ispezione in linea si misura in millisecondi, il Q&A in secondi, l'analisi in minuti. Le persone impiegano 42–55 s a chiamata.***



Fonte: Sintesi dell'autore (capitoli 1–2).

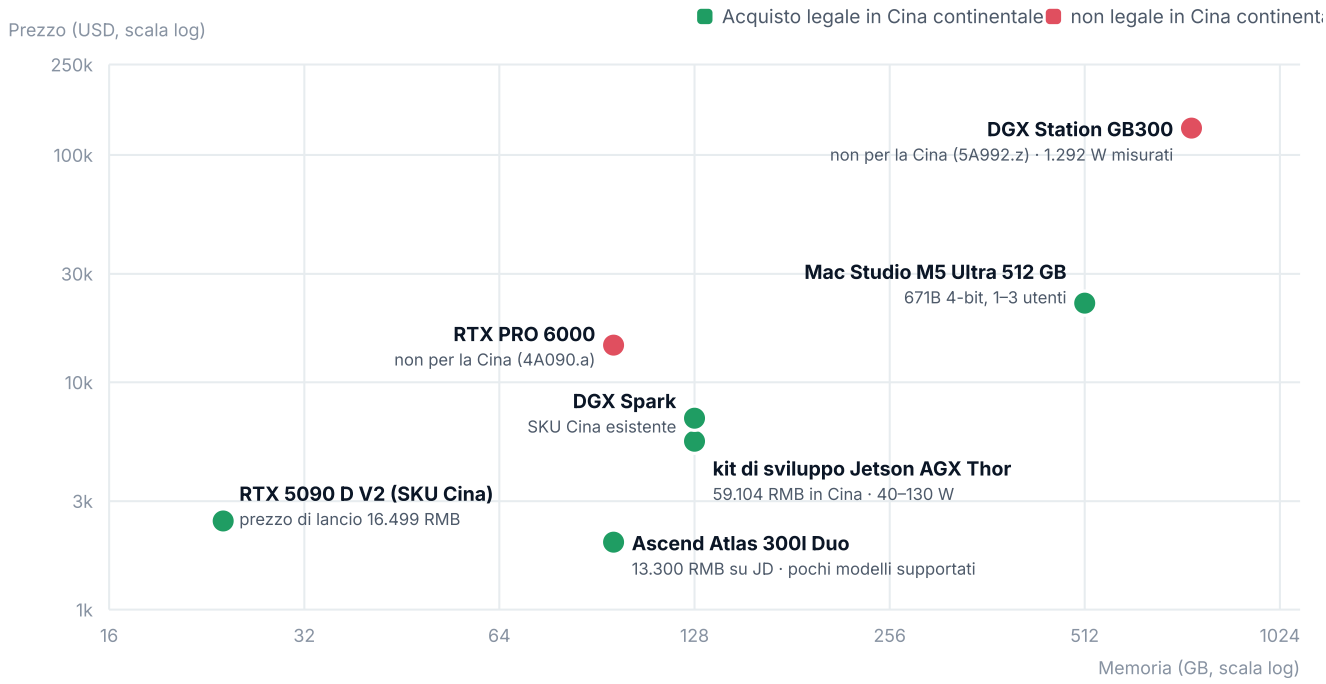
Hardware: Thor per la stazione, un Mac Studio o un appliance domestico per il Q&A di fabbrica, e nessun DGX Station nella Cina continentale

Dispositivo	Memoria e banda	Prezzo (novembre 2026)	Per cosa è adatto	Cina continentale
Kit di sviluppo Jetson AGX Thor / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 al lancio → $5,499 (luglio 2026); T5000 $4,999, T4000 $2,999; kit di sviluppo RS Cina RMB 59,104 tasse incluse	Modelli di visione 2–8B che ricontrollano i ritagli; VLA di classe 3B in tempo reale	Sì (5A992.c, nessuna licenza per fabbriche civili)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; lo stesso chip ASUS GX10 modello Cina RMB 32,999 su JD (ott 2025)	MoE 8–35B multiutente	Sì (5A992.c, modello Cina, canali autorizzati)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (giugno 2026) → ≈$16,005 listino medio eBay a novembre	FP8 27–35B multiutente; MoE di classe 120B a 4 bit	No (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB unificati, 1.2 TB/s; 480 W max sostenuti	Da $22,299 negli Stati Uniti; ≈RMB 173–186k stimati in Cina	DeepSeek 671B a 4 bit per 1–3 utenti	Sì
DGX Station GB300	748 GB (252 GB HBM3e a 7.1 TB/s + 496 GB LPDDR5X a 396 GB/s); 1,292 W misurati alla GPU	$123–140k prezzo di strada a novembre (≈RMB 830–940k prima del dazio); $95–117k ad agosto	Alta concorrenza su MoE 120–235B che sta in HBM; solo pochi utenti su modelli 433–743B	No (5A992.z, TPP 58,484)

Dispositivo	Memoria e banda	Prezzo (novembre 2026)	Per cosa è adatto	Cina continentale
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	RMB 13,300 da un venditore JD	Revisione batch a basso volume; pochi modelli supportati	● Sì (domestico)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Non ottenuto; DeepSeek pieno richiede due unità a W8A8, quattro a BF16	Q&A e analisi multiutente di fabbrica	● Sì (domestico)
RTX 5090 D V2 (modello Cina)	24 GB	RMB 16,499 al lancio (ago 2025), salito poi con i prezzi delle memorie	Modelli di visione 8B, piccoli VLA	● Sì

Fonti: (NVIDIA Cina Thor, lancio NVIDIA Thor, CNX sugli aumenti di prezzo, dati di classificazione export NVIDIA, RS Cina, Implicator, IThome su ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, specifiche Apple, TechRadar, IThome sui prezzi cinesi di M5 Ultra, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, matrice di supporto vLLM-Ascend, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). I prezzi della Cina continentale per il Mac Studio 512 GB sono estrapolati dai premi di configurazione statunitensi e sono a bassa confidenza estimate .

FIG 08 **Calcolo on-premise: memoria, prezzo e cosa si può acquistare legalmente in Cina continentale*Le classificazioni all'esportazione decidono la lista ristretta prima dei benchmark.***



Fonte: Classificazioni all'esportazione NVIDIA, Apple, RS China, JD.com e prezzi di recensione (novembre 2026).

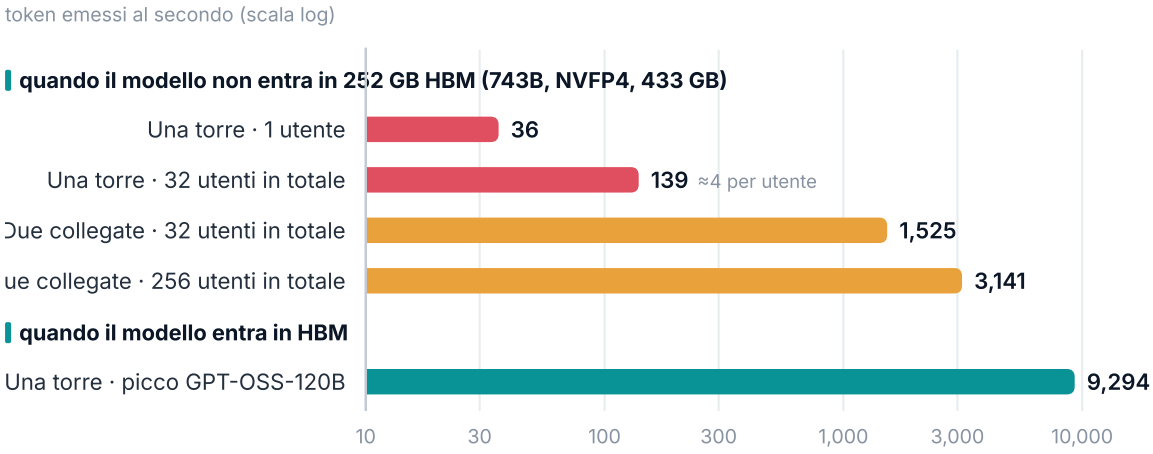
Jetson Thor rende praticabile la distribuzione a livello di stazione. I test TensorRT Edge-LLM di NVIDIA mostrano Qwen3-VL-8B (NVFP4) che fa prefilling di 2,048 token in 104.5 ms e decodifica a 44.4 token/s su Thor (TensorRT Edge-LLM) vendor claim ; per un ritaglio AOI e una risposta di 20 token sono circa 0.5 s con il modello 8B e 0.2 s con il 2B, ovvero 8.5–21 ritagli al secondo in batch estimate . È la lunghezza dell'output, non il numero di parametri, a fissare la latenza, quindi il decoding vincolato a un token di classe o a un JSON breve fa parte del progetto. GR00T N1.7 (3B) gira a 25.1 Hz su Thor (Jetson AI Lab) vendor claim . Un Orin NX impiega circa 1.8 s per ritaglio e un Atlas 300I Duo ha un p95 sopra i 23 s, adatto solo al lavoro batch (CSDN) single source . Nella Cina continentale, previeni i prodotti NVIDIA al 40–60% sopra il prezzo di listino statunitense estimate .

Un Mac Studio da 512 GB rende possibile un modello a grandezza piena su una sola macchina, per pochi utenti. L'M5 Ultra (agosto 2026) offre fino a 512 GB a 1.2 TB/s da circa $22,299 (Apple, TechRadar); DeepSeek 671B a 4 bit (≈380–404 GB) decodifica a 17–20 token/s su un M3 Ultra (Pinggy) verified e si stimano 25–29 sull'M5 Ultra. I

prompt lunghi sono il punto debole: un prompt di retrieval da 10,000 token richiede circa 17–70 s estimate, oltre il limite di interazione di 10 secondi.

DGX Station è la singola macchina più forte, ma non è disponibile nella Cina continentale e costa cinque-sei volte un Mac. Il GB300 in spedizione ha **748 GB**, di cui solo 252 GB sono HBM veloce (Wccftech, Tom's Hardware, ago 2026) verified. Test indipendenti mostrano circa 9,300 token/s di output su GPT-OSS-120B, che sta in HBM, ma un modello 743B che trabocca nella LPDDR5X ha dato a 32 utenti solo 139 token/s in totale, circa 4 ciascuno; due unità collegate hanno raggiunto 1,525 (StorageReview, ago 2026, StorageReview, ott 2026) single source. I prezzi di strada sono saliti a $123–140k entro novembre (CyberPowerPC, monitors.com). Nella Cina continentale, le opzioni realistiche per lo stesso lavoro sono il Mac Studio, server CPU+GPU in stile KTransformers (KTransformers) vendor test o server Ascend domestici.

FIG 09 **Una grande workstation non è un assistente condiviso** ***Quando il modello non entra nella memoria ad alta larghezza di banda, 32 utenti simultanei ottengono circa 4 token al secondo ciascuno.***



Fonte: StorageReview (ago. e ott. 2026); ServeTheHome. Test indipendenti da singola fonte.

La carenza di memorie ha spostato le stesse parti del 20–100% in un anno (DGX Spark è passato da $3,999 a $6,950 (TechRadar)), quindi i preventivi dell'attrezzatura vanno indicizzati piuttosto che fissati.

Gli acceleratori domestici fanno girare modelli a grandezza piena; la manutenzione del software è il collo di bottiglia. Ascend 910B ha 64 GB di HBM, ma vLLM-Ascend marca DeepSeek V3/R1 come "non più mantenuto" e supporta solo pochi modelli sulle schede più piccole (vLLM-Ascend) single source, primary, e ogni nuovo modello richiede uno-due mesi per essere adattato (Caijing/Anue); distribuisci una versione di modello congelata e validata. Servire DeepSeek 671B pieno a 20–100 utenti richiede almeno due server Atlas 800I A2 a otto schede (China Telecom Cloud) verified, eppure le gare pubbliche per appliance LLM in Cina sono ammontate a soli RMB 27.5 milioni in gennaio–settembre 2025 (C114) single source: le fabbriche pagano per integrazione e servizio, non per una scatola di calcolo.

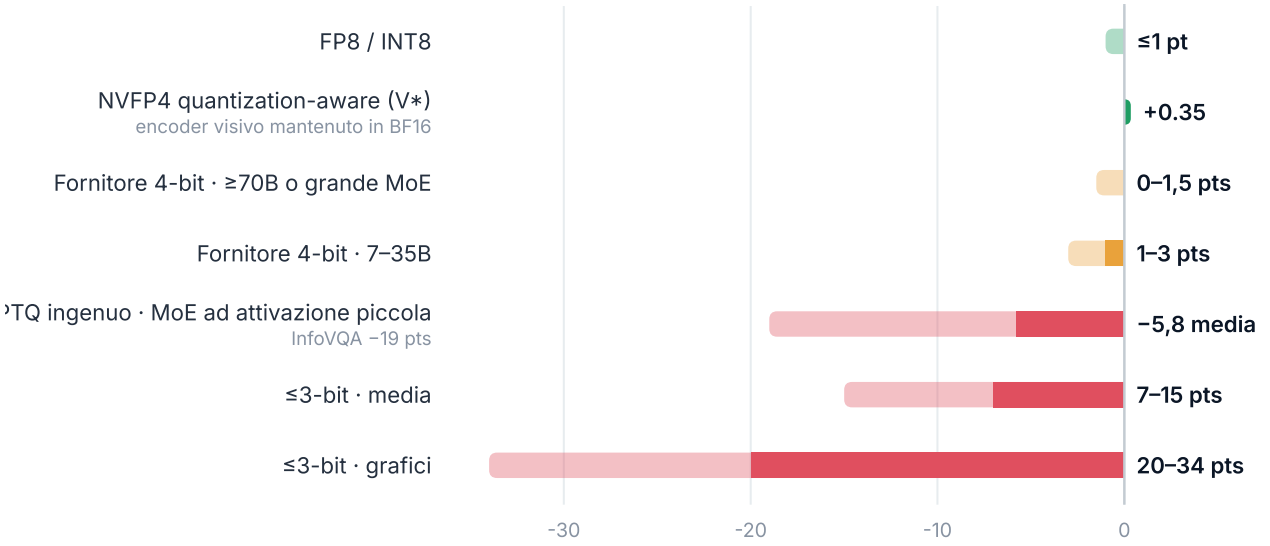
Quantizzazione: 8 bit è sicuro, 4 bit dipende dalla ricetta, 3 bit e sotto sono esclusi

8 bit (FP8 o INT8) è essenzialmente senza perdite per i modelli visione-linguaggio (Red Hat, Hugging Face) verified. A 4 bit, la perdita dipende da dimensione e ricetta: il build AWQ 72B di Qwen perde meno di un punto su MMMU mentre il 7B perde 2.8 (Qwen AWQ) verified, e i compiti ricchi di dettagli calano di più (il GPTQ 4 bit naif costa a Qwen3-VL-30B-A3B 19 punti su InfoVQA (VEQ)). A 3 bit e sotto, la perdita media è di 7–15 punti e i grafici perdono 20–34 (VLMQ) verified; oltre i 64k token, il 4 bit perde fino al 59%, peggio su input non inglesi (arXiv 2505.20276). Mantieni l'encoder di visione a FP16/BF16.

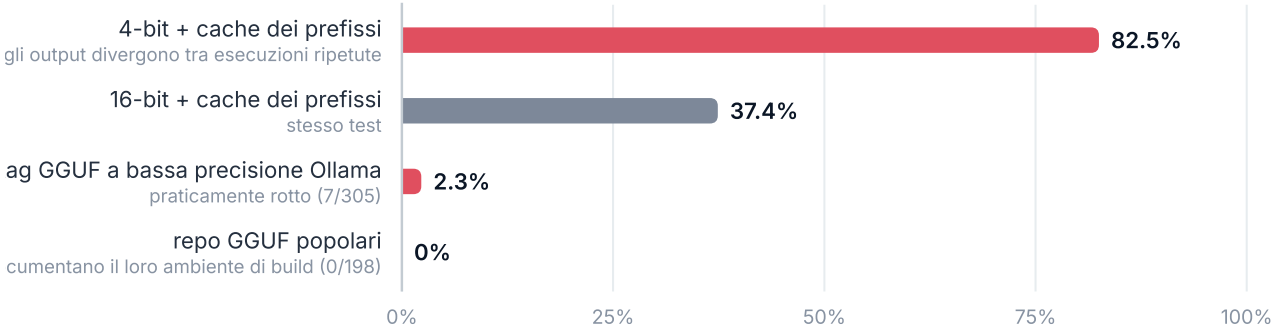
FIG 10 **Quantizzazione: 8-bit è sicuro, 4-bit dipende dalla ricetta, 3-bit e meno è escluso** ***Gli errori ingegneristici nello stack di serving costano più precisione della quantizzazione stessa.***

Precisione: 8-bit sicuro, 4-bit secondo la ricetta, ≤3-bit escluso

variazione del punteggio vs piena precisione (punti)



i guasti ingegneristici pesano più dell'errore di quantizzazione



Fonte: Scheda modello NVIDIA Nemotron, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 e altri.

Due avvertenze contano di più. Primo, **all'11 novembre 2026 nessuno studio o scheda modello ha misurato la quantizzazione su immagini di difetti PCB/SMT**, quindi rivalida sul golden set. Secondo, gli errori dello stack di serving sono spesso più grandi dell'errore di quantizzazione: lo stesso modello ha totalizzato circa 30 punti in meno sotto Ollama che sotto l'implementazione di riferimento (SiliconBench); 7 dei 305 tag ufficiali di modelli di codice di Ollama erano sostanzialmente rotti benché i file fossero intatti (arXiv 2609.05881); nessuno dei 198 repository GGUF popolari ha registrato il proprio build (arXiv 2610.19498); e con il riuso della prefix cache, l'82.5% degli episodi divergeva a 4 bit contro il 37.4% a 16 bit (arXiv 2609.04748) single sources. Il 4 bit quantization-aware del fornitore è vicino a senza perdite sui compiti di dettaglio (NVIDIA Nemotron 3.5) vendor claim, ma i VLA a 4 bit perdono 1.5–4.3 punti sotto perturbazione e girano più lenti su Thor (LoopQ) single source. Gestisci un modello quantizzato come un nuovo strumento di misura: fissa gli hash dei pesi, registra il build, azzera la prefix caching e rivalida dopo ogni cambio di precisione, motore o template.

Precisione	≥70B denso o ≥100B MoE	7–35B	≤4B o compiti ricchi di dettagli	Verdetto
FP8 / INT8	Sicuro (≤1 punto)	Sicuro	Sicuro; il default per l'ispezione	Usa

Precisione	≥70B denso o ≥100B MoE	7–35B	≤4B o compiti ricchi di dettagli	Verdetto
4 bit AWQ/GPTQ (ricetta fornitore, encoder di visione FP16)	0–1.5 punti	1–3 punti; GPTQ naif su MoE piccolo-attivo −5.8 in media, −19 su InfoVQA	2–3 punti; valida prima i compiti di dettaglio	Condizionato
NVFP4 (ricetta fornitore)	≈0.3–1 punto sotto FP8 (dati fornitore)	Build quantization-aware quasi senza perdite su V*, OCRBench	Nessun dato pubblico	Condizionato (preferisci build quantization-aware)
GGUF Q4_K_M	Non testato	Alta varianza sui compiti di dominio	Non per testi piccoli o difetti	Non per il giudizio di dettaglio
GGUF Q5_K_M e superiori	Presunto sicuro	Quasi senza perdite in due test	Il minimo per il dettaglio su llama.cpp	Usa
≤3 bit (inclusi modelli 1T compressi a 1.8–2 bit su una macchina)	Non per decisioni	No	No	No

Quale configurazione per quale lavoro

Configurazione	Ispezione in linea	Revisione AI dei falsi allarmi	Pianificazione e controllo robot	Q&A e analisi
Detector dedicati o modelli di anomalia (YOLO, RF-DETR, classe PatchCore), INT8/FP16, qualsiasi box edge NPU o GPU	● Sì: il giudice principale, 2–12 ms per ritaglio	● Sì (classificatore messo a punto sulla linea)	● Solo percezione	—
Modelli di visione 2–4B, NVFP4/AWQ/FP8 con encoder di visione FP16, su Thor/T4000 o DGX Spark	● Condizionato: ricontrolla i ritagli segnalati dall'AOI a 0.2–0.3 s ciascuno	● Condizionato, sotto 1 s dopo la messa a punto	● Backbone VLA (VLA 2–3B a ≈40–130 ms per chunk)	● Debole
Modelli di visione 7–9B (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 o 4 bit fornitore, su Thor, Spark o una scheda 24–32 GB	● Condizionato: ≈0.5 s per ritaglio segnalato	● Sì: 0.5–1 s per chiamata, miglior rapporto qualità-prezzo	● Pianificazione e scelta delle skill (1–3 s), non nel loop di controllo	● Q&A SOP di base per uno o pochi utenti
MoE 26–35B (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● No (troppo pesante)	● Sì, come spiegatore o per i casi difficili	● Pianificazione	● Sì (8 utenti su uno Spark, ≈0.1 s primo token)
MoE 106–241B a 4 bit su un Mac Studio, RTX PRO 6000 o Spark	● No	● No	● Pianificazione offline	● Sì, per uno o pochi utenti
DeepSeek 671B o Kimi K2.x 1T a 4 bit su un Mac 512 GB, server KTransformers o DGX Station	● No	● No	● Solo generazione offline di programmi	● Condizionato: 1–3 utenti e prompt brevi; i prompt lunghi richiedono 16–170 s

Configurazione	Ispezione in linea	Revisione AI dei falsi allarmi	Pianificazione e controllo robot	Q&A e analisi
Qualsiasi cosa a ≤3 bit; GGUF Q4_K_M per il lavoro di dettaglio su modelli piccoli; Ollama per molti utenti	● No	● No	● No	● No

Fonti: i dati TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research e KTransformers sopra, e (test DGX Spark, test RTX PRO 6000, Hardware Corner su Atlas, Ultralytics YOLO26, RF-DETR) **test community e fornitore, per lo più single source** .

