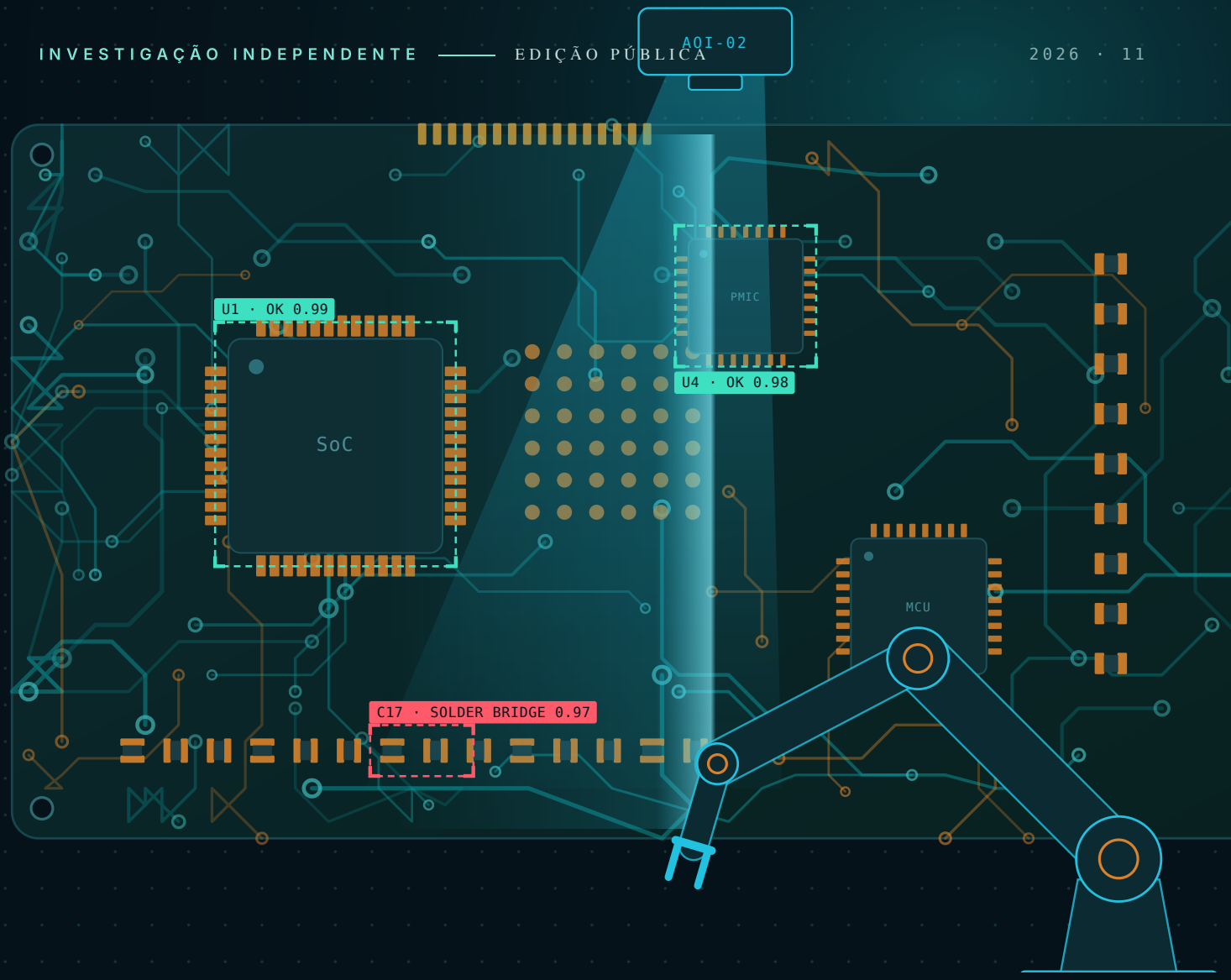


FABRICO DE ELETRÓNICA · INSPEÇÃO COM IA × ROBÓTICA

Inspeção visual com IA e braços robóticos em fábricas de eletrónica

Avaliação 2026

Configurações comuns · Viabilidade regional · Implantação no local · Modelos professor-aluno
Retorno do investimento · Concorrência · O que só um piloto pode responder

CONCLUSÃO PRINCIPAL

Vale a pena fazer revisão com IA, de forma restrita. Os braços robóticos incorporados podem esperar.