Quindi, alle due preoccupazioni solite: **non deve essere lento** se il modello corrisponde al lavoro (i modelli di visione piccoli e medi su Thor o Spark soddisfano ogni budget in linea e di revisione; i modelli 671B–1T restano fuori linea), e **la quantizzazione non deve renderlo sbagliato** con FP8/INT8 di default, o 4 bit di ricetta fornitore sopra i 7B validato sul golden set. I rischi maggiori sono l'accuratezza zero-shot industriale dei modelli generali di circa 57 punti e gli errori dello stack di serving. Dai alle richieste di ispezione dimensioni di batch fisse e una coda propria (la temperatura 0 non è riproducibile a bit sotto batching (documentazione vLLM)), mantieni un hot standby N+1, e fai fallire in apertura ogni passo in linea.

Insegnante e studente

03 Insegnante e studente: i modelli frontier possono pre-etichettare i dati di ispezione; per i bracci robotici i veri insegnanti sono ancora le persone

Candidati insegnanti di novembre 2026: la visione generica è raddoppiata, mancano i punteggi sui difetti industriali

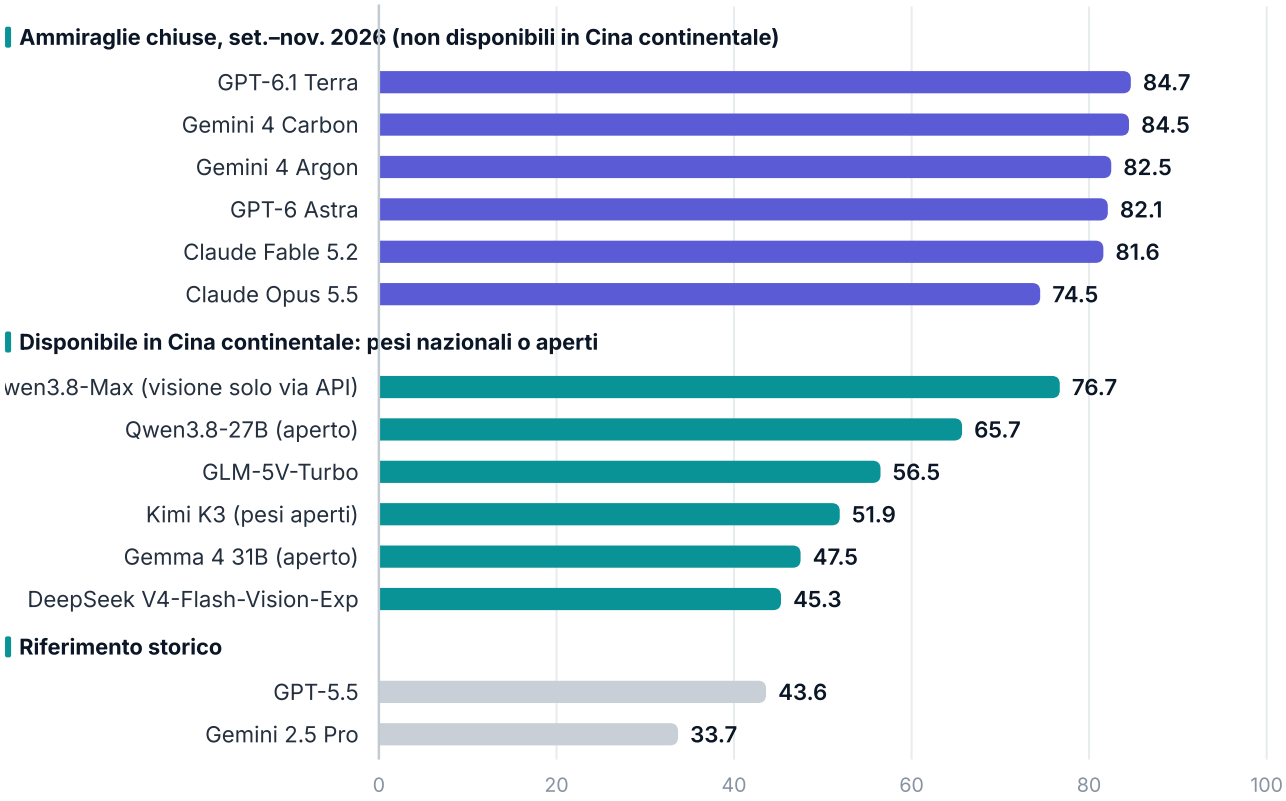
Ogni modello di punta chiuso che potrebbe fare da insegnante nel novembre 2026 accetta immagini: Claude Fable 5.2 (generalmente disponibile il 9 novembre, $8/$40 per milione di token) e Opus 5.5 ($4/$20) (Anthropic, panoramica modelli Claude) verified; GPT-6 Astra ($10/$50) e GPT-6.1 Terra ($1/$5) (prezzi OpenAI); e Gemini 4 Argon e Carbon, aperti a fasi (Google Carbon). Utilizzabili nella Cina continentale sono Qwen3.8-Max (visione via API) e modelli aperti come Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp e Gemma 4 (blog Roboflow, Hugging Face).

Il divario chiave va dichiarato con chiarezza: nessuno di questi modelli ha un punteggio pubblicato su MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB o qualsiasi dataset PCB/SMT. MMAD non ha ancora una leaderboard (MMAD GitHub), e i due paper più recenti sulle anomalie industriali (agosto–settembre 2026) usano ancora Gemini 2.5 Pro e Claude Sonnet 4 come baselines generali più recenti (InspectorGPT, Anomaly-LR). Una ricerca full-text su arXiv dell'11 novembre ha trovato zero occorrenze di "Fable 5.2" o "Gemini 4 Carbon" in questo contesto (API arXiv) single source, assenza di evidenza. I sostituti più vicini:

- Il modello frontier più recente con un punteggio MMAD è GPT-5.5: 69.95% su "difetto o no" e 76.20% mediato su sette compiti su un sottoinsieme di 500 immagini, contro 95.24% e 86.65% per gli esperti umani (AnomalyClaw) single source; GPT-4o aveva totalizzato 68.63% e 74.92% nel 2024 (MMAD, ICLR 2025), circa 1.3 punti di progresso in due anni.
- Su LDU-Bench, immagini SEM reali di revisione dalla litografia IC e l'unico benchmark di elettronica con modelli 2026, il triage buono-o-cattivo è decente (macro F1 0.932 per GPT-5.4, 0.884 per GLM-5V-Turbo, 0.876 per Claude Opus 4.6, 0.504 per Gemma 4 31B), ma classificare 11 morfologie di difetto raggiunge solo 0.20–0.41 (LDU-Bench) single source.
- Sulle valutazioni di visione indipendenti di Roboflow (aggiornate al 10 novembre 2026), la mAP@50 di object detection è 84.5 per Gemini 4 Carbon, 84.7 per GPT-6.1 Terra e 81.6 per Fable 5.2, contro 33.7, 43.6 e 38.6 per Gemini 2.5 Pro, GPT-5.5 e Claude Opus 4.8 (Roboflow Vision Evals) single source, valutatore indipendente, dataset privati.

FIG 11 **Candidati maestro: i punteggi generici sono raddoppiati in una generazione*Ma nessuna delle ammiraglie 2026 ha un punteggio pubblicato su un benchmark di difetti industriali.***

rilevamento generico Roboflow mAP@50



Benchmark di difetti industriali (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): zero punteggi pubblicati per queste ammiraglie 2
i punteggi di rilevamento generici sono raddoppiati in una generazione, ma non è giudizio di difetto; testate sul vostro set di riferimento

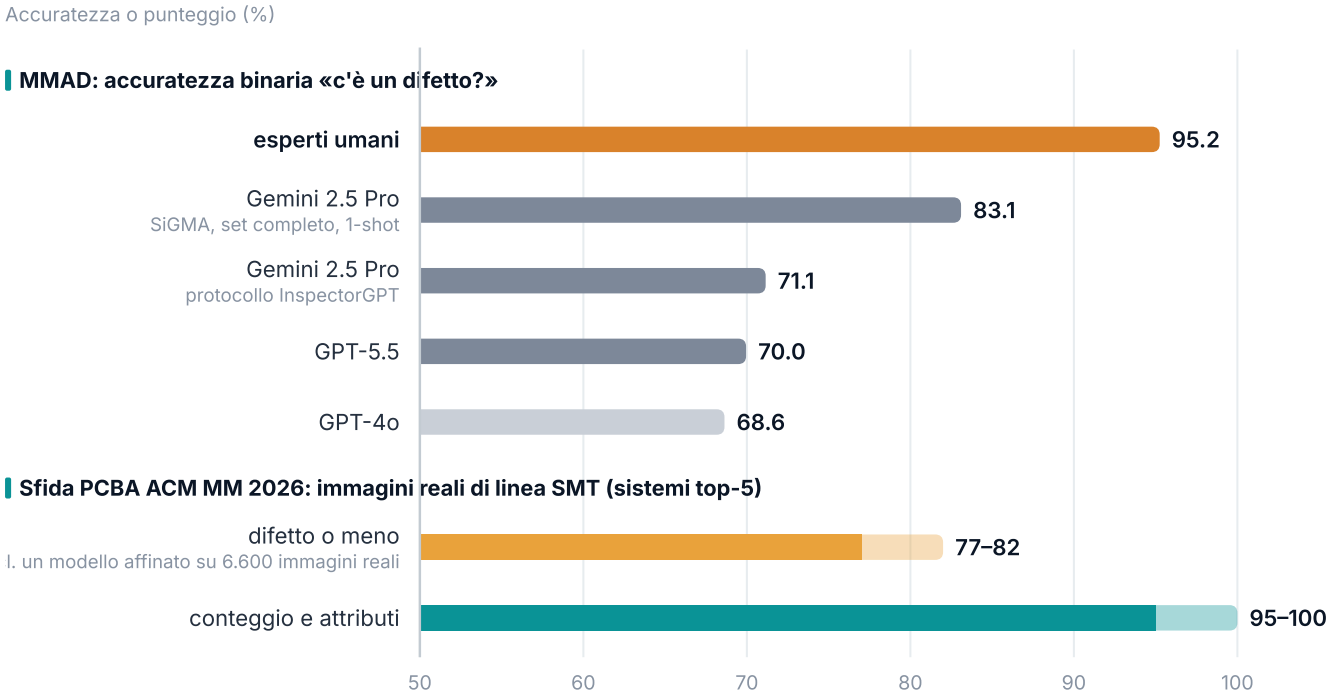
Fonte: Roboflow Vision Evals; note di rilascio dei modelli; pagine di MMAD, MVTec AD 2 e altri benchmark (all'11 nov. 2026).

Candidato insegnante (nov 2026)	Prezzo per M token (in/out)	Roboflow detection mAP@50 / conteggio	Benchmark difetti industriali	Utilizzabile nella Cina continentale
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Nessuno	No
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Nessuno	No
GPT-6 Astra	$10 / $50	82.1 / 80.2	Nessuno	No
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Nessuno	No
Gemini 4 Carbon	$2 / $10 (promozionale)	84.5 / 90.1	Nessuno	No
Qwen3.8-Max (visione solo via API)	$2 / $6	76.7 / 81.1	Nessuno	Sì
Qwen3.8-27B (aperto)	Locale	65.7 / 64.9	Nessuno	Sì
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench triage 0.884	Sì
Gemma 4 31B (aperto)	Locale	47.5 / 51.4	LDU-Bench triage 0.504	Sì (pesi locali)
Riferimento: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD binario 69.95% / 68.63%	No

Candidato insegnante (nov 2026)	Prezzo per M token (in/out)	Roboflow detection mAP@50 / conteggio	Benchmark difetti industriali	Utilizzabile nella Cina continentale
Esperti umani	—	—	MMAD binario 95.24%	—

I paper più recenti testano ancora la generazione precedente, ma indicano la stessa direzione. Sul set MMAD completo, gli esperti umani totalizzano il 95.2% su difetto-o-no e Gemini 2.5 Pro l'83.1% (SiGMA), eppure lo stesso modello ha totalizzato solo il 71.14% in un altro paper (InspectorGPT): il protocollo di test sposta il punteggio quanto una generazione di modello. Nella sfida CVPR 2026 VAND 4.0, i modelli visione-linguaggio pronti all'uso hanno raggiunto un picco di 61.88 AP contro 90.84 per i metodi supervisionati, e gli organizzatori hanno concluso che "non possono attualmente sostituire i classificatori fine-tuned" (VAND 4.0) single source. Nella sfida PCBA di ACM MM 2026, i primi cinque hanno totalizzato 76.92–82.47 su difetto-o-no mentre il tipo di componente e il conteggio dei pin hanno raggiunto 95–99.5 (arXiv 2609.21276) single source: riconoscere è molto più facile che giudicare.

FIG 12 **«C'è un difetto?» resta difficile per i modelli generali** ***Il miglior modello generale è oltre 12 punti dietro gli esperti umani su MMAD; i sistemi affinati fanno meglio su immagini di linea reali.***



Fonte: MMAD; SiGMA; InspectorGPT; sfida PCBA ACM MM 2026.

Ne seguono tre conclusioni. Vedere e contare sono raddoppiati in una generazione, ma giudicare l'anomalia si muove lentamente, dipende pesantemente dal protocollo e porta pregiudizi sistematici; quanto guadagnano i nuovi modelli si può misurare solo sulle immagini di uno stabilimento. Dentro la loro distribuzione di addestramento, gli specialisti 7B fine-tuned battono ancora i modelli generali (Anomaly-LR, UniPCB). E i dati di linea reali fanno crollare i punteggi: sui pannelli elettronici assemblati, l'AUROC a livello di immagine è sceso da 97–98 su MVTec AD a 31.5–78.6 (arXiv 2606.01992) single source. Per le aziende della Cina continentale il più forte insegnante disponibile, Qwen3.8-Max, sta circa 8 punti sotto il frontier globale.

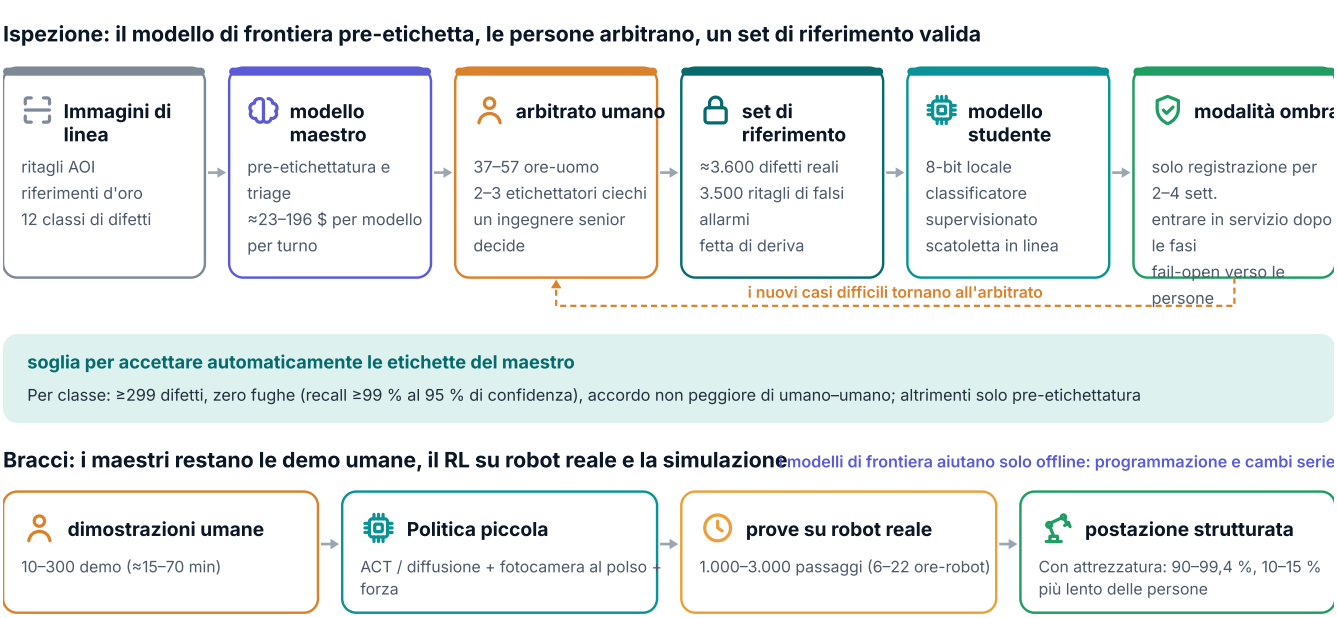
Fare decine o centinaia di round: il denaro non è il problema, la qualità delle etichette sì

Il denaro non è il problema. Ai prezzi di listino di novembre 2026, 1,000 immagini da un megapixel costano circa $3 (GPT-6.1 Terra) a $32 (GPT-6 Astra) per round, e cento round su 100,000 immagini circa $32,000–135,000 prima

degli sconti sul volume estimate (documentazione visione Claude, guida visione OpenAI, Roboflow Vision Evals). Per etichette semplici una persona costa all'incirca quanto Terra; il vantaggio dell'insegnante è velocità e scala, non il prezzo. Gli NDA dei clienti di solito tengono le immagini di prodotto dentro lo stabilimento, quindi il golden set è etichettato in loco.

Il vero ostacolo è la qualità, e più round non correggono gli errori sistematici. Roboflow esegue ogni compito tre volte e i punteggi si muovono solo di ±0.1–1.4 punti (Roboflow), quindi un modello ripete per lo più i suoi errori e il voto a maggioranza non può salvarlo. I round aiutano solo quando ognuno aggiunge informazione: una o due buone immagini di riferimento aggiungono circa 3–4 punti, di più non aggiungono nulla (AD-Copilot); far discutere un modello contro un buon riferimento su più turni aggiunge 6.23 punti di AUROC cross-dominio per GPT-5.5 e 2.95 su MMAD, con picco dopo circa tre chiamate (AnomalyClaw) single sources. "Uno studente addestrato su etichette dell'insegnante si avvicina alle etichette umane" è stato dimostrato solo su immagini naturali: una ResNet-101 addestrata su etichette Qwen3-VL di COCO e Objects365 ha raggiunto il 95.7–99.4% del risultato con etichette umane, e gli autori affermano che i dati di dominio non sono testati (arXiv 2602.20972) single source.

FIG 13 **Due pipeline maestro–studente** ***Per l'ispezione un modello di frontiera pre-etichetta e le persone arbitrano; per i bracci robot i veri maestri restano le dimostrazioni umane.***



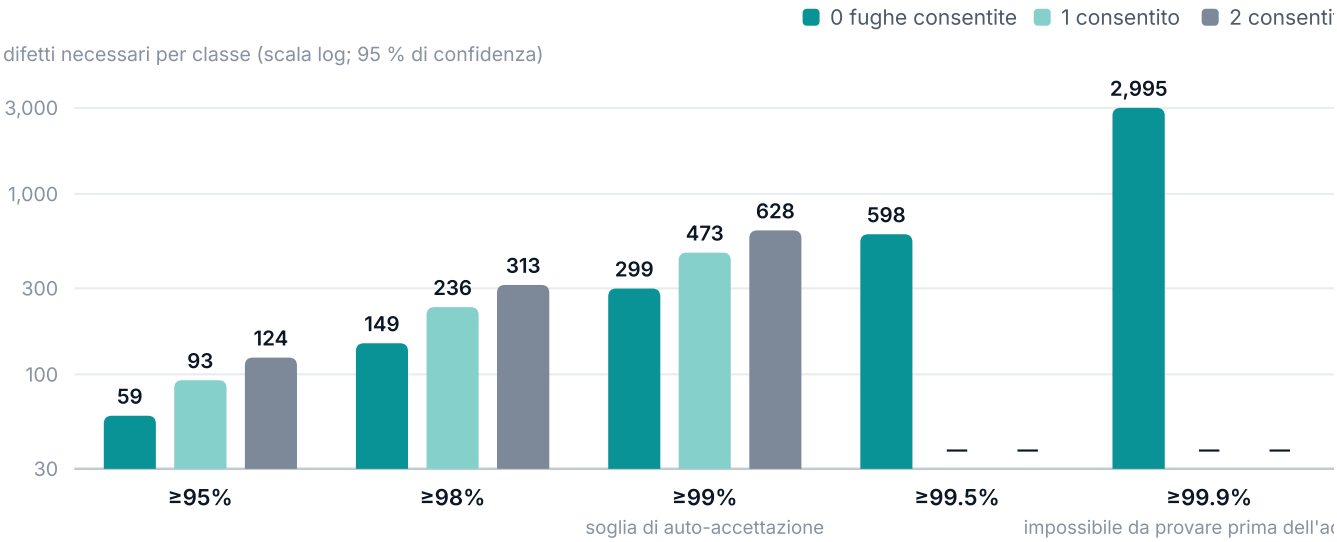
Fonte: Sintesi dell'autore (capitolo 3); costi e dimensioni del campione sono stime.