Tan Shuai
tanshuai.comi@tanshuai.com

Dados até 11 de novembro de 2026
China continental · Taiwan · Sudeste Asiático · Estados Unidos

CONTEÚDO

Conteúdo



◆ **SUMÁRIO EXECUTIVO** — Vale a pena, mas de forma restrita, agora e com parceiros — 5

01 **A linha hoje** — O que as fábricas de eletrónica executam hoje e o que é tecnicamente viável — 9

02 **No local** — Executar a IA dentro da fábrica: salas limpas, isolamento de rede, hardware e quantização — 17

03 **Professor e aluno** — Professor e aluno: os modelos de fronteira podem pré-etiquetar dados de inspeção; para os braços robóticos, os verdadeiros professores continuam a ser pessoas — 26

04 **Custos e retorno** — Custos e retorno: a revisão com IA amortiza-se em meses, a alimentação de máquinas em anos, os braços de retirada nunca — 35

05 **Mercado e concorrência** — Mercado, concorrência e como os compradores pagam — 43

06 **IA incorporada** — IA incorporada: primeiro muda como os robôs são ensinados, depois como se movem — 52

07 **Questões em aberto** — O que o registo público não consegue responder e como um piloto resolve isso — 58

◆ **Sobre o autor** — Tan Shuai — 61

Etiquetas de confiança no texto

- **verified** duas ou mais fontes independentes, ou dois documentos primários independentes, concordam
- **single source** apenas uma fonte, incluindo documentos primários
- **vendor claim** publicado pelo vendedor do produto ou serviço
- **estimate** aritmética ou inferência do autor a partir de dados citados

Como ler este relatório

Os números-chave trazem uma etiqueta de confiança. **verified** significa que duas ou mais fontes independentes concordam, ou que dois documentos primários independentes concordam; **single source** significa apenas uma fonte (incluindo documentos primários); **vendor claim** significa publicado pelo vendedor do produto ou serviço, mesmo que amplamente repetido; **estimate** significa a minha própria aritmética ou inferência a partir de dados citados, com os dados de entrada indicados. "Primário" descreve o tipo de fonte (relatório anual, prospeto, regulamento, norma, estatísticas oficiais); não altera a classe de confiança. Quando as fontes entram em conflito, os valores são apresentados lado a lado e não são calculadas médias. As conversões de moeda usam a paridade central do Banco Popular da China de 9 de novembro de 2026, $1 = RMB 6.7188 (Xinhua), e para Taiwan a fixação do SMBC no mesmo dia, $1 = NT$32.12 (SMBC). Os dados de mão de obra e subsídios do continente usam o Delta do Rio das Pérolas, onde a indústria de eletrónica está mais concentrada (dados salariais oficiais de Shenzhen), com o Delta do Rio Yangtzé para comparação. As alegações de entrevistas do setor são tratadas como alegações a verificar, nunca como factos. As questões que o registo público não consegue responder estão reunidas no capítulo 7.