Un protocollo praticabile estimate:

1. **Raccogli il golden set dall'archivio AOI dello stabilimento e dalle schede NG conservate**, stratificato in 12 classi di difetto reali (mancante, offset, tombstone, polarità, parte sbagliata, ponte/cortocircuito, saldatura insufficiente, saldatura in eccesso, sfera di saldatura, terminale sollevato/aperto, parte danneggiata, materiale estraneo) e quattro cause di falso allarme (ombra, flussante, serigrafia, tolleranza), più una "fetta di deriva" di almeno 300 difetti da una seconda linea o con illuminazione, colore della scheda o codici articolo diversi.
2. **Etichetta in cieco con due o tre ispettori addestrati IPC-A-610**, con un ingegnere senior che decide e raggi X, test elettrico o sezioni che risolvono i campioni ambigui. Accetta una definizione di classe solo se l'accordo tra ispettori raggiunge κ ≥0.75 (≥0.90 per le classi critiche); l'accordo umano su 11 morfologie di difetto elettronico in LDU-Bench era solo 0.727, quindi alcune classi dovranno essere unite o ridefinite. L'accordo per attributi in stile AIAG (efficacia ≥90%, mancati ≤2%, falsi allarmi ≤5%, κ >0.75; >0.90 per il rilascio) è un minimo sensato (Fabrico, iFactory).
3. **Dimensionalo con limiti binomiali esatti (Clopper–Pearson) unilaterali al 95%** (intervalli di confidenza binomiali, regola del tre:

Recall da dimostrare	0 mancati ammessi	1 ammesso	2 ammessi
≥95%	59 difetti	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **Quanto deve essere grande il set di riferimento*Provare il 99 % di recall con zero fughe richiede 299 difetti per classe; il 99,9 % non può essere provato prima dell'acquisto.***



Fonte: Intervalli di confidenza binomiali esatti (95 %).

Sessanta difetti per classe bastano per uno screening passa/non passa; ne servono 300 per decidere l'accettazione automatica. Il set completo, circa 3,600 difetti più 3,500 ritagli di falsi allarmi, richiede circa 37–57 ore-uomo per etichettare e arbitrare, e far girare un candidato insegnante su di esso costa $23–196; il costo reale è di due-sei settimane di raccolta, fissate dalla classe più rara estimate. Invia ritagli AOI di singolo componente, non schede intere, con il bersaglio ad almeno 42–48 pixel di lato (arXiv 2610.09313).

1. **Accetta le etichette dell'insegnante automaticamente solo per una classe con zero errori su ≥299 difetti golden**, accordo modello-verità non peggiore dell'accordo umano-umano, e almeno il 50% dei falsi allarmi chiusi. Le classi critiche come cortocircuiti, parti mancanti, polarità e parti sbagliate richiedono il 99.5–99.9% (598–2,995 difetti); altrimenti qualsiasi cosa l'AI non può chiudere con alta confidenza va a una persona. Un tasso di errori sfuggiti sotto lo 0.1% dei difetti non può essere dimostrato prima dell'acquisto (2,995 difetti con zero errori; lo 0.01% richiederebbe 29,956), quindi la linea mantiene il campionamento umano delle chiamate chiuse dall'AI, l'ICT/FCT a valle come rete di sicurezza e SPC sugli errori sfuggiti. Le etichette verificate vengono poi distillate in un detector o classificatore convenzionale e rivalidate su ogni nuovo prodotto.

Anche i termini e le regole sui dati decidono l'insegnante. Anthropic vieta di usare i suoi servizi per addestrare modelli concorrenti (termini commerciali Anthropic); OpenAI consente di usare gli output per addestrare classificatori interni non venduti a terzi (termini commerciali OpenAI); DeepSeek consente esplicitamente la distillazione (termini DeepSeek) primary terms. Nessuno definisce un "modello concorrente", quindi l'uso commerciale richiede revisione legale, e le regole sui dati dei clienti di solito tengono le immagini in loco. Il default più sicuro è un insegnante a pesi aperti eseguito localmente, con i modelli cloud frontier solo quando il cliente lo consente per iscritto.

Studenti di ispezione: piccoli e veloci; i colli di bottiglia sono ottica, dati e validazione

Lo studente non è il collo di bottiglia: YOLO26n impiega 3.80 ms per immagine in INT8 su un Orin Nano Super da $249–399 (Ultralytics) vendor test, NVIDIA riporta il 98.51% su tre classi di difetto PCB dopo aver adattato un modello fondazionale su circa 600 immagini (NVIDIA) vendor claim, e i difetti sintetici aggiungono solo 0.9–2.5 punti di mAP (arXiv 2605.04635). I vincoli sono ottica, dati di linea, validazione e licenze: Ultralytics YOLO è AGPL-3.0 o con licenza enterprise (Ultralytics), quindi i prodotti chiusi hanno bisogno di una licenza o di un detector Apache-2.0 come RF-DETR.

Studenti robot: le stazioni strutturate sono vicine al livello produttivo; il lavoro flessibile e dinamico no

Nel novembre 2026 ci sono tre tipi di insegnante robot, e solo uno può diventare uno studente locale: VLA aperti e fine-tunabili (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). I modelli robot chiusi (π0.6/π0.7 via API beta, Gemini Robotics 2 per tester fidati) non possono essere distribuiti in loco (The AI Insider, DeepMind). I modelli generali frontier sono lenti e scarsi nella manipolazione fine: nel test di Artificial Analysis hanno totalizzato 46–69 con il 28–45% di successo nella presa e circa 30 minuti per compito, “troppo lenti e costosi per qualsiasi uso produttivo” (AA-Robotics); GPT-6 Astra è riuscito in una slot di precisione solo 2 volte su 20 (RoboCurve); e comandare direttamente i giunti dà 0–5.5% su LIBERO a 0.2–0.4 Hz contro circa 83 Hz necessari (Anthropic) verified su tre fonti. Il 64.9% di Fable 5.2 sul benchmark simulato RoboDojo è un’esecuzione interna (Anthropic) vendor claim.

Gli studi di terze parti da ottobre concordano: circa 33–44% su LIBERO-10 per i tre agenti principali (arXiv 2610.30051); agenti che dichiarano falsamente il successo sul 24–39% dei sotto-compiti (arXiv 2611.08209); e 104 di 108 compiti quasi-statici su robot reale per Opus 5.5, ma a 280 secondi ciascuno e senza inserimento o assemblaggio (AMP1) single sources. I modelli frontier appartengono all’offline: scrivere programmi, orchestrare compiti e redigere criteri di successo.

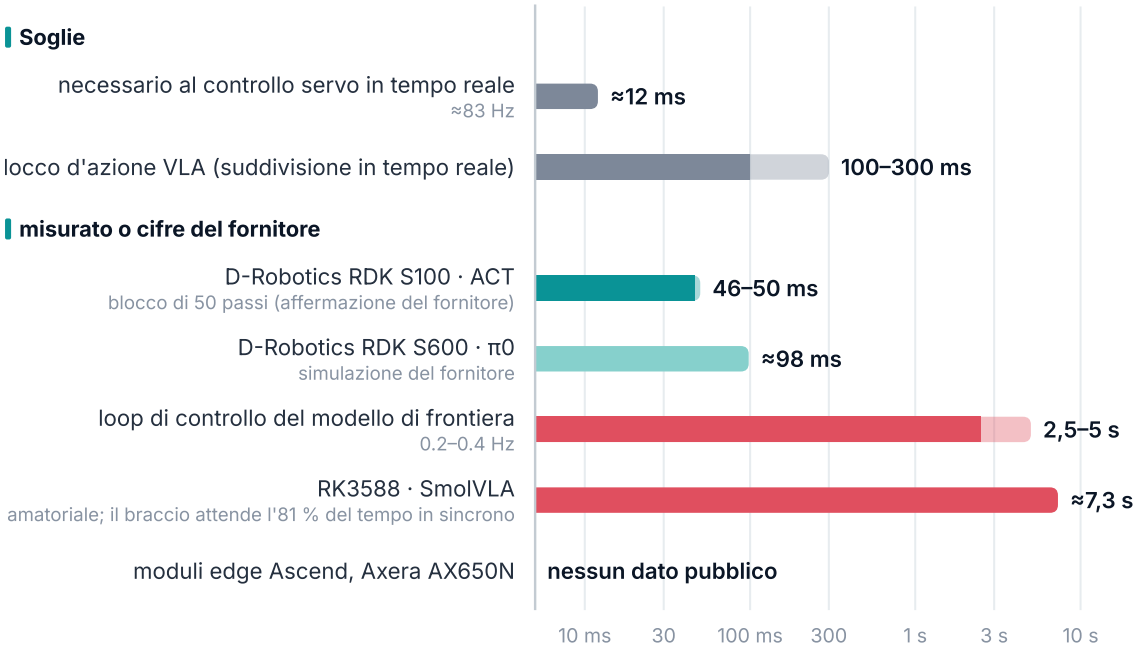
La qualità degli studenti ha un margine netto: buona dentro la distribuzione di addestramento, in crollo fuori. π0.7 “spesso supera il 90%” sui compiti visti e raggiunge il 60–80% su quelli mai visti (paper π0.7) vendor research; ContactDP raggiunge il 97.8% sui connettori addestrati ma il 28.6–66.7% su quelli mai visti (ContactDP); lievi cambi di fotocamera o posa fanno scendere il 95% sotto il 30% in LIBERO-Plus (LIBERO-Plus, CVPR 2026); la presa appresa da nastri in movimento ha un picco del 58.89% (ForeTime-VLA) single sources, un lavoro che il classico tracking visivo fa in produzione da anni.

Le evidenze quasi-produttive vengono tutte da stazioni strutturate, dimostrazioni umane e correzione su robot reale. Neuromeka ha mantenuto una macchina a stati e movimento pre-insegnato e ha appreso solo presa, inserimento cavi e saldatura, ciascuno da meno di 20 minuti di dati: 322 su 324 in produzione, il 12.8% più lento di una persona (Neuromeka) single source, vendor. Fraunhofer ha raggiunto il 92.8% su quattro connettori da 68 minuti di dimostrazioni (Fraunhofer); HIL-SERL di Berkeley è riuscito 100 volte su 100 nell’inserimento di RAM, SSD e USB dopo 1–2.5 ore di RL su robot reale (HIL-SERL); AgiBot ha eseguito 18–20 secondi per passo dentro un takt di 20 secondi su una linea di test per tablet (IThome) vendor claim.

La distillazione compra velocità, non generalizzazione: π0.5 gira a 0.70 Hz su Jetson Orin grezzo e a 10.30 Hz dopo scheduling asincrono (EagleVLA); la distillazione a strati lo porta da circa 3 Hz a circa 10 Hz (Shallow-π) verified. Uno studente eredita solo ciò che l’insegnante ha coperto; ogni nuovo prodotto richiede ancora dati di compito.

FIG 15 **Latenza di politica edge** ***Una piccola politica ACT gira in ≈50 ms su una scheda edge; un VLA portato su un chip amatoriale impiega 7,3 s.***

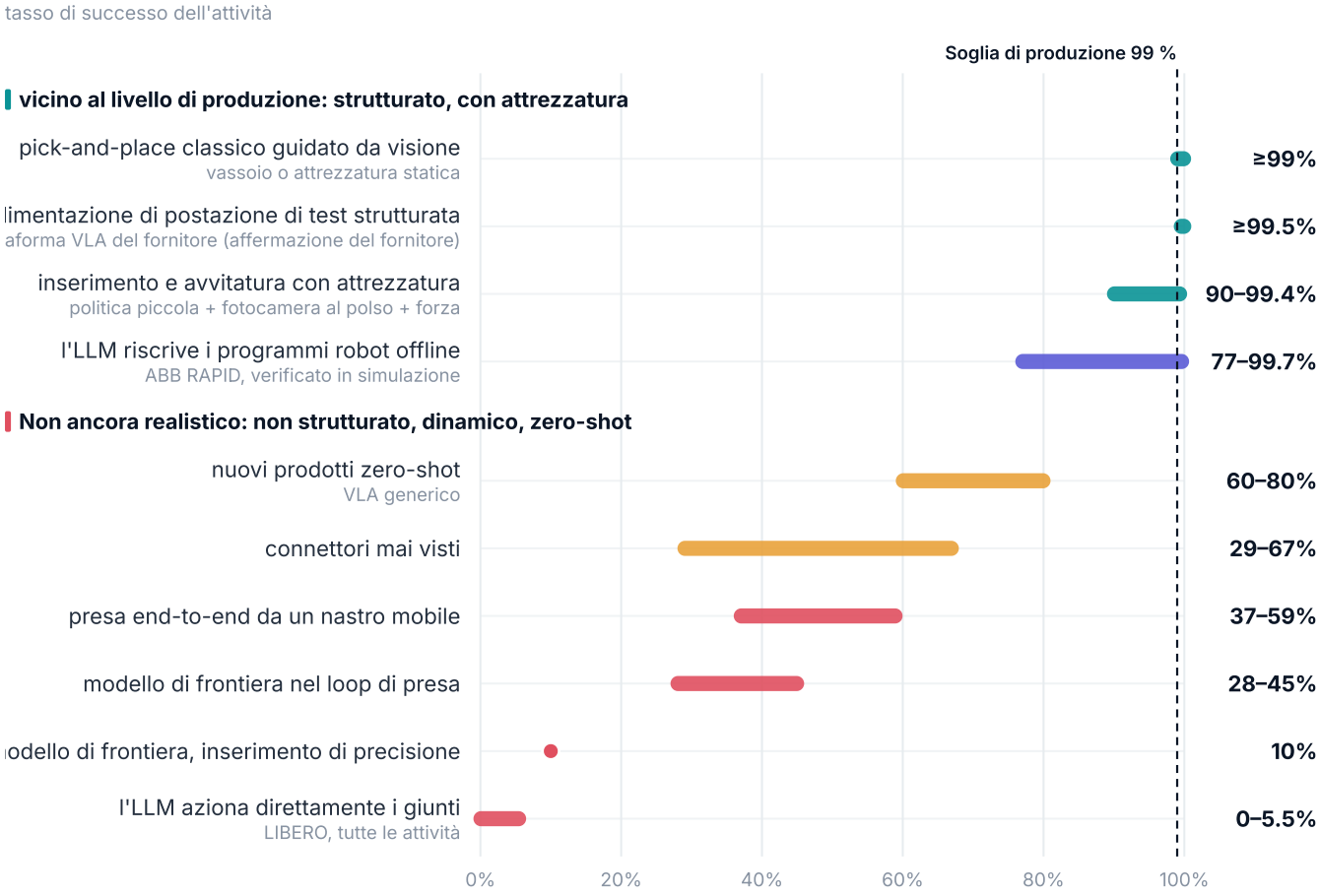
latenza di inferenza per blocco d'azione o ciclo di controllo (scala log)



Fonte: Documentazione e forum D-Robotics (cifre del fornitore); port comunitari GitHub; ricerca Anthropic.

Sui chip edge cinesi, l'unica policy appresa misurata è una policy ACT su RDK S100 di D-Robotics a 46–50 ms per chunk di 50 passi, senza tasso di successo pubblicato (documentazione D-Robotics RDK S) vendor claim; un port hobby di SmolVLA su un RK3588 impiega circa 7.3 s per chunk (GitHub). La quantizzazione non compra velocità gratis: nessun build π0.5 a 4 bit era più veloce del 16 bit a batch size 1 (LoopQ). Le dimostrazioni sono costose (Reuters ha riportato circa 300 tentativi per movimento utilizzabile per un teleoperatore novizio (Notebookcheck)), e la validazione costa di più: dimostrare un tasso di guasto ≤0.1% al 95% di confidenza richiede circa 3,000 successi consecutivi, 17–22 ore-robot estimate. Nessuna policy di manipolazione appresa ha certificazione di sicurezza.

FIG 16 **Studenti robot: vicini al livello di produzione solo se la postazione è strutturata***Le postazioni con attrezzatura raggiungono 90–99 %+; i compiti non strutturati, dinamici e zero-shot non sono ancora realistici.*



Fonte: Articoli e materiale dei fornitori citati nel capitolo 3; alcune cifre sono affermazioni del fornitore.

Compito elettronico	Studente praticabile	Successo atteso	Velocità vs una persona	Costo per cambio	Verdetto 2026–2028
Pick-and-place ad alta varietà, kitting, alimentazione macchine (vassoio o maschera statici)	Percezione appresa più movimento programmato o pianificato	≥99% con guida visiva classica	Come le stazioni convenzionali	Ore	Fattibile ora
Connettore, FFC e inserimento cavi su maschera, avvitatura, saldatura	Macchina a stati + piccola policy BC/diffusion/ACT + fotocamera da polso + forza/coppia (idealmente tocco) + correzione umana	90–99.4%	10–15% più lento	10–300 demo (≈15–70 min) + 1,000–3,000 run di validazione (6–22 ore-robot)	Fattibile, R&D intensivo; mesi per la prima stazione
Tending di stazione di test strutturata e smistamento OK/NG	Piattaforma VLA fornitore + gemello digitale + RL su robot reale	≥99.5% vendor claim, contrastante	18–20 s per passo, dentro un takt di 20 s	≤4 h di riaddestramento + 15 min di calibrazione	Fattibile su piattaforma fornitore
Programmazione dei cambi	LLM scrive o modifica programmi robot offline, verificati in simulazione	77–99.7% su riscritture ABB RAPID	Nessun effetto sul takt	Da minuti a ore per programma	Fattibile come strumento ingegneristico

Compito elettronico	Studente praticabile	Successo atteso	Velocità vs una persona	Costo per cambio	Verdetto 2026–2028
Prelievo appreso end-to-end da nastri in movimento	Studente VLA	37–59%; ≤37% su nastri veloci	—	≥458 demo per setup	**Non prima del 2028; usa tracking classico più percezione appresa**
Nuovi prodotti senza nuovi dati (zero-shot)	VLA generico	60–80%; connettori mai visti 29–67%	—	—	**Non realistico**
LLM frontier nel loop che comanda il braccio	Agenti GPT-6, Claude, Gemini	Presa 28–45%; inserimento di precisione 10%	≈30 min o 280 s per compito vs ≈20 s	—	**Non realistico**
Cablaggi a percorso libero e assemblaggio flessibile multi-passo	—	Nessuna evidenza di produzione 2026	—	—	**Non prima del 2028**

Fonti: gli studi sopra e (University of Passau sulla riscrittura RAPID, GraspVLA, Komax) **per lo più single source e dati fornitore** . Nel 2026–2028, il valore di un grande modello per i robot è come priore pre-addestrato e programmatore offline, non come qualcosa che "insegna" a uno studente. Ogni caso quasi-produttivo che ho trovato (Neurometa, AgiBot da Longcheer, Fraunhofer, ContactDP) si basa su dimostrazioni umane, rilevamento di forza o tocco, simulazione o gemelli digitali e RL supervisionato. Nessuno si basa su etichette di un modello frontier.

Costi e rientro

04 Costi e rientro: la revisione AI rientra in mesi, l'alimentazione macchine in anni, i bracci di prelievo mai

Costo del lavoro: gli Stati Uniti sono circa 5.5 volte Shenzhen, Taiwan circa 1.5 volte, gran parte del Sud-est asiatico meno della metà

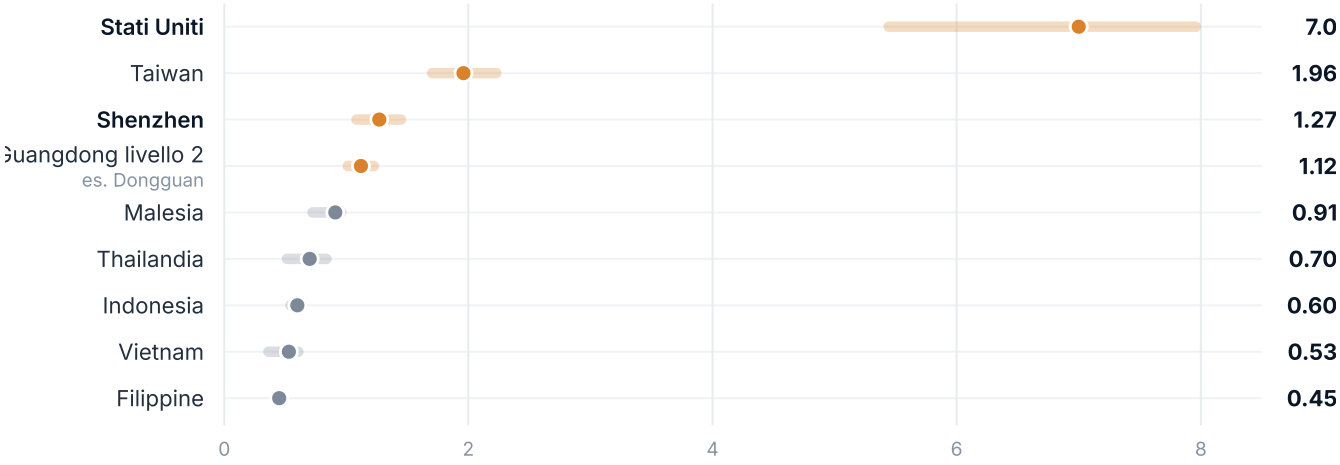
Il ritorno sulla sostituzione del lavoro parte da quanto costa il lavoro. La guida salariale ufficiale di Shenzhen pone la mediana 2024 di un ispettore qualità manifatturiero a **RMB 74,074** l'anno; i salari sono l'86.92% del costo totale del lavoro nella produzione elettronica, quindi il costo pienamente caricato è di circa RMB 85,000. L'indagine JETRO 2025 pone il costo annuo di un operaio cinese a $12,667, circa RMB 85,100, entro lo 0.2% (guida salariale Shenzhen, JETRO 2025) verified .

Regione	Costo annuo pienamente caricato di un ispettore o addetto (basso / base / alto)	Base
Cina continentale, Shenzhen	RMB 70k / **85k** / 100k ($10.4k / **12.7k** / 14.9k)	Mediana ufficiale × 1.15; combacia con JETRO verified
Aree salariali di livello 2 del Guangdong (Dongguan ecc.)	RMB 65k / **75k** / 85k	Mediana Guangdong 2023 × 1.15 più due anni di aumenti estimate
Delta del Fiume Azzurro (Suzhou, Kunshan, Wuxi)	Come Shenzhen	Nessuna base per manodopera di ispezione più economica; il salario minimo massimo del Jiangsu è circa il 5.6% sopra quello di Shenzhen estimate
Taiwan	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	Indagine salariale occupazionale del Ministero del Lavoro più ≈1.18 oneri legali; JETRO come estremo alto estimate
Malaysia	$6.8k / **9.1k** / 10.0k	JETRO e salari minimi estimate
Thailandia	$4.7k / **7.0k** / 8.8k	Idem
Indonesia	$5.0k / **6.0k** / 6.3k	Idem
Vietnam	$3.2k / **5.3k** / 6.5k	Idem
Filippine	$4.0k / **4.5k** / 4.9k	Idem
Stati Uniti	$54k / **70k** / 80k	Mediana ispettori BLS maggio 2025 $48,570 ÷ salari al 65.5% della retribuzione estimate

Fonti: (JETRO 2025, guida salariale Guangdong, ufficio risorse umane Wuxi, indagine salariale occupazionale Taiwan, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **Quanto costa all'anno un addetto a ispezione o movimentazione** ***A pieno carico, gli USA sono ≈5,5× Shenzhen e Taiwan ≈1,5×.***

costo annuo caricato per addetto CQ o movimentazione (10k $; punto = base, barra = basso–alto)



Fonte: Guida salariale ufficiale di Shenzhen, Ministero del Lavoro di Taiwan, JETRO, BLS; stime dell'autore.

Tre correzioni fissano la base. Primo, Shenzhen ha alzato il minimo contributivo pensionistico da RMB 4,775 a 5,521 al mese a luglio 2026 (Shenzhen Bendibao) single source, quindi un ispettore di Shenzhen nel 2026 costa più probabilmente RMB 80–90k. Secondo, un posto non è una persona: due turni da 12 ore richiedono circa 2.0 persone (sei giorni a settimana) a 2.4 (tutto l'anno) nel Delta del Fiume delle Perle, circa 2.7–3.1 nel Sud-est asiatico sotto i tetti di straordinario, e 4.2–4.5 per 24×7 negli Stati Uniti e a Taiwan estimate. La revisione AI di solito consente a un revisore di coprire da due a quattro linee, quindi il risparmio realistico è **0.5–1.5 persone a tempo pieno per linea, non una persona per linea per turno**. Terzo, gli acquirenti a Taiwan, Vietnam e Malaysia non inseguono solo il costo: nell'indagine JETRO, il 61.1% delle imprese a Taiwan, il 48.2% in Vietnam e il 44.9% in Malaysia ha detto che assumere era diventato più difficile negli ultimi due anni, contro il 18.6% nella Cina continentale (JETRO 2025) single source.

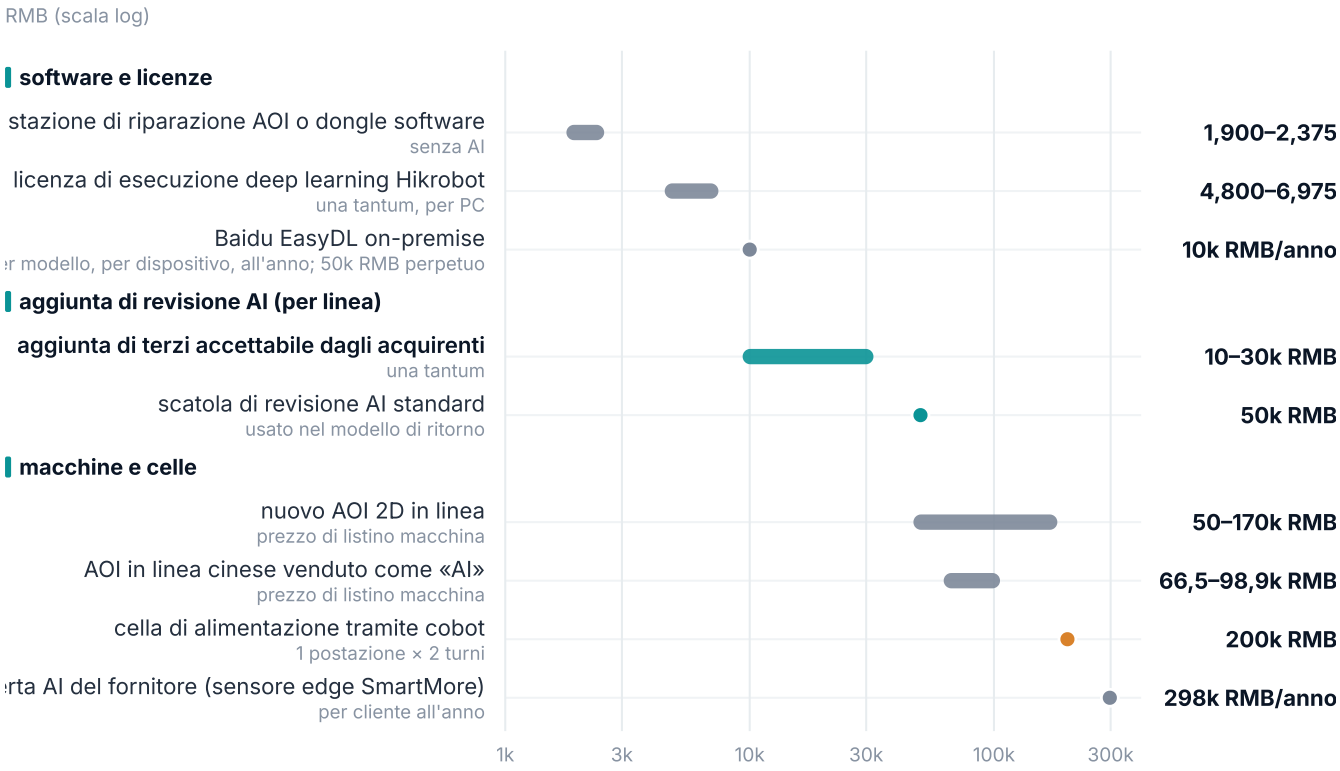
Prezzi dell'attrezzatura: quattro basi che non vanno mediate

I prezzi degli AOI divergono soprattutto perché misurano cose diverse. Le inserzioni online (novembre 2026, miste nuove, usate e a noleggio) vanno da circa RMB 17–80k per 2D offline o desktop, RMB 50–170k per 2D in linea e RMB 80–475k per 3D in linea, fino a RMB 700k per 3D coreano (1688 AOI in linea, 1688 AOI offline, 1688 AOI 3D) single source, prezzi di listino. Una gara di novembre 2025 ha fissato il tetto per un AOI 3D più un SPI 3D a RMB 800k (avviso di gara). Il prospetto di Sinic-Tek implica RMB 428–512k per AOI 3D nel 2019–2022, circa RMB 340–470k secondo la sua relazione 2025 (prospetto Sinic-Tek, relazione annuale Sinic-Tek 2025) primary + estimate. Un fornitore estero stima $3,000–20,000 per 2D desktop, $30–70k 2D in linea, $50–90k 3D entry e $100–200k+ 3D di fascia alta, con la stessa macchina in linea a $30–50k in Cina (Averroes) vendor estimate da un venditore di software AI concorrente.

La revisione AI in sé non ha un prezzo pubblico in RMB; i fornitori quotano caso per caso dopo una valutazione gratuita in loco (Sohu su Hengtianyi), e le ricerche su inserzioni e-commerce, gare, articoli di fornitori e media taiwanesi non hanno prodotto alcuna quotazione pubblica (1688, 51SMT). Le ancore: circa €27k per linea impliciti dal caso Siemens estimate; $21,924 di listino per il modulo di rete di revisione remota di un distributore statunitense (Manncorp) vendor list price; circa RMB 298k per cliente all'anno per i sensori edge AI di SmartMore nel 2025 (36Kr) single source. Un AOI 2D in linea domestico venduto come "AI" ha un listino di RMB 66.5k (1688 AOI in linea), e un buffer NG per la re-ispezione manuale circa RMB 50k (1688 revisione AOI). Quindi il tetto per un add-on di revisione di terze parti è di circa RMB 10–30k per linea, o una tariffa annuale simile; sopra, l'acquirente preferirebbe sostituire la macchina con un AOI domestico con AI integrata estimate.

Ancora	Prezzo (RMB)	Cosa compra	Fonte
Stazione di riparazione AOI o dongle software (TRI, Zhenhuaxing più vecchi)	≈1,900–2,375	Chiave di licenza, senza AI	1688 [prezzo di listino]
Runtime di visione deep-learning generico (dongle Hikrobot VisionMaster)	≈4,800–6,975 una tantum per PC	La piattaforma; il cliente costruisce e mantiene i modelli	1688 [prezzo di listino]
Distribuzione on-premises Baidu EasyDL	10,000 per modello per dispositivo all'anno, o 50,000 perpetui	Servizio di inferenza confezionato	documentazione Baidu AI [pagina prezzi 2022]
AOI in linea domestico venduto come "AI"	66,500–98,900 per macchina	Sostituzione della macchina	1688 [prezzo di listino]
Sensori edge AI SmartMore	≈298,000 per cliente all'anno	Soluzione fornitore su più macchine	36Kr single source

FIG 18 **La scala dei prezzi***Da poche migliaia di RMB di licenza software a 200–300k RMB per una cella robotica o un'offerta AI del fornitore.*



Fonte: Annunci 1688, documentazione Baidu AI (listino 2022), 36Kr, un'intervista del settore (non verificata).

Una tariffa annuale in stile Baidu, con un server di inferenza per una-quattro linee e uno-quattro modelli ciascuno, implica un abbonamento di RMB 2.5–40k per linea all'anno, realisticamente circa RMB 5–10k; al 20% di una licenza perpetua, abbonamento e acquisto pareggiano al quinto anno, una struttura che vale la pena copiare estimate. Gli acquirenti sono abituati a tariffe software quasi nulle: l'intero fatturato "altro" di Sinic-Tek (ricambi, servizio, noleggio, accessori) era il 7.65% delle vendite 2025, circa RMB 2,300–3,100 per macchina installata all'anno (relazione annuale Sinic-Tek 2025) primary + estimate. Un abbonamento AI di RMB 5–10k per linea all'anno è da due a cinque volte ciò che gli acquirenti ora spendono per il software AOI, quindi va giustificato con le ore di revisione che libera (circa RMB 75–90k per persona a tempo pieno all'anno), non con il confronto con le tariffe software esistenti.

Per i bracci robotici, il prezzo medio dei cobot cinesi è sceso a circa RMB 58,600 nel 2025 (GGII) estimate, e la media a sei assi di Dobot è scesa del 19.1% quell'anno man mano che i modelli più economici guadagnavano quota (Huxiu) verified, anche se i prezzi di listino degli stessi modelli sono saliti del 5–15% nel 2026 (Sina Finance). Negli Stati Uniti, un FANUC CRX-10iA ha un listino di $51,434 (Vention) e le raccolte pongono un UR7e a $30–48k (cobot-cost) estimate. Il braccio è solo il 40–50% di una cella statunitense (Standard Bots); in Cina, fotocamere, pinze, sicurezza e ingegneria arrivano facilmente a due-quattro volte un braccio da RMB 40k, quindi prevedi il costo ingegneristico assoluto piuttosto che un multiplo estimate.

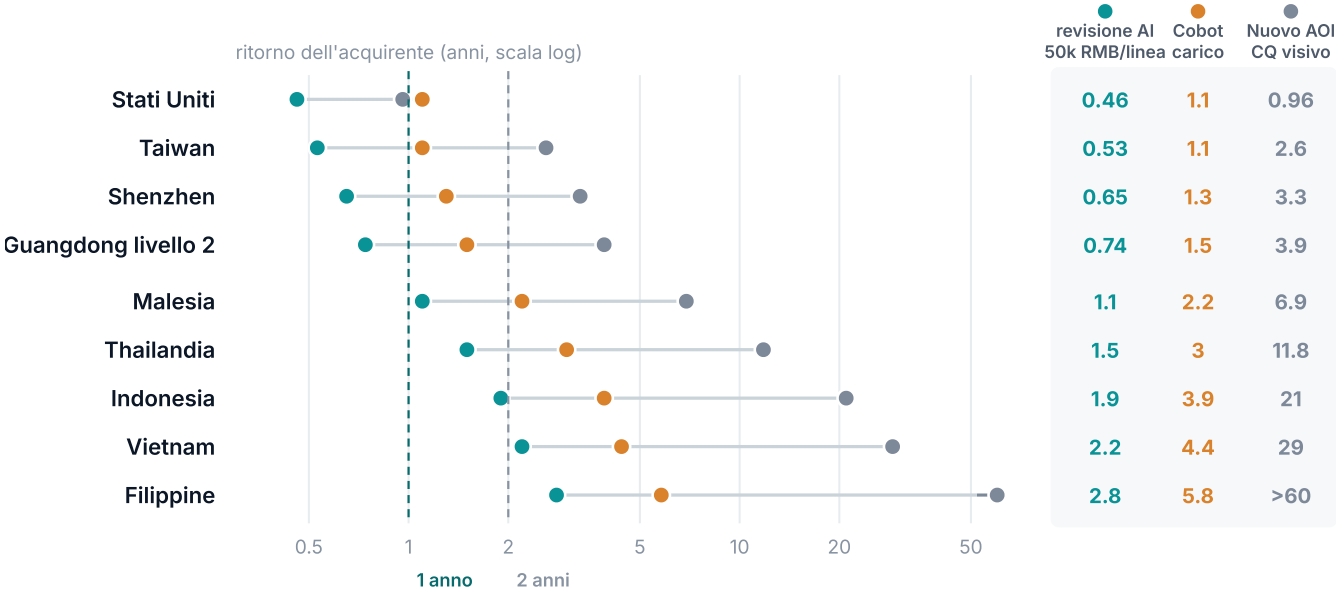
Rientro: revisione in mesi, alimentazione in anni, prelievo mai

Rientro semplice = investimento ÷ (persone risparmiate × costo del lavoro caricato − costo di esercizio annuo), per una linea a due turni senza sussidi o dazi. L'attrezzatura di Taiwan è prezzata a 1.3 volte la Cina, Vietnam, Malaysia e Thailandia a 1.15 volte, Indonesia e Filippine a 1.2 volte, e gli Stati Uniti a prezzi esteri estimate.

Scenario	Investimento (Cina / Stati Uniti)	Persone risparmiate	Costo di esercizio annuo
(a) Box di revisione AI o box fotocamera AI	RMB 50k / $30k per linea	1.0 per linea	15%
(b) Nuovo AOI 2D in linea che sostituisce i controlli visivi manuali (resta 0.5 revisori per turno)	RMB 200k / $60k	1.0	12%
(c) Cella di carico cobot (una stazione × due turni)	RMB 200k / $140k	2.0	8%
(d) Cella integrata detect-and-pick (ispezione 1.0 + prelievo 0.07)	RMB 300k / $200k	1.07	8%
(e) Un braccio robotico aggiunto solo per estrarre le schede NG	RMB 150k / $100k	0.07	8%

Regione	(a) Revisione AI	(b) Nuovo AOI	(c) Cella cobot	(d) Detect-and-pick	(e) Braccio di prelievo
Shenzhen	0.65 anni (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	Mai
Guangdong livello 2	0.74	3.9	1.5	5.3	Mai
Taiwan	0.53	2.6	1.1	3.6	Mai
Malaysia	1.1	6.9	2.2	9.2	Mai
Thailandia	1.5	11.8	3.0	15	Mai
Indonesia	1.9	21	3.9	25	Mai
Vietnam	2.2 (1.6–4.5)	29	4.4	34	Mai
Filippine	2.8	>60 (costo di esercizio ≈ risparmi)	5.8	≈100	Mai
Stati Uniti	0.46	**0.96**	**1.1**	3.4	Mai

FIG 19 **Ritorno per regione***La revisione AI ripaga in mesi negli USA, a Taiwan e Shenzhen e in 1–3 anni nel Sud-est asiatico; l'alimentazione tramite cobot, in anni.*



Un braccio robotico solo per estrarre le schede NG non ripaga mai, da nessuna parte
Con resa di primo passaggio oltre il 90 %, l'estrazione occupa solo 0,003–0,14 FTE per turno

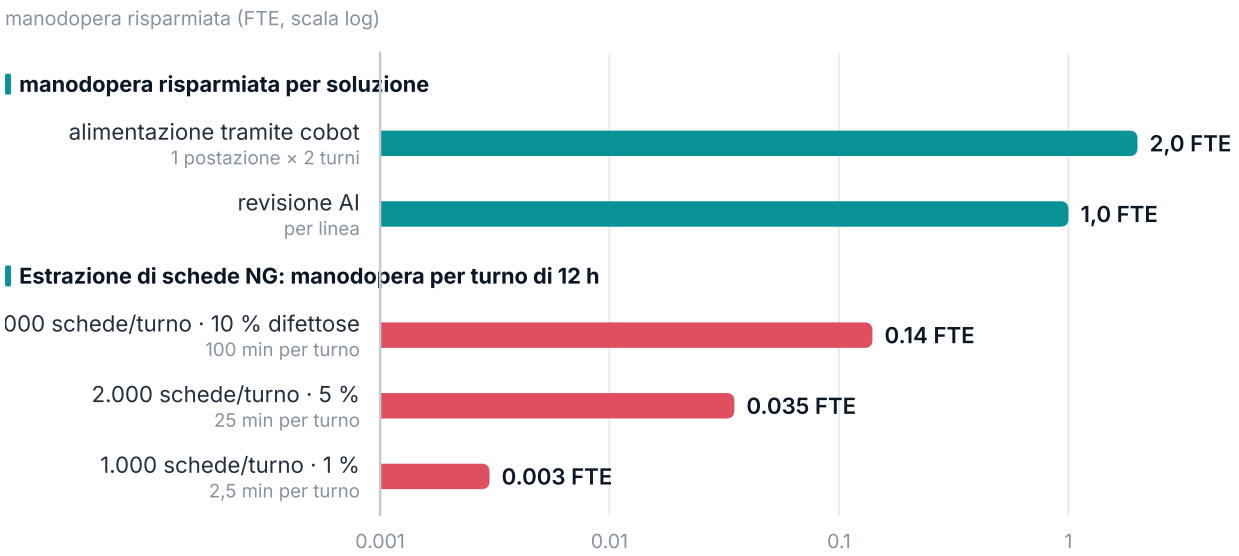
Fonte: Modello dell'autore: scatola di revisione AI da 50k RMB per linea; cella di alimentazione cobot a 200k RMB o 140k $, due turni.