FIGURAS

Figuras



01	Onde a IA se encaixa hoje no chão de fábrica	7
02	Uma linha SMT em volume: as pessoas ficam atrás do AOI	10
03	Cerca de 97–99 % dos alarmes AOI são falsos	11
04	Uma câmera de uso geral não é um AOI	13
05	Uma implantação isolada (air gap)	17
06	As zonas limpas e ESD decidem onde o computador pode ficar	18
07	Limites de velocidade por uso	20
08	Computação local: memória, preço e o que se pode comprar legalmente na China continental	21
09	Uma grande estação de trabalho não é um assistente compartilhado	22
10	Quantização: 8-bit é seguro, 4-bit depende da receita, 3-bit e menos é descartado	23
11	Candidatos a professor: as pontuações genéricas dobraram em uma geração	27
12	«Há defeito?» continua difícil para modelos gerais	28
13	Dois pipelines professor–aluno	29
14	Qual deve ser o tamanho do conjunto de referência	30
15	Latência de política edge	32
16	Alunos robô: perto do nível de produção só quando o posto é estruturado	33
17	O que custa por ano um trabalhador de inspeção ou movimentação	36
18	A escada de preços	37
19	Retorno por região	39
20	Por que extrair placas ruins não se paga	40
21	Tarifa dos EUA sobre robôs, AOI e computadores edge	41
22	Os tamanhos de mercado diferem em mais de dez vezes	44
23	Ascendente: 40.000 linhas → 16.000 atendíveis → ≈320 M RMB pontuais	45
24	A janela de retrofit é de cerca de cinco anos	46
25	A cunha de revisão está cercada por três lados	46
26	A margem está no hardware industrializado	48
27	Como os compradores pagam	49
28	Dias de vendas a receber	50
29	A pilha de IA corpórea	52
30	O que as implantações robóticas em fábrica realmente divulgam	54
31	Mão de obra líquida substituída por robô	54
32	Cronologia de normas e regulamentação	56

SUMÁRIO EXECUTIVO

Vale a pena, mas de forma restrita, agora e com parceiros

CONCLUSÃO PRINCIPAL

O veredito: para as fábricas de eletrónica, o que está maduro em 2026 e pelo que as fábricas pagarão é a revisão com IA e a inspeção de nicho sobre imagens de AOI, mais a alimentação programada de máquinas com robôs, não "um modelo grande a olhar para placas mais robôs incorporados a apanhá-las".

0.46–0.65 anos

Retorno da revisão com IA nos EUA, em Taiwan e em Shenzhen, a RMB 50 mil por linha

≥95% · <20 PPM

Cobertura e fugas da revisão com IA interna de uma EMS de topo: a fasquia para vender a EMS de nível 1

100–700 RMB M

Mercado de retrofit de terceiros aproveitável na China continental, pagamento único, numa janela de cerca de cinco anos

0 pontuações

Pontuações publicadas em benchmarks de defeitos industriais dos modelos de topo de novembro de 2026: professores, não inspetores

1.1–1.5 anos

Retorno de uma célula de alimentação de máquinas com cobot a dois turnos nos EUA, em Taiwan e em Shenzhen

0.003–0.14 FTE

Mão de obra que um braço robótico de retirada poupa por turno numa linha SMT: nunca se amortiza

Inspeção
ROI

A aprendizagem profunda supervisionada aplicada às imagens que um AOI existente já produz reduz os falsos alarmes ou as horas de revisão em 60–90% (alegações de fornecedores). A DBG, uma grande empresa de EMS, refere no seu relatório anual que o seu sistema interno cobre mais de 95% dos componentes sinalizados, com fugas abaixo de 20 PPM e mais de 85% menos revisão humana; essa é a fasquia que um produto de terceiros tem de superar para vender a EMS de nível 1. A RMB 50,000 por linha, a revisão com IA amortiza-se no prazo de um ano na China continental, em Taiwan e nos EUA (Shenzhen 0.65 anos, Taiwan 0.53, EUA 0.46), as células de alimentação de máquinas com cobot em 1.1–1.5 anos e o Sudeste Asiático em 2–6 anos, enquanto um braço robótico adicionado a uma linha SMT apenas para retirar placas defeituosas nunca se amortiza em lado nenhum, porque, acima de 90% de rendimento à primeira passagem, a retirada ocupa apenas 0.003–0.14 de uma pessoa por turno estimate.