Tutti i valori sono stime dagli input sopra. Sensibilità: dimezza le persone risparmiate e ogni rientro all'incirca raddoppia; fai un solo turno e (b)–(d) raddoppiano; il deprezzamento del 100% al primo anno negli Stati Uniti accorcia il rientro di circa il 20%, mentre i dazi sull'attrezzatura cinese, giapponese o coreana lo allungano di circa il 10–40%; un sussidio locale del 15% in Cina lo accorcia di circa il 15%. Non ho trovato alcuna indagine indipendente sulle soglie di rientro degli acquirenti; i fornitori dicono che l'AOI rientra "tipicamente in 9–18 mesi" (Averroes) e il caso Siemens dichiara otto mesi vendor claims. Uso uno e due anni come linee di screening. Il modello esclude i guadagni di qualità (garanzia, resi) e i costi di turnover e formazione, quindi è prudente.

La parte controintuitiva del "detect and pick" è che il lavoro sta nella revisione, non nel togliere le poche schede NG reali dalla linea. A 15 secondi di movimentazione per scheda NG estimate:

Schede per turno	Tasso di difetto	Tempo di prelievo per turno di 12 ore	Persone
1,000	1%	2.5 min	0.003
2,000	5%	25 min	0.035
4,000	10%	100 min	0.14

FIG 20 **Perché estrarre le schede difettose non ripaga** ***Sopra il 90 % di resa di primo passaggio, l'estrazione occupa solo 0,003–0,14 di persona per turno.***



Fonte: Stime dell'autore.

Il caso base (2,000 schede per turno, 5% difettose, due turni) è 0.07 di una persona, circa RMB 6,000 l'anno a Shenzhen, meno del costo di esercizio annuo di circa RMB 12,000 del braccio, quindi non rientra mai da nessuna parte; anche il caso peggiore (4,000 schede per turno, 10% difettose) richiede circa 13 anni a Shenzhen e 9 negli Stati Uniti estimate. Se la preoccupazione è fermare la linea per prelevare, un deviatore NG o un nastro buffer standard la risolve senza braccio. Le persone del settore dicono "sopra il 90% di resa al primo passaggio, ferma la linea e togli la scheda a mano", e **questo è giusto per il prelievo, ma non per l'ispezione**. Il progetto di prodotto su una linea SMT è quindi "ispeziona più marca o devia", con i bracci robotici tenuti per i punti dove ogni parte viene manipolata (carico di maschere di test, gestione di vassoi e caricatori, imballaggio), dove l'ispezione diventa un add-on quasi gratuito e si applica lo scenario (c).

Sussidi, tasse e dazi spostano il rientro del 15–25%, ma non la classifica

Cina continentale: il sostegno abbassa per lo più il costo di finanziamento; le singole macchine raramente si qualificano direttamente. Un sussidio nazionale sui prestiti paga 1.5 punti di interesse per un massimo di due anni in 14 filiere prioritarie (Yicai) verified, circa RMB 9,000 su una macchina da RMB 300k; il premio di upgrade del Guangdong fino al 20% richiede un progetto registrato con almeno RMB 7 milioni di nuove attrezzature ed esclude Shenzhen (regole Guangdong), e Dongguan ha effettivamente pagato il 4.42% a luglio 2026 (assegnazione Dongguan) single source, primary. Un dispositivo di ispezione AI è approvato solo come attrezzatura "con software di supporto" dentro il progetto registrato dell'acquirente, quindi il caso base non presuppone sussidi.

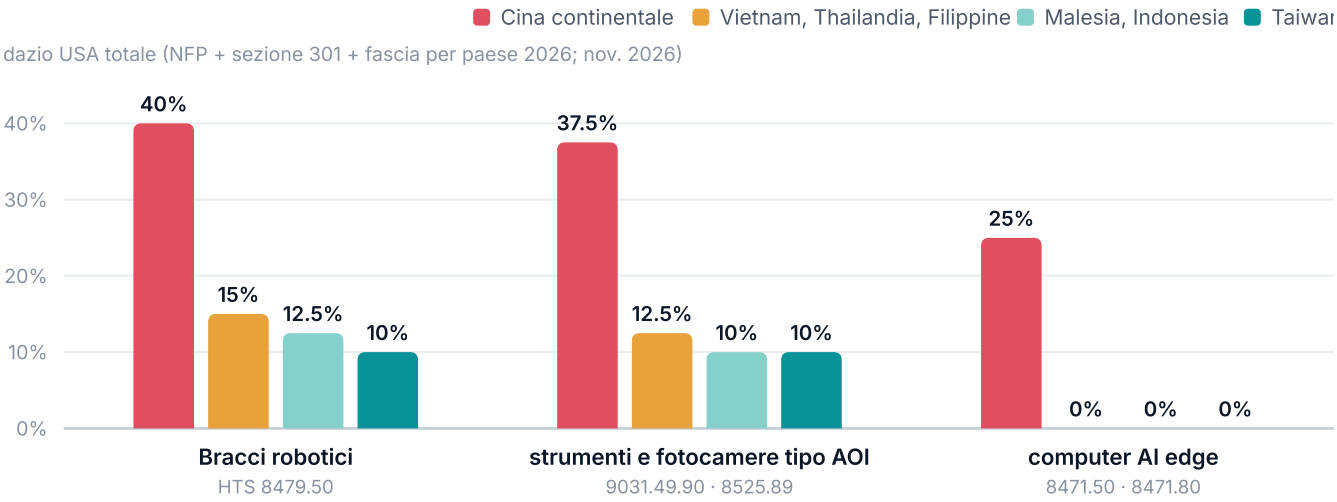
Taiwan: gli incentivi vanno all'acquirente. Sotto l'articolo 10-1 dello Statuto per l'Innovazione Industriale, modificato a maggio 2025 per includere l'AI, un acquirente può portare in credito il 5% della spesa 2025–2029 su nuovi macchinari intelligenti o AI (inclusa l'ispezione basata su CNN/DNN) contro l'imposta sul reddito dell'anno, o il 3% l'anno per tre anni (Statuto, regole di credito) verified; le sovvenzioni lato fornitore sono chiuse al capitale della Cina continentale (MOEA) single source.

Sud-est asiatico: solo Malaysia e Thailandia premiano l'acquisto di automazione. L'allowance di capitale del 200% della Malaysia sui primi RM10 milioni vale circa il 24% del capex per un acquirente redditizio (MIDA, PwC Malaysia) verified; il BOI thailandese dà un'esenzione fiscale triennale fino al 50–100% di un investimento di upgrade, con il tetto più alto quando sono coinvolti fornitori di automazione thailandesi (BOI) single source, primary, quindi entrare in Thailandia funziona meglio con un integratore locale. Vietnam, Indonesia e Filippine premiano la costruzione di stabilimenti piuttosto che l'acquisto di automazione (Vietnam Law Magazine, PwC Filippine).

Stati Uniti: il deprezzamento accorcia il rientro; i dazi lo allungano. I beni acquistati dopo il 19 gennaio 2025 ottengono il deprezzamento permanente del 100% al primo anno, circa uno scudo fiscale federale del 21% per un acquirente redditizio (IRS Notice 2026-11) single source, primary; l'indagine della Sezione 232 su robot e macchinari industriali si è conclusa il 28 ottobre 2026 con negoziati anziché nuovi dazi, da rivalutare entro 180 giorni (Casa Bianca, Mayer Brown) verified. L'assemblaggio finale nel Sud-est asiatico riduce la maggiorazione su bracci e strumenti di tipo AOI da circa il 37.5–40% al 10–15% (capitolo 1).

Regione	Strumenti principali	Effetto sul rientro dell'acquirente estimate
Cina continentale	Sussidio sui prestiti di 1.5 punti fino a due anni; premi di upgrade provinciali e cittadini (il Guangdong richiede ≥RMB 7M di progetti registrati; Dongguan ha pagato effettivamente il 4.42%); misure Shenzhen 2026–2029 (aliquote non pubbliche)	Circa 0–5% per una singola macchina; 4–15% dentro un grande progetto
Taiwan	Credito Statuto §10-1, 5% o 3% × 3 anni	Circa 5% più corto, se l'acquirente paga abbastanza tasse
Malaysia	Allowance di capitale del 200% sull'automazione (primi RM10M)	Circa 15–25% più corto per acquirenti redditizi
Thailandia	Esenzione fiscale triennale BOI fino al 50–100% dell'investimento	Circa 15–25% più corto per acquirenti redditizi fuori da una tax holiday
Vietnam, Indonesia, Filippine	Incentivi fiscali sul reddito per la costruzione di stabilimenti; deprezzamento attrezzature extra del 20% nelle Filippine	Quasi nullo per singoli acquisti di automazione
Stati Uniti	Deprezzamento del 100% al primo anno; ≈37.5–40% di dazio combinato sull'attrezzatura della Cina continentale	Deprezzamento ≈20% più corto; dazi ≈10–40% più lunghi

FIG 21 **Dazio USA su robot, AOI e computer edge*I bracci fabbricati in Cina continentale pagano ≈40 %; spediti dal Sud-est asiatico o da Taiwan, 10–15 %.***



Fonte: Capitolo 99 della Tariffa Armonizzata USITC; Federal Register; USTR (novembre 2026).

Per regione: nella Cina continentale, un box di revisione AI rientra in mesi e una cella cobot in 1.3–1.5 anni, mentre un nuovo AOI acquistato solo per tagliare il lavoro richiede 3–4 anni e il detect-and-pick integrato 4.5–5.3, entrambi oltre uno screening di due anni. Taiwan fa leggermente meglio di Shenzhen (0.53 anni per la revisione, 1.1 per una cella cobot), e lì assumere è più difficile. In Malaysia e Thailandia, un box AI richiede 1–1.5 anni e una cella cobot 2–

3, circa il 15–25% più veloce con gli incentivi fiscali locali per acquirenti redditizi. In Indonesia, Vietnam e Filippine solo un box AI molto economico rientra entro 2–3 anni; la domanda lì viene più dai requisiti di qualità dei clienti e dalle carenze di manodopera. Negli Stati Uniti ogni opzione tranne un braccio di prelievo puro rientra in 0.2–3.4 anni, con dazi, la rivalutazione della Sezione 232 e la certificazione NRTL come variabili principali.

05 Mercato, concorrenza e come pagano gli acquirenti

Mercato e concorrenza

Dimensione del mercato: i numeri pubblicati differiscono di multipli; conta solo un conteggio bottom-up

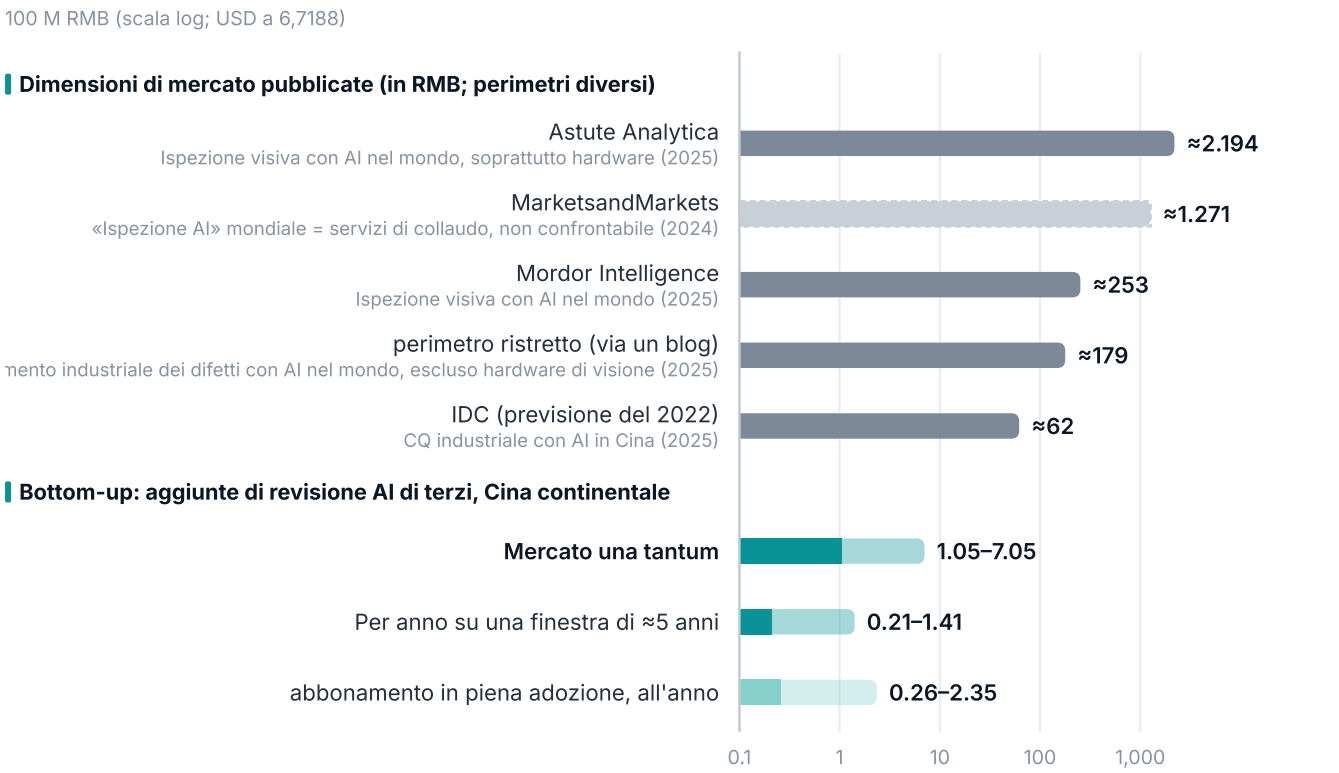
Ogni numero di mercato va letto con il suo perimetro. Il mercato cinese della visione artificiale era di RMB 18.15 miliardi nel 2024 (Zhiyan) verified, con l'elettronica 3C come maggior acquirente; le stime globali per la visione artificiale differiscono di circa tre volte e per l'AOI di circa 6.5 volte (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Bottom-up dai ricavi di Koh Young, TRI, Jutze, Sinic-Tek e ViTrox, l'attrezzatura globale di ispezione SMT è dell'ordine di $1–2 miliardi estimate.

Non esiste un mercato autonomo per il "software di ispezione visiva AI". I numeri pubblicati differiscono di oltre dieci volte e tutti includono hardware o servizi:

Perimetro	Editore	Cifra
"Rilevamento difetti industriali AI" globale (stretto, escluso hardware di visione)	Azienda non nominata, via un blog iFactory	$2.66 miliardi nel 2025, 8.6% CAGR al 2035, elettronica 34%
"Ispezione visiva AI" globale	Mordor Intelligence	$3.76 miliardi nel 2025, $12.8 miliardi nel 2031, 22.67% CAGR
"Ispezione visiva AI" globale (per lo più hardware)	Astute Analytica	≈$32.66 miliardi nel 2025
"Ispezione AI" globale (servizi di test e certificazione; non comparabile)	MarketsandMarkets	$18.92 miliardi nel 2024
"Ispezione qualità industriale AI" cinese	IDC, previsione fatta nel 2022	$958 milioni nel 2025 (≈RMB 6.2 miliardi), 91.5% in comunicazioni, elettronica e due altri settori

Fonti: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) ciascuna single source, per lo più sintesi di rapporti a pagamento.

FIG 22 **Le dimensioni di mercato differiscono di oltre dieci volte** ***Solo il perimetro ristretto e un conteggio bottom-up sono utili alle decisioni.***



Fonte: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; stima bottom-up dell'autore.

Bottom-up, c'è un solo conteggio pubblico delle linee SMT: New Venture Research stimava circa 73,231 linee nel mondo nel 2020, il 47% in Cina (19% nel 2000), circa il 70% di fascia alta e il 30% di fascia bassa (EMSNow, sintesi NVR SMT2020) single source, 2020 ; NVR ora appartiene a IPC e i dati più recenti sono a pagamento (NVR). Crescendo del 3–5% l'anno con una quota cinese del 40–47% si ottengono circa 87,000–98,000 linee nel mondo nel 2026 e 35,000–46,000 nella Cina continentale (base 40,000) estimate .

Input	Basso	Base	Alto	Base
Linee SMT Cina continentale (2026)	35,000	40,000	46,000	Base NVR 2020 × crescita × quota
Quota senza revisione AI	75%	80%	85%	Finestra di retrofit, sotto
Quota raggiungibile da un terzo (non interna, non legata all'AI OEM, può pagare)	40%	50%	60%	Assunzione: ≈30% linee di fascia bassa più alcune linee di fascia alta a marche miste, escludendo build interne in stile DBG
Linee servibili	**10,500**	**16,000**	**23,500**	Prodotto
Prezzo una tantum per linea	RMB 10k	RMB 20k	RMB 30k	Scala e tetto di prezzo (capitolo 4)
Mercato una tantum servibile (base installata)	**RMB 105M**	**RMB 320M**	**RMB 705M**	Prodotto
All'anno su una finestra di ≈5 anni	RMB 21M	RMB 64M	RMB 141M	La finestra si riduce dell'8–12% l'anno
Abbonamento (RMB 2.5k / 5k / 10k per linea all'anno, adozione piena)	RMB 26M l'anno	RMB 80M l'anno	RMB 235M l'anno	Ancora tariffa annua in stile Baidu

FIG 23 **Bottom-up: 40.000 linee → 16.000 servibili → ≈320 M RMB una tantum** ***La nicchia della revisione AI di terzi in Cina continentale è reale, ma piccola.***



Un nuovo entrante che conquista il 2–5 % delle linee servibili in 3 anni: ≈320–800 linee, ≈2–5 M RMB all'anno
Come abbonamento (5k RMB per linea all'anno): 1,6–4 M RMB ricorrenti, solo lo 0,3–2,3 % della previsione IDC del CQ industriale AI in Cina

Fonte: NVR (base 2020); stime dell'autore.

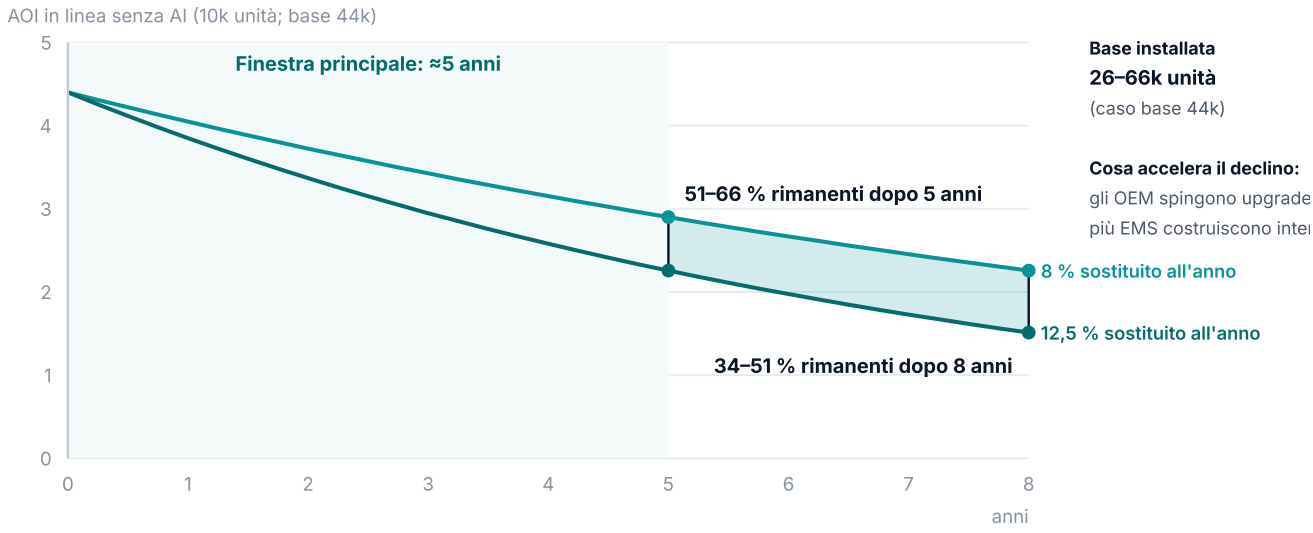
Tutte queste sono stime; né la quota né il prezzo sono stati misurati pubblicamente. La nicchia vale solo lo 0.3–2.3% l'anno della previsione IDC per tutta l'ispezione qualità industriale AI cinese nel 2025 (circa RMB 6.2 miliardi). Un nuovo entrante che vincesse il 2–5% delle linee servibili di base in tre anni avrebbe circa 320–800 linee: RMB 6.4–16 milioni in totale a RMB 20,000 per linea, circa RMB 2–5 milioni l'anno, o RMB 1.6–4 milioni l'anno di ricavi ricorrenti a RMB 5,000 per linea estimate. Questo basta per una business unit dentro un'azienda, non per un'azienda autonoma di scala venture. Ogni 1,000 linee installate a Taiwan, nel Sud-est asiatico o negli Stati Uniti valgono all'incirca RMB 3–15 milioni una tantum ai prezzi della Cina continentale; i prezzi sono più alti a Taiwan e negli Stati Uniti, ma lì gli acquirenti usano più spesso AOI 3D OEM con opzioni AI estimate.

La finestra di retrofit è di circa cinque anni

Nessun produttore di AOI pubblica la vita della macchina; le aziende EMS deprezzano l'attrezzatura in 3–10 anni (relazione annuale DBG 2025), e gli AOI più vecchi perdono aggiornamenti software e ricambi dopo la fine vita (I-Connect007) single source. Uso 8–12 anni (base 10).

A 1.0–1.65 AOI in linea per linea, la Cina continentale ha circa 35,000–78,000 AOI in linea in servizio (base circa 55,000), coerente con le spedizioni note (Circuits Assembly, relazione annuale Sinic-Tek 2025). Nessuna fonte misura quanti usano la revisione AI; assumendo che le macchine domestiche la offrano da circa il 2023 e che il 40–60% delle consegne recenti la usi, **circa 26,000–66,000 AOI in linea nella Cina continentale (base circa 44,000) non usano l'AI** estimate. Se ogni nuova macchina ora esce con l'AI, quel bacino si riduce dell'8–12.5% l'anno: il 51–66% resta dopo cinque anni e il 34–51% dopo otto, più velocemente se i produttori di AOI spingono upgrade AI verso le macchine più nuove. **La maggior parte dell'opportunità cade nei prossimi cinque anni circa**, con al centro gli EMS di medie dimensioni con AOI 2D domestici più vecchi, marche miste e nessun percorso di upgrade AI.

FIG 24 **La finestra di retrofit è di circa cinque anni** ***Gli AOI in linea senza AI si riducono dell'8–12,5 % all'anno man mano che vengono sostituiti o aggiornati.***

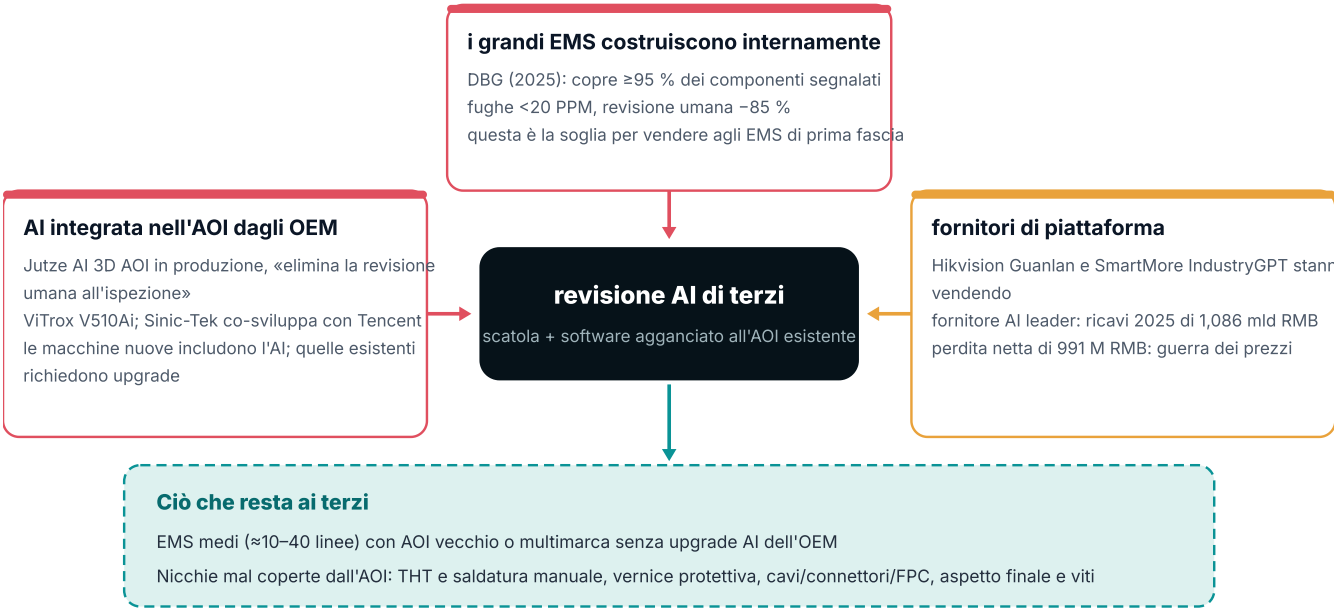


Fonte: Stime dell'autore; base installata e quote non verificate in modo indipendente.

I concorrenti sono già forti: il cuneo della revisione è accerchiato

Il cuneo della revisione AI si sta restringendo da tre direzioni. I leader di AOI e visione stanno integrando l'AI nelle nuove piattaforme: Jutze ha "adottato pienamente gli algoritmi AI" (relazioni con gli investitori Jutze, giugno 2026), il fatturato AI industriale di OPT è cresciuto del 96.77% nel 2025 (relazione annuale OPT 2025) single source, primary, e ViTrox dichiara fino al 95% di accuratezza per la revisione autonoma (ViTrox) vendor claim. I grandi produttori costruiscono la propria (iThome). E le aziende AI sostenute da venture capital vendono su scala: i clienti dei sensori edge di SmartMore sono cresciuti da 65 a 251 in due anni (36Kr) single source, Hikvision vende il suo modello Guanlan per l'ispezione (Xinhua), e l'integrazione "occhio, cervello e mano" è già offerta da SmartMore, Hikrobot e MechMind (36Kr, Securities Star).

FIG 25 **Il cuneo di revisione è accerchiato su tre lati** ***Lo spazio per i terzi è negli EMS medi e nelle nicchie di ispezione mal coperte dall'AOI.***



Fonte: Rapporti annuali di Jutze, Sinic-Tek e DBG; materiale pubblico di Hikvision e SmartMore; compilato dall'autore.

Lo spazio lasciato a un nuovo entrante è stretto: EMS di medie dimensioni con AOI 2D/3D più vecchi e nessun upgrade AI dal produttore; linee che mescolano più marche di AOI (un fornitore di retrofit supporta già Omron, VI, Koh Young e Mirtec (Sohu su Hengtianyi)); nicchie che gli AOI SMT coprono male, come giunti a foro passante e saldati a mano, conformal coating, cavi, connettori e FPC, e aspetto finale e viti; e piccoli stabilimenti che non possono permettersi un AOI da RMB 230–280k.

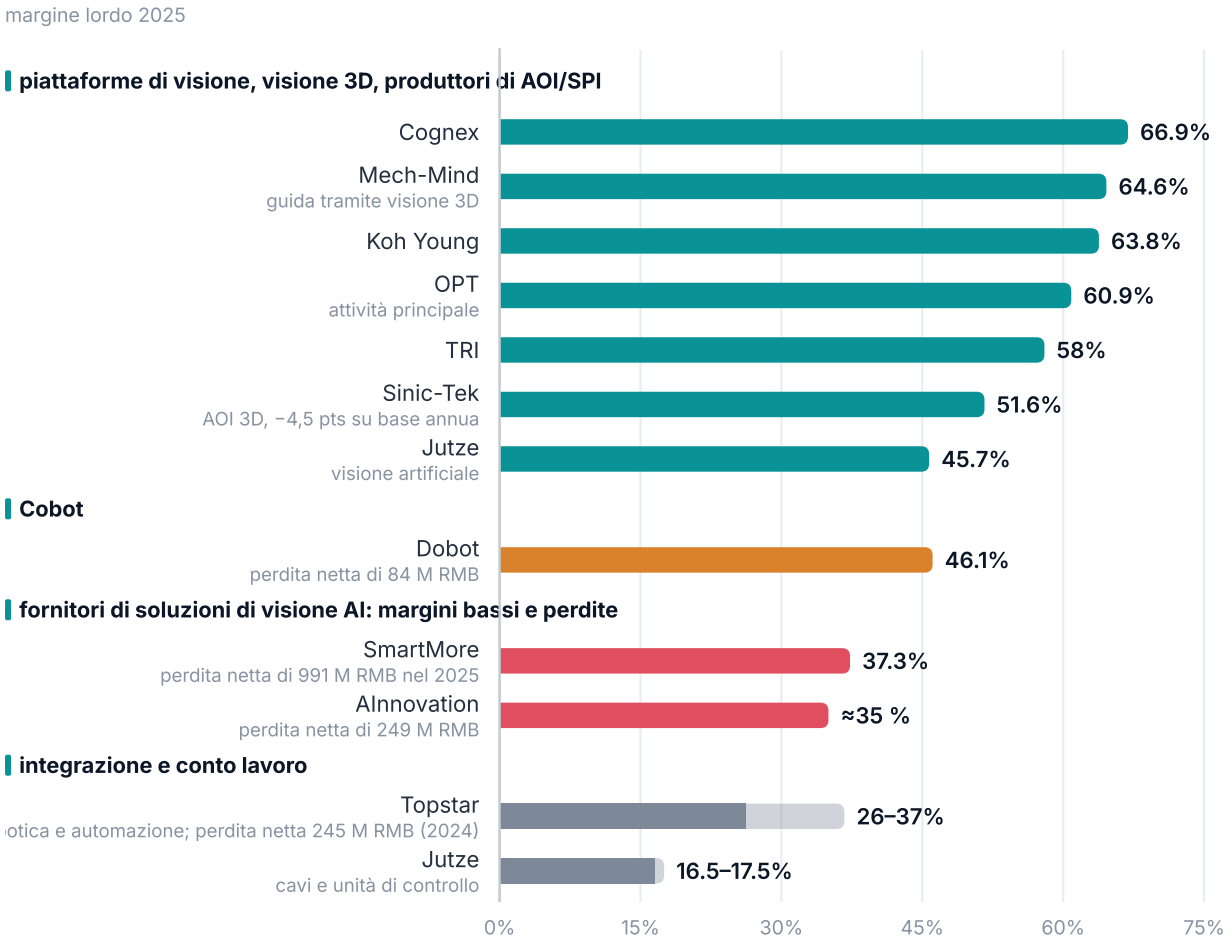
Il comparabile quotato più vicino dà un avvertimento. Weiyi Zhizao, che vende attrezzature di ispezione AI e robot, ha visto il fatturato 3C calare del 79.9% anno su anno in gennaio–maggio 2026, ha riconosciuto "concorrenza crescente e pressione sui prezzi", ha rinviato la sua quotazione a Hong Kong il 25 ottobre 2026 e ha comunicato 15 progetti terminati prima del completamento, 12 dei quali progetti di robot incarnati (prospetto Weiyi Zhizao, avviso di rinvio) single source, primary. AInnovation ha chiuso il 2025 con il 29.8% dei suoi crediti svalutato (relazione annuale AInnovation 2025). Il cuneo dell'ispezione 3C è affollato, e il potere di prezzo per linea è debole.

Il margine sta nell'hardware prodotto in serie

Posizione nella catena	Aziende rappresentative e margine lordo 2025	Profitto e incassi
Piattaforme di visione	Cognex 66.9%	Redditizio; dismissione di ≈$22M di ricavi a basso margine
Attrezzatura AOI/SPI	Koh Young 63.8%; TRI 58%; core business OPT 60.85%; Sinic-Tek AOI 3D 51.62% (−4.53 punti YoY); visione artificiale Jutze 45.74%	Redditizio; giorni di credito 139 (Jutze) a 247–274 (OPT)
Guida visiva 3D	Mech-Mind 64.6%	—
Fornitori di soluzioni di visione AI	SmartMore 37.3%; AInnovation ≈35%	In perdita: SmartMore perdita netta RMB 991M nel 2025, AInnovation RMB 249M
Cobot	Dobot 46.1%	Dobot perdita netta RMB 84M; ricavi UR da $326M (2022) a ≈$232M (2025, previsione)
Integrazione di sistema	Topstar robotica e automazione 26.2–36.7%	Perdita netta RMB 245M (2024); giorni di credito 180–240
Hardware su contratto	Cavi e unità di controllo Jutze 16.5–17.5%	—

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

FIG 26 **Il margine sta nell'hardware industrializzato***Piattaforme di visione e produttori di apparecchiature realizzano il 46–67 % di margine lordo; i fornitori di soluzioni AI il 35–37 % e perdono denaro.*



Fonte: Rapporti annuali o prospetti 2025.

Un nuovo entrante mantiene il margine solo trasformando l'ispezione in un prodotto, attrezzatura o software; la messa a punto progetto per progetto spinge i margini a circa il 35% con perdite persistenti. Anche la meccanica gioca contro le piccole aziende: il fatturato è riconosciuto solo dopo l'accettazione in loco, gli anticipi sono piccoli (Jutze, OPT), i piloti sono di solito gratuiti (A3/Mindtrace), e i proprietari di marchi spesso specificano l'attrezzatura di ispezione sulle linee dei loro produttori a contratto.

Come pagano gli acquirenti: 30% alla firma, 60% dopo l'accettazione, 10% trattenuto per un anno

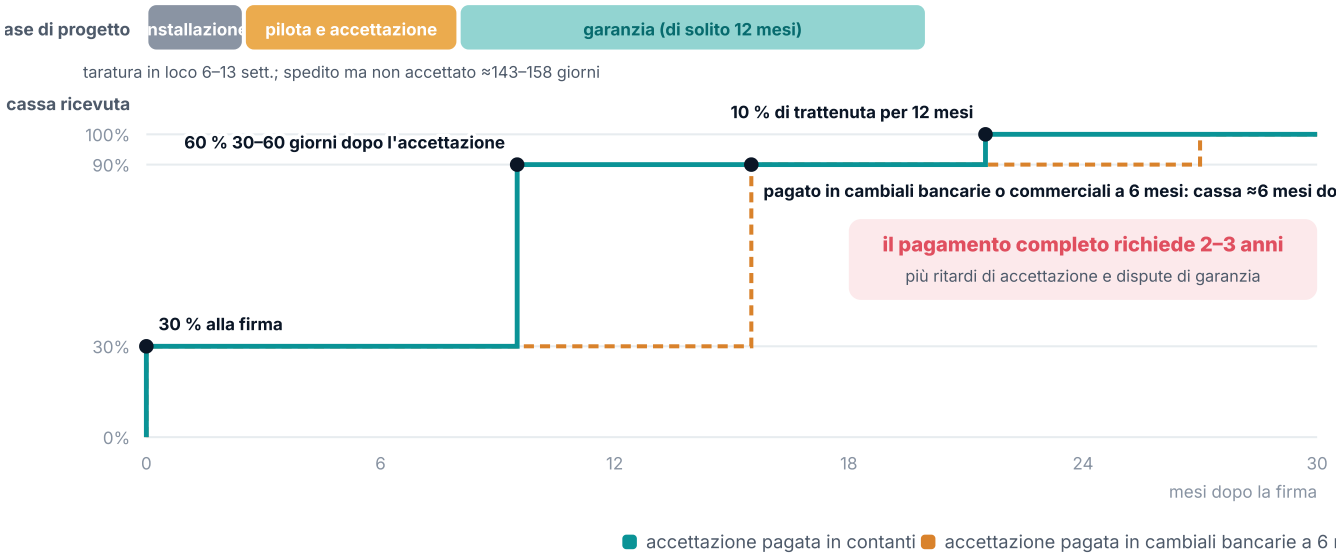
I termini di pagamento che gli stabilimenti di elettronica della Cina continentale usano per l'attrezzatura di ispezione sono esplicitati in 13 contratti di acquirenti nominati divulgati nella risposta all'indagine di quotazione di Sinic-Tek (risposta all'indagine Sinic-Tek 8-1-1) single source: un documento primario, più acquirenti indipendenti, 2020–2022 :

Acquirente (anno)	Firma	Consegna	Accettazione	Trattenuta o saldo
Un'unità Hikvision (2021)	30% entro 15 giorni	—	65% entro 30 giorni dall'accettazione finale	5% di trattenuta di garanzia pagato 12 mesi dopo l'accettazione; parte pagata in note di accettazione bancarie a 6+ mesi
Nanchang Huaqin (2020–2021)	30% entro 10 giorni lavorativi	—	60% a 60 giorni dall'accettazione	10% dopo una garanzia di un anno

Acquirente (anno)	Firma	Consegna	Accettazione	Trattenuta o saldo
Guangdong Yinuo (2021, AOI 3D)	30% entro 10 giorni lavorativi	—	60% entro 30 giorni	10% dopo tre mesi di funzionamento senza guasti
Xiamen Hongxin, Jingmen Hongyi (2022)	30% entro 10 giorni lavorativi	—	60% entro 30 giorni lavorativi dall'accettazione finale, note bancarie a 6 mesi	10% dopo una garanzia di un anno
AKM Panyu, Suzhou (2021–2022)	30% entro 7–10 giorni lavorativi	—	60% un mese dopo l'accettazione e la fatturazione, note bancarie a 180 giorni	10% dopo 360 giorni
BYD Precision, Shenzhen, Huizhou (2020)	—	30% all'arrivo	70% dopo l'accettazione e la fatturazione	Tutto in note commerciali a 6 mesi
Shenzhen H&T (2020)	—	—	90% dopo l'accettazione e la fatturazione	10% dopo sei mesi di uso senza guasti

Altri documenti primari mostrano la stessa struttura: i grandi progetti di Mech-Mind pagano il 30% alla firma, il 30% prima della consegna, il 30% dopo l'accettazione e il 10% di trattenuta di garanzia su 12–15 mesi (prospetto Mech-Mind); GKG afferma che la trattenuta nel settore dell'attrezzatura è di solito il 5–10% del contratto, pagata un anno dopo l'accettazione (risposta all'indagine GKG) single sources, primary. La norma, quindi, è circa il 30% alla firma (20–30%, zero in alcuni stabilimenti come BYD), il 60–65% dopo l'accettazione (pagato 30–60 giorni dopo, spesso in note bancarie o commerciali a sei mesi, il che ritarda la liquidità utilizzabile di un altro semestre), e il 5–10% di trattenuta per 3–12 mesi (12 il più comune) verified su più contratti.

FIG 27 **Come pagano gli acquirenti *30 % alla firma, 60 % dopo l'accettazione, 10 % trattenuto per un anno: il pagamento completo richiede 2–3 anni.***



Fonte: Risposta a richiesta Sinic-Tek, Mech-Mind, GKG, JT e altri documenti; schema su condizioni usuali, tempistica stimata.

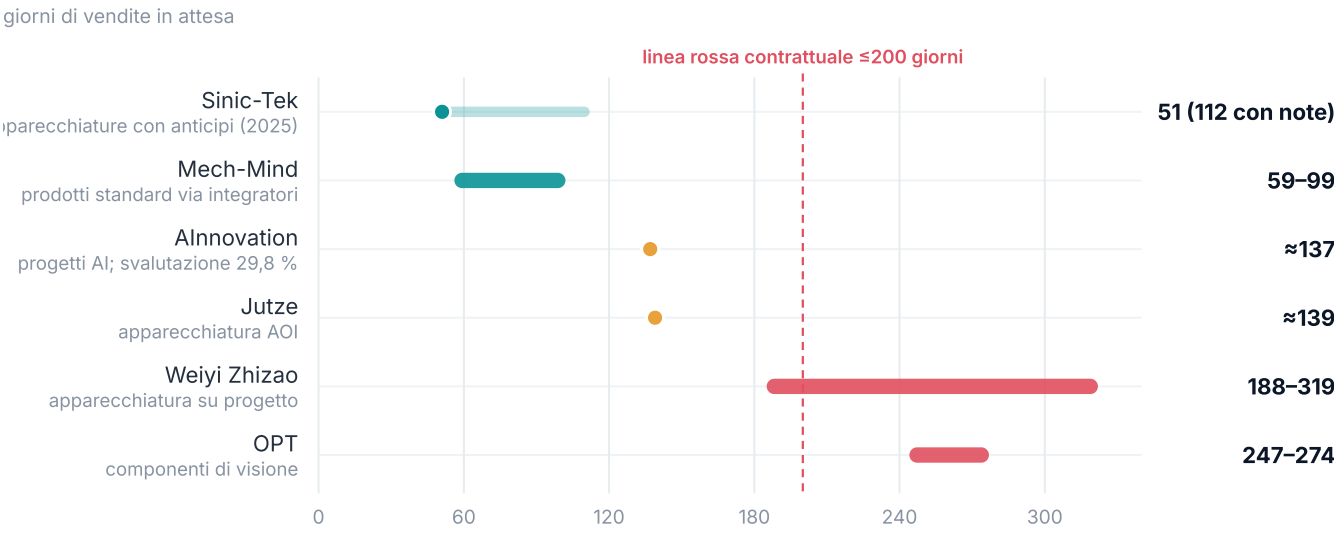
Accettazione e collaudi sono lenti e pagati dal venditore. I beni spediti ma non accettati di Sinic-Tek impiegavano circa 143–158 giorni a girare, perché i grandi EMS e i clienti di marca richiedono test di integrazione di intera linea e una prova estesa (risposta all'indagine Sinic-Tek 8-1-2); i progetti AI di Weiyi richiedono 4–8 mesi dalla valutazione alla distribuzione (prospetto Weiyi Zhizao) single sources, primary. Non ci sono dati di conversione per i piloti di revi-

sione AI; Sinic-Tek convertiva solo circa il 18–19% delle sue macchine demo direttamente in vendite. Pianifico uno su tre a uno su cinque per i piloti gratuiti e circa l'80% per i piloti a pagamento dopo la shortlist estimate.