Modelos
ROBÔS

Os grandes modelos multimodais ainda não podem ser o inspetor principal. Até 11 de novembro de 2026, nenhum dos modelos de topo (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 e outros) tem uma pontuação publicada em qualquer benchmark de defeitos industriais ou de PCB, e no desafio PCBA da ACM MM 2026 os cinco melhores sistemas, treinados em 6,600 imagens reais de linha, obtiveram apenas 77–82 em "há um defeito". Os modelos de fronteira podem pré-etiquetar e fazer triagem. Para os braços robóticos, os verdadeiros professores continuam a ser as demonstrações humanas, a aprendizagem por reforço em robôs reais e a simulação: os alunos chegam a 90–99.4% em estações fixas e estruturadas, mas apenas a 30–80% em transportadores em movimen-

to, peças nunca vistas e novos produtos; um agente de fronteira completou 104 de 108 tarefas quase estáticas com robôs reais, a 280 segundos cada.

No local
IMPLANTAÇÃO

A implantação isolada da rede em fábricas limpas e controladas é totalmente praticável. Uma sala SMT comum é normalmente uma área protegida contra ESD e não uma sala limpa classificada: uma caixa Jetson sem ventoinha ou com NPU nacional fica na linha e os servidores GPU ficam na sala de servidores. Segundo as próprias classificações de exportação da NVIDIA, o Jetson Thor e o DGX Spark podem ser vendidos a uma fábrica comum do continente sem licença dos EUA; o DGX Station não pode. A precisão por defeito é de 8 bits; o 4-bit com quantização consciente do fornecedor é utilizável com condições; 3 bits ou menos está fora.