Azienda	Modello di business	Giorni di vendite non incassate
Sinic-Tek (2025)	Attrezzatura con acconti	≈51 (≈112 incluse le note)
Mech-Mind	Prodotti standard tramite integratori	59–99
Jutze	Attrezzatura AOI	≈139
AInnovation (2025)	Progetti AI	≈137 a fine anno, con 29.8% di svalutazione
Weiyi Zhizao	Attrezzatura su progetto	188–319
OPT	Componenti di visione	247–274

Fonti: le relazioni annuali e i prospetti sopra; crediti (incluse le attività contrattuali) ÷ ricavi × 365 primary + estimate.

FIG 28 **Giorni di vendite in attesa** ***Prodotti standard 51–99 giorni; apparecchiature su progetto 188–319 giorni.***



Fonte: Documenti aziendali; crediti (incl. attività contrattuali) ÷ ricavi × 365.

Le licenze di revisione AI vendute tramite produttori di AOI o integratori dovrebbero incassare come i 60–100 giorni di Mech-Mind; le vendite dirette su progetto come i 190–320 giorni di Weiyi. Mettendo tutto insieme come cronologia di cassa: pilota gratuito da uno a tre mesi, qualificazione presso un grande ODM fino a circa dodici mesi, costruzione, consegna e accettazione da tre a sette mesi; il 30% arriva all'ordine, il 60% circa quattro-nove mesi dopo e l'ultimo 10% circa 16–21 mesi dopo. **Dal primo pilota presso un grande EMS al pagamento completo possono passare due-tre anni**; gli EMS di medie dimensioni senza gare formali possono accorciare la qualificazione a settimane estimate. I criteri di accettazione nei contratti divulgati coprono solo l'integrazione di intera linea, il throughput (un progetto taiwanese richiedeva 100 unità all'ora) e mesi di funzionamento senza guasti; nessuno specifica tassi di errori sfuggiti o falsi allarmi (risposta all'indagine Anda) single source, primary. I criteri di accettazione scritti per la revisione AI vanno concordati prima del pilota.

Quattro schemi di fallimento

I fallimenti documentati cadono in quattro schemi: spedire prima che il prodotto fosse affidabile (Rethink (The Robot Report)); uno strato software sull'hardware di un leader che l'ecosistema del leader ha poi sostituito (READY Robotics (The Robot Report)); finire i soldi al fondo di un ciclo (Dextrous Robotics (Robots & Startups)); e aziende model-

lo finite in acqui-hire (Covariant (GeekWire)). Le lezioni: produttizza in modo mirato invece di mettere a punto progetto per progetto; tieni capitale circolante per ricavi basati sull'accettazione e incassi lunghi; fornisci servizio locale per stabilimenti che girano 24 ore; e non scalare prima che il prodotto sia provato.

Tempistica: l'ispezione non guadagna nulla ad attendere; l'hardware dei bracci diventa più economico

Gli argomenti per attendere sono deboli. L'hardware non è diventato più economico nel 2026: i produttori di robot hanno alzato i prezzi del 5–15%, alcuni prodotti del 25–50% (Sina Finance) single source, e i kit NVIDIA sono saliti una terza volta (36Kr). I VLA di livello produttivo non arriveranno dentro la finestra (IFR, Instrumental). Nel frattempo ogni sostituzione di AOI rimuove un obiettivo di retrofit e ogni distribuzione di AI OEM blocca i dati e il processo di una linea; la capacità dei modelli si sta commoditizzando, e ciò che scarseggia è l'accesso agli stabilimenti e ai dati di fabbrica etichettati. Anche il supporto locale ha finestre: le regole 2026 di Shenzhen sulla "città pioniera dell'AI", con fino a RMB 2 milioni per progetti dimostrativi di applicazione AI e voucher per modelli che coprono il 30% degli acquisti, durano fino al **30 giugno 2027** (Yicai, maggio 2026) termini single source.

Ciò che si sta ancora svalutando velocemente sono gli umanoidi e le mani destre (le mani sono scese da circa RMB 50,000 a sotto 10,000 in due anni (Sina Finance)), e i cobot entry man mano che il mix di prodotto scende; questo spinge a rinviare la parte incarnata, non la parte di ispezione. **In sintesi: ispezione adesso, come prodotto mirato; celle robot adesso, tramite partner e integrazione; AI incarnata più tardi, con solo una stazione R&D oggi.**

AI incarnata

06 AI incarnata: prima cambia come si insegnano i robot, poi come si muovono

Ciò che cambia è percezione e programmazione; il movimento resta deterministico

Ogni leader dei bracci robotici ora commercializza la "AI fisica", ma ciò che è stato spedito cade in tre strati: percezione appresa senza insegnamento (AI Masterless Vision di FANUC, in spedizione a gennaio 2027 (FANUC); la selezione neurale della presa di KUKA (KUKA)); grandi modelli che scrivono programmi e pianificano compiti (iiQWorks.Copilot di KUKA, che avverte che il codice generato dall'AI va verificato in un gemello digitale (KUKA); Yaskawa con Gemini Robotics-ER, dove "l'AI generativa decide cosa fare" e il robot decide "come muoversi" (Yaskawa)); e piattaforme di calcolo e dati (l'AI Accelerator di UR con un Jetson Orin (Universal Robots)) vendor claims. Nessun leader pubblica tassi di successo in produzione per le sue funzioni AI, e i leader giapponesi collaborano con Google, SoftBank, Microsoft e altri piuttosto che costruire VLA frontier (indice notizie FANUC): lo strato di policy apprese arriverà dalle aziende di modelli, i bracci e i controller certificati dai produttori di bracci.

FIG 29 **Lo stack di AI incarnata** ***I modelli grandi pianificano offline, le politiche piccole girano nella cella, movimento e sicurezza restano deterministici.***

Livello	Cosa fa e fino a dove arriva l'evidenza	Scala temporale · chi la controlla
Pianificazione L4	**modelli di frontiera (offline)** scrivere o modificare programmi robot offline, cambi serie, spiegare i guasti riscrive i programmi ABB RAPID al 77–99,7 %, distribuito dopo simulazione loop di controllo a soli 0,2–0,4 Hz: resta fuori dal tempo reale	**da secondi a minuti** giganti dei modelli
Competenze / politi L3	**ACT, politiche di diffusione, piccoli VLA** postazioni strutturate con attrezzatura 90–99,4 % chip edge: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**blocchi d'azione 100–300 ms** produttori di chip · produttori d
controllo del movim L2	**Controller robot** traiettorie, servo, controllo di forza, rilevamento collisioni codice deterministico; le politiche apprese fissano solo gli obiettivi	**millisecondi (il tempo reale richiede >83 Hz)** produttori di robot
Sicurezza L1	**PLC di sicurezza, scanner, limiti di forza e velocità** certificato secondo ISO 10218-1/-2:2025 e regole locali Niente ML nelle funzioni di sicurezza: valutazione da organismo notificato UE dal 20 gen. 2027	**tempo reale rigido** robot certificato + integratore

Le piccole aziende vincono nell'integrazione e nell'hardware di rilevamento: inserire percezione appresa e piccole politiche in celle robotiche ce

Fonte: Sintesi dell'autore (capitoli 3 e 6); ISO 10218:2025; Regolamento Macchine UE.

Quanto manca al lavoro non strutturato, flessibile, in piccoli lotti: stazioni strutturate in 1–3 anni, assemblaggio flessibile 5+

I numeri reali di fabbrica hanno iniziato ad apparire, ma sono riportati dai fornitori, stretti e da stazioni strutturate. Gli otto umanoidi di AgiBot sulla linea di test finale per tablet di Longcheer hanno registrato 64,828 compiti al 99.99% su oltre 64 ore (AgiBot, Yicai Global) vendor e cliente, ma è lavoro di tending che un braccio convenzionale con visione può fare. UBTech afferma che il suo Walker S2 lavora al 30–50% dell'efficienza umana (Irish Times/FT) single

source. Gli 11 mesi di Figure 02 da BMW Spartanburg hanno prodotto cifre di volume (1,250+ ore, 90,000+ parti), mentre i suoi numeri di ciclo, successo e intervento sono obiettivi, non risultati (Figure, BMW Group) verified, solo numeratori. Nucleus afferma che circa il 40% del suo video di fabbrica è teleoperato (Interesting Engineering) vendor claim, e un'inchiesta Reuters ha citato circa il 20% della produzione di un lavoratore per alcuni umanoidi cinesi (Notebookcheck) single source. Nessuna distribuzione pubblica insieme tasso di successo, tempo ciclo e tasso di intervento, e nessuna è stata sottoposta a audit, eppure sono esattamente questi tre ciò di cui un calcolo di rientro ha bisogno.

Distribuzione	Tasso di successo	Takt	Interventi o teleoperazione	Uptime o intervallo di guasto	Run di volume
Figure 02 @ BMW Spartanburg	Solo obiettivo (>99% per turno)	Solo obiettivo (84 s)	Solo obiettivo (0 per turno)	Non comunicato (avambraccio primo punto di guasto)	1,250 h
Agility Digit @ GXO	Non comunicato	Non comunicato	Non comunicato	Non comunicato	100,000 tote
Apptronik Apollo @ Mercedes-Benz	Non comunicato	Non comunicato	Non comunicato	Non comunicato	Non comunicato
Agile ONE @ Audi	Non comunicato	Non comunicato	Non comunicato	Non comunicato	Non comunicato
Nucleus (stabilimento non nominato)	Non comunicato	Non comunicato	≈40% teleoperato (metodo non comunicato)	Non comunicato	≈2 h di video
AgiBot G2 @ Longcheer	99.99% di 64,828 compiti	18–20 s per passo	Non comunicato	Un guasto di comunicazione, ≈2 min per riprendere	64+ h
Cella ibrida Neuromeka	322/324	12.8% più lenta di una persona	Non comunicato	Non comunicato	324 cicli
Physical Intelligence π0.6 presso un cliente	—	—	96.4% autonomo su un turno	Non comunicato	Un turno

Fonti: questo capitolo e il capitolo 3 per lo più vendor claims.

FIG 30 **Cosa divulgano davvero le distribuzioni robot in stabilimento** ***La maggior parte divulga solo il volume; nessuna divulga insieme tasso di successo, takt e tasso di intervento.***

	tasso di successo	Takt	Interventi / teleoperazione	Disponibilità / guasti	Volume
Figure 02 @ BMW	solo obiettivo >99 %/turno	solo obiettivo 84 s	solo obiettivo 0/turno	n/d	1.250 h
Agility Digit @ GXO	n/d	n/d	n/d	n/d	100.000 contenitori
Apptronik Apollo @ Mercedes	n/d	n/d	n/d	n/d	n/d
Agile ONE @ Audi	n/d	n/d	n/d	n/d	n/d
leus (stabilimento non nominato)	n/d	n/d	≈40 % teleoperato*	n/d	≈2 h di video
AgiBot G2 @ Longcheer	99.99%	18–20 s per fase	n/d	un guasto, 2 min	64+ h
cella ibrida Neuromeka	322/324	12,8 % più lento	n/d	n/d	324 cicli
π0.6 @ sede cliente	—	—	96,4 % autonomo	n/d	1 turno

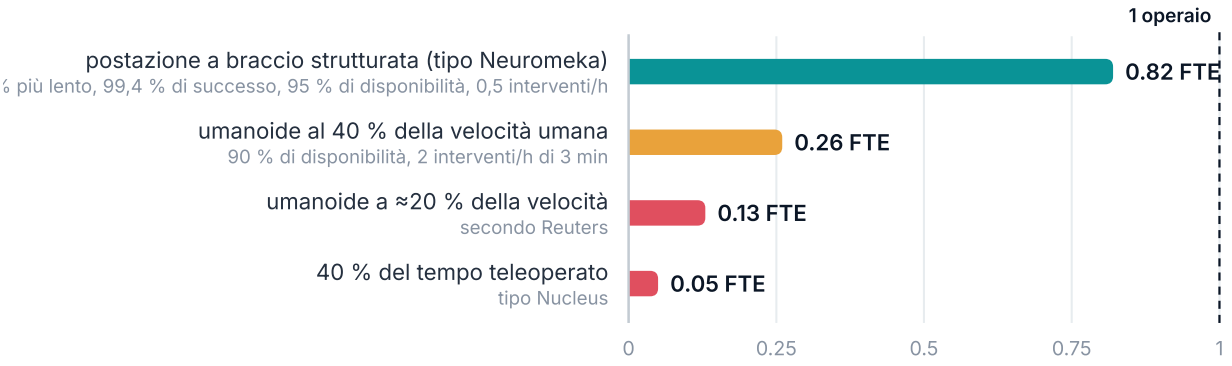
* metodo non divulgato

Divulgato (soprattutto affermazioni del fornitore) · solo obiettivi · n/d

Fonte: Materiale pubblico di BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence e altri.

La teleoperazione conta più del tasso di successo per l'economia. Persone nette sostituite per robot per turno ≈ (ciclo umano ÷ ciclo robot) × uptime × successo, meno tempo di intervento e teleoperazione estimate, formula mia. Una stazione strutturata in stile Neuromeka (12.8% più lenta, 99.4% di successo, 95% di uptime, 0.5 interventi di 2 minuti all'ora) sostituisce circa 0.82 persone; un umanoide al 40% della velocità umana, 90% di uptime e due interventi di 3 minuti all'ora circa 0.26; al 20% di Reuters, circa 0.13; e se, come Nucleus, un operatore lo teleopera il 40% del tempo, circa 0.05, quasi nulla. A circa RMB 600k per umanoide e RMB 85k per persona all'anno nella Cina continentale, gli ultimi tre rientrano in decenni; solo la stazione con braccio strutturato ha un rientro realistico (capitolo 4) estimate.

FIG 31 **Manodopera netta sostituita per robot** ***Una cella a braccio strutturata sostituisce ≈0,82 di lavoratore; un umanoide teleoperato il 40 % del tempo, ≈0,05.***



Fonte: Stime basate sulle formule dell'autore.

Lontano dalle stazioni strutturate, i modelli generali totalizzano ancora il 40–90% sui nuovi compiti: Gemini Robotics 2 raggiunge il 74.2% sul pick-and-place di laboratorio, il 78.9% sul kitting di utensili e il 44% nel fare un nodo a una borsa (eesel/DeepMind); su RoboChallenge Table30, 30 compiti su robot reale, π0.5 totalizza il 43.7% e il miglior modello aperto il 50.3% (RoboChallenge, QbitAI).

Le previsioni esterne divergono, ma nessuna sostiene un uso diffuso entro due anni. L'IFR afferma che l'AI fisica generale "non è ancora di livello industriale" e che la maggior parte degli umanoidi sono "ancora prototipi o piloti"

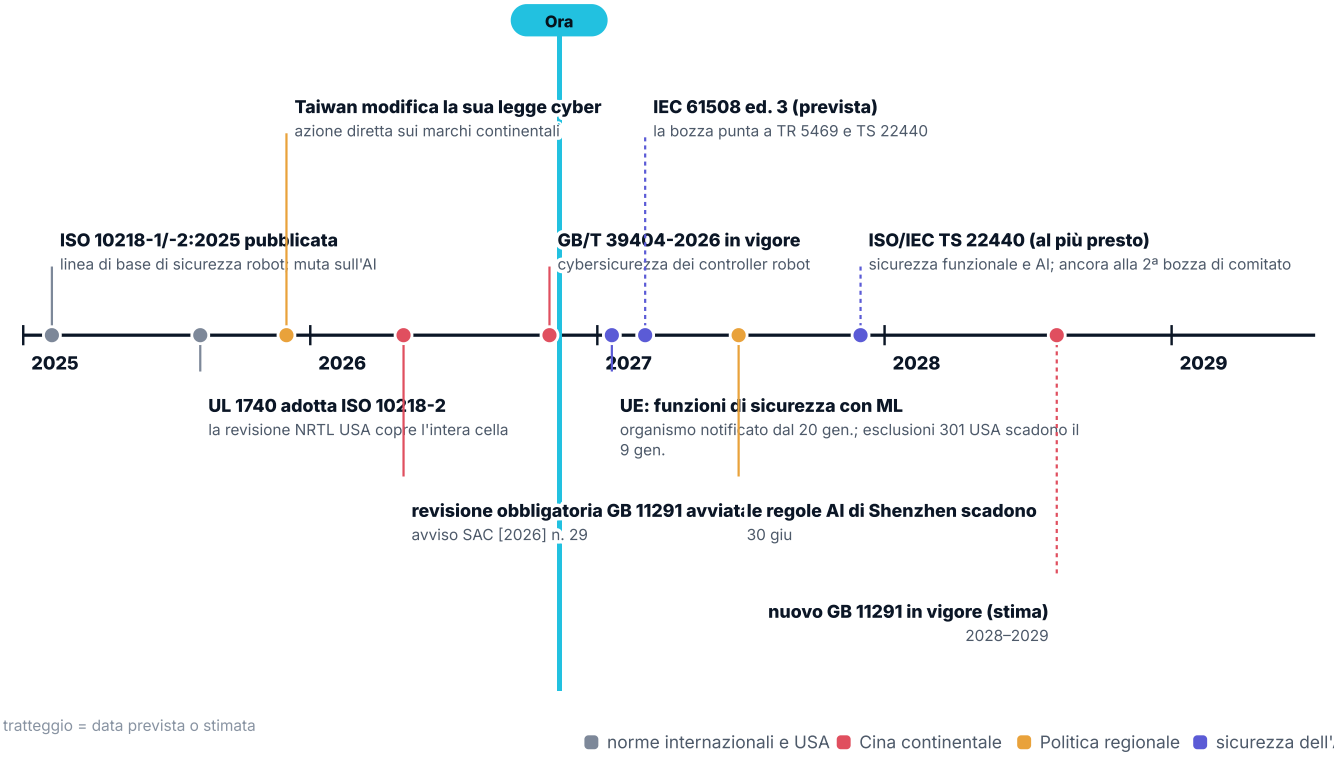
(IFR); Interact Analysis stima che solo circa il 10% degli umanoidi prodotti nel 2025 sia entrato in uso reale, con il punto di svolta commerciale "dopo il 2032" (automation.com); UBS vede un "momento EV" degli umanoidi improbabile prima del 2030 (Yicai Global); insider cinesi stimano altri due-tre anni prima dell'uso su scala di fabbrica (Gasgoo) single sources. Il mio giudizio estimate: in stabilimenti di elettronica e automotive ben dotati con il fornitore in loco, **i compiti di stazione strutturata** (tending, smistamento OK/NG, inserimento su maschera) raggiungeranno ≥99% a velocità quasi umana in circa 1–3 anni; **il kitting a parti miste** con VLA generali in circa 3–5; **i cablaggi a percorso libero e l'assemblaggio flessibile multi-passo** in più di 5. Il rientro degli umanoidi oggi non funziona: un Walker S2 a circa RMB 600k (10jqka) single source, al 30–50% dell'efficienza umana su due turni, equivale a circa 0.6–1.0 lavoratori, un rientro semplice di 7–12 anni ai costi di lavoro della Cina continentale prima di integrazione, manutenzione e fermo estimate.

Dove è bloccato

Anello	Evidenza nel 2026
Dati	Il settore stima che servano almeno 10 milioni di ore di dati reali di alta qualità e solo circa 500,000 ne erano state raccolte all'inizio del 2026 (Gasgoo); Reuters riporta circa 500,000 ore contro circa 100 milioni necessarie (Notebookcheck); un teleoperatore novizio ha bisogno di circa 300 tentativi per movimento utilizzabile; gli stabilimenti non condivideranno i dati di processo; le policy addestrate in simulazione crollano nettamente sui robot reali
Affidabilità	Stazioni strutturate 98–99.99% (vendor claims, guasto indefinito); modelli generali 40–90% sui nuovi compiti; connettori mai visti 29–67%
Velocità	Umanoidi al 30–50% della velocità umana (UBTech), alcuni a circa il 20% (Reuters); agenti frontier circa 280 s per compito quasi-statico su robot reale; loop di controllo LLM a 0.2–0.4 Hz contro circa 83 Hz necessari (Anthropic)
Costo	Walker S2 in media circa RMB 600k; rientro di 7–12 anni estimate
Certificazione di sicurezza	ISO 10218-1/-2:2025 rende la sicurezza funzionale più esplicita e assorbe i requisiti collaborativi, ma non dice nulla sull'AI (The Robot Report); ISO/IEC TS 22440 (sicurezza funzionale e AI) era ancora una seconda bozza di comitato nel novembre 2026 (ISS, ISO Open Data); la guida IEC 61508-3 sconsiglia l'AI sopra SIL 1 (Analog Devices); sotto il Regolamento macchine UE, le funzioni di sicurezza con machine learning richiedono la valutazione di un organismo notificato dal 20 gennaio 2027 (EUR-Lex)
Mani e utensili	Le mani destre gestiscono circa 30 N (≈3 kg) per dito contro circa 10 kg per una mano umana (Gasgoo); i cambi utensile sono ancora manuali
Integrazione	Letture errate dei sensori, disallineamenti di protocollo e collegamenti MES in loco; tassi di intervento quasi mai comunicati

L'implicazione per la sicurezza è diretta: gli standard né vietano le policy apprese né le riconoscono. ISO 10218-1/-2:2025 non dice nulla sull'AI; ISO/IEC TS 22440 (sicurezza funzionale e AI) è ancora a bozza di comitato e improbabile prima di fine 2027 o 2028 estimate; ISO/IEC TR 5469:2024 è solo guida, che raccomanda fallback non AI, vincoli di supervisione e voto ridondante; IEC 61508 edizione 3 è attesa all'inizio del 2027 (ISO Open Data, BSI, 61508 Association) verified. Nessuno standard pubblicato offre ancora una via accettata per certificare la percezione basata su machine learning come funzione di sicurezza. Anche le regole regionali si muovono: la Cina sta rivedendo GB 11291 (in vigore circa 2028–2029) (piano di revisione 20262543); UL 1740 ha portato ISO 10218-2 nella revisione NRTL statunitense da agosto 2025; l'UE richiede la valutazione di un organismo notificato delle funzioni di sicurezza con machine learning dal 20 gennaio 2027; e nessun umanoide è certificato per l'uso in fabbrica senza recinzione mentre ISO 25785-1 è ancora in bozza (i-SCOOP). L'architettura difendibile: il machine learning prende solo decisioni di processo ("quale parte", "questa scheda è buona"), mentre uno strato di sicurezza deterministico certificato (tipicamente PL d / Categoria 3) e un PLC di sicurezza decidono velocità, forza e arresti. Tratta qualsiasi affermazione di "AI con fotocamera per la sicurezza senza recinzione" come non provata finché un certificato PL d o SIL 2 non nomina il componente di machine learning.

FIG 32 **Cronologia di norme e regolamentazione** ***Una via di certificazione per l'AI dentro le funzioni di sicurezza non si assesterà prima del 2027–2029.***



Fonte: ISO, IEC, UL, EUR-Lex, SAC e banche dati regolatorie nazionali.

Quale strato è l'opportunità: i modelli appartengono ai giganti, i box edge ai produttori di chip, le piccole aziende all'integrazione e all'hardware di sensing

Lo strato dei modelli ha bisogno di capitale e dati e appartiene ai giganti (AgiBot ha addestrato GO-2 su "decine di migliaia di ore" (iFeng Tech); i modelli più forti sono chiusi o solo cloud). Lo strato dei box edge appartiene ai produttori di chip: Jetson Thor è lo standard di fatto, e nella Cina continentale le schede RDK di D-Robotics (D-Robotics) e i coprocessori Rockchip (CSDN) competono, per lo più senza risultati VLA pubblicati.

Le piccole aziende hanno due aperture. Una è l'hardware di sensing che compra direttamente affidabilità: il tocco porta il successo di inserimento di ContactDP dal 58.7% al 97.8%, e i sensori di forza/coppia e le fotocamere da polso fanno un lavoro simile; anche i kit di raccolta dati vendono (10jqka), benché mani, sensori e bracci stiano tutti calando di prezzo. L'altra è lo strato di integrazione tra produttori di bracci e aziende di modelli, possedendo i dati di compito, l'involucro di sicurezza e il collegamento MES di ogni stazione, che i produttori di bracci oggi non vendono come prodotto. I concorrenti cadono in quattro gruppi: leader dei bracci (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot e altri), aziende visione-più-braccio (Mech-Mind, Hikrobot, SmartMore), aziende di AI incarnata (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) e nuovi entranti dei chip automotive.

Accettazione dell'acquirente: nessun numero senza denominatore

Poiché nessun fornitore pubblica tassi di intervento, intervalli di guasto o conteggi di successo di lungo periodo, questi appartengono ai termini di accettazione del contratto. Statisticamente, 3,000 cicli di produzione consecutivi con zero guasti dimostrano un tasso di guasto di al massimo lo 0.1% al 95% di confidenza; 10 guasti su 3,000 dimostrano solo ≤0.57%; dimostrare ≤1%, ≤0.5%, ≤0.1% e ≤0.01% con zero guasti richiede 299, 598, 2,995 e 29,956 cicli **estimate, binomiale esatta**. Con questo standard, il 322/324 di Neuromeka ha un intervallo bilaterale al 95% del 97.8–99.9% e non dimostra ≥99%; i circa 6 guasti di AgiBot su 64,828 compiti implicano 99.98–100%, ma solo se "guasto" è definito come lo definisce l'acquirente **estimate**. Questi sono i requisiti di accettazione che porrei a qualsiasi

cella robot o incarnata; le soglie attingono agli obiettivi di Figure da BMW e ai casi Neuromeka e AgiBot, e sono giudizio, non uno standard estimate:

Voce	Richiedi	Soglia di superamento suggerita
Definizioni	Definizioni firmate di ciclo, successo, guasto (successo parziale, ri-presa, caduta, difetto sfuggito), intervento (tocco, teleoperazione remota, reset, e-stop) e come si calcola l'autonomia	Le definizioni vengono prima di qualsiasi metrica; nessuna percentuale senza formula
Interventi e teleoperazione	Per ora-robot, per tipo e durata; tempo medio tra interventi; minuti di teleoperazione come quota del tempo di esecuzione; operatori per robot	Stazioni strutturate ≤0.5 interventi per ora-robot (a ≥2 h di distanza); zero teleoperazione durante l'accettazione
Successo di lungo periodo	≥3,000 cicli di produzione consecutivi su ≥3 turni, ≥2 codici articolo e normale variazione di illuminazione e fornitura; guasti classificati; limite inferiore esatto	Limite inferiore ≥99.5% per stazioni in linea che fermano la linea, ≥99.0% offline; difetti di qualità sfuggiti riportati separatamente
Tempo ciclo	P50/P90/P99 per parte incluso il recupero, contro ≥30 cicli umani cronometrati alla stessa stazione	≤1.15 × P50 umano, o dentro il takt di linea
Uptime e guasti	Ore di esecuzione, registro dei fermi, guasti per componente, tempo di sostituzione parti	≥95% di uptime su ≥2 settimane; MTTR ≤20 min con ricambi in loco
Recupero	Iniezione di guasti sceneggiata: cadute, occlusione, cambio di illuminazione, arresto del nastro, perdita di MES o rete, perdita di alimentazione, e-stop	≥90% di recupero autonomo, di nuovo in produzione entro 2 min
Cambio	Ore, dimostrazioni, abilità dell'operatore e costo di raccolta dati per nuovo codice articolo o maschera	Stazioni strutturate: ≤1 turno e ≤100 demo per nuovo codice articolo
Sicurezza	Braccio e cella secondo ISO 10218-1/-2:2025, valutazione del rischio di cella, regole locali (GB 11291 nella Cina continentale; CNS 14490 più valutazione scritta a Taiwan; ANSI/A3 R15.06-2025 e UL 1740 negli Stati Uniti)	Nessuna funzione di sicurezza implementata da machine learning; uno strato di sicurezza certificato (tipicamente PL d / Cat 3) limita velocità e forza; nessun robot con gambe senza recinzione finché ISO 25785-1 non è pubblicata e valutata
Calcolo e comportamento offline	Hardware di inferenza, latenza P99, comportamento senza rete, versionamento del modello, rivalidazione dopo ogni aggiornamento	L'inferenza critica per il takt non dipende dal cloud; piena ri-accettazione dopo ogni cambio di modello, quantizzazione o motore
Evidenza	Video non modificato e log grezzi per l'intera finestra di accettazione; testimoniato dal team qualità del cliente o da una terza parte	Nessun reel promozionale o sintesi solo del fornitore
Termini commerciali	Pietre miliari di pagamento legate alle metriche di accettazione; penali di disponibilità quando fatturato per ora-robot; proprietà dei dati e regole di accesso remoto	Pagamento finale e trattenuta rilasciati solo dopo l'accettazione dei 3,000 cicli e ≥4 settimane di funzionamento stabile
Economia	Persone nette sostituite = rapporto di velocità × uptime × successo − tempo di intervento − quota di teleoperazione	Rifiuta qualsiasi proposta la cui sostituzione netta non sostiene il rientro dichiarato

Fonti: (Figure, Neuromeka, Yicai Global, i-SCOOP); le soglie sono mie proposte.

Questioni aperte

07 Cosa non può rispondere il registro pubblico, e come lo risolve un pilota

Le domande a cui le fonti pubbliche non possono rispondere cadono in cinque gruppi: a quanto si vende davvero la revisione AI e cosa pagheranno gli acquirenti; come i produttori di AOI raggruppano l'AI e aggiornano le vecchie macchine; quanto vecchi sono gli AOI installati e quanti usano già l'AI; quante persone e ore richiede davvero la revisione per linea; e come i piloti convertono e come pagano gli acquirenti. A queste si aggiunge come i modelli più recenti, i build quantizzati e i bracci robotici performano davvero nelle condizioni di un dato stabilimento. Ognuna può essere trasformata in un test con una soglia chiara. Le soglie sotto sono regole decisionali che ho derivato dalle evidenze di questo rapporto estimate ; ricalcolale una volta arrivati i primi dati di stabilimento.

Questione aperta	Come risolverla	Come appare un buon risultato
Prezzo e disponibilità a pagare per la revisione AI	Una richiesta di quotazione standard (per esempio quattro linee con AOI 2D più vecchi misti e 3D più nuovi, due turni, circa 2,000 schede per turno, 12 classi di difetto, solo on-premises, integrazione MES o IPC-CFX) inviata a più fornitori terzi; un'indagine di sensibilità al prezzo a quattro domande su 20–30 stabilimenti EMS	La maggior parte delle quotazioni entro RMB 40–120k una tantum o 20–40k l'anno per quattro linee; almeno metà degli stabilimenti accetta RMB 10–30k per linea con rientro entro 12 mesi
Raggruppamento AI dei produttori di AOI e upgrade per macchine più vecchie	Brevi interviste strutturate con circa una dozzina di produttori e distributori di AOI	Se la revisione AI è standard, opzionale o assente sui modelli 2025–26; se le macchine pre-2020 possono essere aggiornate e a che prezzo
Età degli AOI installati e quota che usa l'AI	Indagine in loco su circa 30 stabilimenti EMS, stratificata per dimensione e regione, con elenchi attrezzatura esaminati sotto NDA	Età mediana AOI ≥5 anni e ≥60% senza AI significa che il bacino di retrofit è reale
Manodopera e ore di revisione	Interviste e 2–4 settimane di log di allarmi AOI, revisione, riparazione e turni in due o più stabilimenti, con campionamento del lavoro	≥1 revisore dedicato per linea per turno e ≥3 ore di revisione per linea per turno, con almeno 0.5 FTE per linea che lo stabilimento è disposto a riassegnare
Performance della revisione AI stessa	Golden set più modalità ombra su una linea reale (capitolo 3): accordo sui dati nella settimana 0, baseline ed etichettatura nelle settimane 1–2, esecuzione ombra nelle settimane 2–4, esecuzione assistita opzionale nelle settimane 4–8	Falsi allarmi o ore di revisione tagliati ≥50–60%; limiti di errori sfuggiti per classe non peggiori della baseline umana; zero errori sulle classi critiche; latenza p95 ≤½ takt; fail-open automatico verso le persone
Qualità delle etichette-insegnante dei modelli frontier	Far girare due o tre candidati sullo stesso golden set, tre formulazioni di prompt ciascuno, una buona immagine di riferimento, ritagli di singolo componente ≥42–48 px	Accettazione automatica solo per classi con zero errori su ≥299 difetti e accordo a livello umano; altrimenti solo pre-etichettatura e triage
Build quantizzati	Cambiare solo la precisione; fissare hash dei pesi, template, preprocessing, decoding, motore e dimensione batch; disabilitare o azzerare la prefix caching; verificare che il build di riferimento riproduca i risultati del fornitore; valutare le coppie su golden set e fetta di deriva	Nessun nuovo errore sulle classi critiche; ≤1 nuovo errore in 1,000 difetti; clearance giù di ≤2 punti; ≥99.5% di decisioni identiche su tre run; altrimenti ricadi su FP8/INT8

Questione aperta	Come risolverla	Come appare un buon risultato
Hardware da linea, pulizia ed ESD	≥2 unità di ciascun box candidato stressate al doppio del tasso di allarmi atteso; soak di 72 ore al flusso d'aria reale dell'armadio; verifica ESD del coordinatore del cliente; test ISO 14644-14 per clienti ISO 7 e più puliti	p95 ≤½ takt; nessun crash o throttling su 72 ore; firma ESD; rapporto particellare alla classe giusta
Termini di pagamento, regole di pilota e conversione	Clausole di pagamento dai recenti contratti di attrezzatura di ispezione negli stabilimenti esaminati; interviste su regole di pilota, criteri di accettazione e tempo di qualificazione	Circa 30/60/10 con cassa entro 60 giorni dall'accettazione; piloti ≤8–12 settimane con un responsabile di firma nominato; ≥1 su 3 prove in conversione; un quadro che copre ≥5–10 linee nel primo anno
Dati, IT e responsabilità	Interviste e revisione contrattuale con consulente legale	Un box locale può connettersi e accedere ai dati; responsabilità limitata ai difetti del software
Conformità regionale prima di andare all'estero	Taiwan: stato scritto del cliente sotto il Cyber Security Management Act, autovalutazione E187; Stati Uniti: ruling doganale su come è classificata una cella, preventivi NRTL, verifica ITAR; Sud-est asiatico: collegamento locale BOI ed eleggibilità allowance malese, casa madre del cliente; Cina continentale: ambito CCC per PC e alimentatori raggruppati	Il cliente non è un ente controllato, e il rientro resta ≤2 anni con dazio e certificazione inclusi
Accettazione di una cella robot o incarnata	La tabella di accettazione a 12 voci del capitolo 6: definizioni firmate, ≥3,000 cicli di produzione consecutivi su ≥3 turni e ≥2 codici articolo, iniezione di guasti sceneggiata, ≥30 cicli umani cronometrati, testimoniata	Limite inferiore ≥99.5% in linea (≥99.0% offline); ciclo ≤1.15 × umano; ≤0.5 interventi all'ora e zero teleoperazione; ≥95% di uptime; MTTR ≤20 min; ≥90% di recupero
Policy ACT e VLA sui chip edge cinesi	Quattro settimane: port e micro-benchmark (≥1,000 inferenze, distribuzione della latenza, potenza, soak di 60 minuti); simulazione a ciclo chiuso con latenza misurata iniettata; A/B su robot reale contro una GPU; endurance di 8–24 ore (≥3,000 cicli)	Chunk ACT p95 ≤100 ms, VLA ≤300 ms; successo su robot reale entro 3 punti dalla GPU e ciclo ≤1.2 × umano; altrimenti mantieni la policy su Jetson o una GPU
Prezzi di cobot, AMR e integrazione	Quotazioni modello per modello da più distributori di cobot, e quotazioni di cella completa da due o tre integratori per il carico di maschere di test (una stazione, due turni)	Cella completa ≤RMB 200k e rientro ≤1.5 anni su due turni a Shenzhen
Conteggi di linee SMT a Taiwan, nel Sud-est asiatico e negli Stati Uniti	Dati IPC (ex New Venture Research) o base installata dei distributori locali di attrezzatura	Abbastanza linee servibili per regione da giustificare l'ingegneria applicativa locale

Di tutte queste, quattro decidono i numeri più sensibili di questo rapporto: il prezzo che gli acquirenti pagheranno, la manodopera che la revisione AI libera davvero, se la revisione AI manca davvero non più delle persone, e quanto veloce torna la cassa. Vale la pena risolverle prima di qualsiasi impegno significativo, da parte di uno stabilimento che considera un acquisto o di un fornitore che considera un prodotto.

Una sequenza di pilota che funziona per entrambe le parti

Per uno stabilimento, e per un fornitore che vuole un riferimento credibile, l'ordine è lo stesso:

1. **Settimane 0–2: accordarsi su dati e definizioni.** Firma l'accordo sui dati, definisci le classi di difetto e cosa conta come difetto sfuggito, e registra una baseline: falsi allarmi per scheda, secondi di revisione per chiamata, difetti sfuggiti trovati a valle.
2. **Settimane 1–4: costruisci il golden set** dall'archivio AOI e dalle schede NG conservate, etichettato in cieco e arbitrato (capitolo 3).

3. **Settimane 2–4 (o più): modalità ombra.** Il modello registra le sue decisioni mentre le persone decidono come prima; confronta per classe.
4. **Settimane 4–8: esecuzione assistita su turni casuali,** con il modello che chiude solo le classi che hanno superato il gate e tutto il resto che va alle persone.
5. **Accettazione, per iscritto, prima di scalare:** i risultati per classe, la latenza, il comportamento fail-open, e chi possiede modelli ed etichette. Scala linea per linea, e rivalida su ogni nuovo prodotto, versione di modello, quantizzazione o cambio di motore.

Limiti di questo rapporto

Questa è ricerca documentale. Non ho condotto interviste primarie con gli stabilimenti per questa edizione; molti dati di linea provengono dai fornitori e non sono stati verificati in modo indipendente. Diversi siti governativi, banche dati di gare e piattaforme di reclutamento non erano accessibili, quindi le aliquote di sussidio per l'upgrade 2026 di Shenzhen, la guida salariale del Delta del Fiume Azzurro, i prezzi di transazione della revisione AI, i revisori per linea e i tassi di conversione dei piloti sono stime o analogie; ognuno appare sopra come questione aperta. I tassi di cambio sono la parità centrale della People's Bank of China del 9 novembre 2026 ($1 = RMB 6.7188) e, per Taiwan, il fixing SMBC dello stesso giorno ($1 = NT$32.12); il dong vietnamita usa dati di maggio 2026. Tutte le soglie sono regole decisionali derivate dalle evidenze citate, non standard.

L'autore

Tan Shuai

Tan Shuai è un tecnologo e fondatore con oltre vent'anni di lavoro tra chip, hardware e software di sistema, e 20 brevetti concessi. Il suo lavoro lo ha portato attraverso la Silicon Valley, New York, Chicago, Taiwan, la Greater Bay Area (Shenzhen e Hong Kong), Vietnam e Malaysia, con una lunga attenzione alla produzione avanzata e alle catene di fornitura. Oggi lavora come ricercatore e sviluppatore AI, concentrandosi su come l'AI viene effettivamente utilizzata in fabbrica.

Accoglie volentieri conversazioni con chi lavora sull'AI nella produzione, sia che si tratti di costruire un team, cercare un partner o confrontare appunti: i@tanshuai.com · tanshuai.com

Note e avvertenze

Questo rapporto è ricerca documentale indipendente basata su documenti pubblici disponibili fino all'11 novembre 2026. È fornito solo a scopo informativo e non è consulenza di investimento, legale, fiscale o ingegneristica. Nomi di prodotti, prezzi, rilasci di modelli, classificazioni di export, dazi e regolamenti cambiano rapidamente; verifica qualsiasi cosa su cui fai affidamento contro la fonte primaria. Le cifre marcate vendor claim o estimate vanno lette come tali. I nomi di aziende e prodotti appartengono ai loro proprietari e sono usati solo a scopo identificativo; nessuna azienda ha revisionato o sponsorizzato questo rapporto.

© 2026 Tan Shuai. Puoi condividere questo rapporto invariato con attribuzione.

PRODUZIONE ELETTRONICA · ISPEZIONE AI × ROBOTICA

Ispezione visiva AI e bracci robotici nelle fabbriche di elettronica: valutazione 2026

Ogni cifra chiave riporta la propria fonte e un tag di confidenza: verified, single source, vendor claim o estimate. Ciò a cui il registro pubblico non può rispondere è raccolto nel capitolo 7, con come un progetto pilota risolve ogni questione.

Tan Shuai
tanshuai.comi@tanshuai.com

EDIZIONE PUBBLICA · DATI AGGIORNATI ALL'11 NOVEMBRE 2026