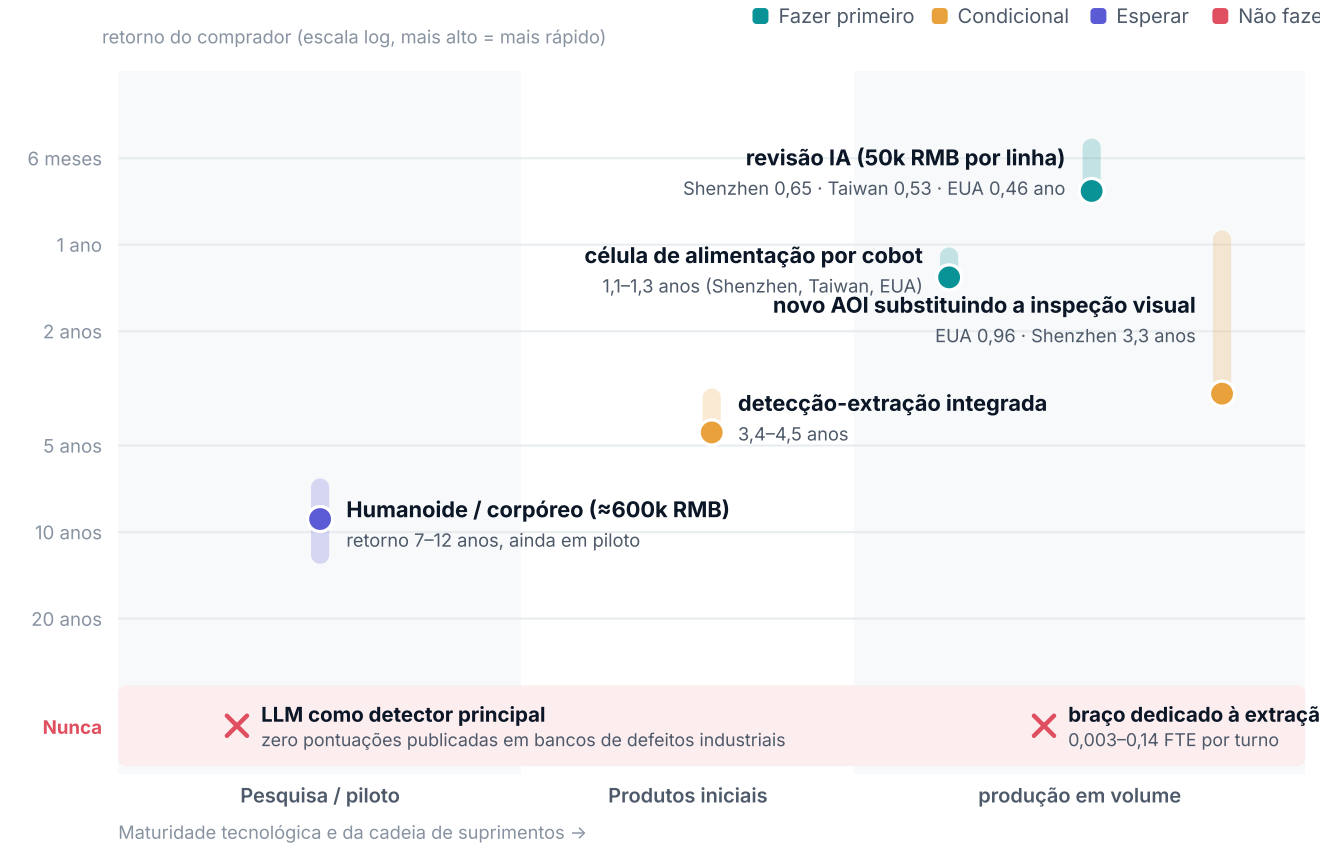
Mercado
CONCORRÊNCIA

O mercado é real mas limitado. A China continental tem cerca de 35,000–46,000 linhas SMT e cerca de 26,000–66,000 AOIs inline sem IA; o mercado aproveitável para retrofits de terceiros é de cerca de RMB 100–700 milhões de pagamento único, numa janela de cerca de cinco anos, e os fabricantes de AOI começaram a integrar a revisão nas máquinas novas estimate. A concorrência já é forte: a cunha de revisão está cercada pelos fabricantes de AOI, pelas equipas internas de EMS e pelos fornecedores de plataformas, e a margem está no hardware transformado em produto.

Questão	Julgamento	Números-chave (fontes e confiança nos capítulos)
① **O que é que as fábricas de eletrónica comuns executam hoje?**	Uma linha SMT de volume é "SPI 3D + AOIs antes e depois do reflow + raio-X por amostragem + ICT/FCT + reparação e inspeção visual", com revisão humana atrás de cada AOI. A IA entra na revisão de três formas: integrada pelos fabricantes de AOI, construída internamente pelas EMS de topo ou adicionada por terceiros. A montagem e a manipulação usam braços, SCARAs e AMRs com perceção aprendida e movimento programado; a IA incorporada ainda está em fase piloto	Takt ≈13–22 s por placa; um ODM orçamentou RMB 30.6M de AOI para 18 linhas, ≈3–5 AOIs por linha; ≈97–99% dos alarmes do AOI são falsos alarmes
② **Até onde vai a tecnologia, pode o dinheiro comprar zero pessoas e o que pode cada região usar?**	A inspeção é de grau industrial como "detetor dedicado + revisão com IA + algumas pessoas"; mesmo dinheiro ilimitado não chega a zero pessoas. O continente usa apenas modelos nacionais ou pesos abertos locais; as fábricas privadas em Taiwan podem usar quase tudo, mas as fabs e as infraestruturas críticas excluem na prática as marcas do continente; o Sudeste Asiático tem as menores restrições; os EUA têm a pilha completa mais direitos aduaneiros e certificação	A revisão com IA reduz os falsos alarmes ou as horas de revisão em 60–90% vendor claims ; uma câmara 4K sobre um painel inteiro ≈78 µm por píxel contra 8–20 µm de um AOI; os braços fabricados no continente pagam ≈40% de direito aduaneiro combinado nos EUA, os instrumentos do tipo AOI ≈37.5%
③ **Pode funcionar no local, em salas limpas e fábricas isoladas da rede, em que hardware, com rapidez e precisão suficientes?**	Sim; as restrições são de procedimento, não de computação. Uma caixa sem ventoinha de ≤100 W na linha, servidores GPU na sala de servidores; Thor para as estações, um Mac Studio ou um aparelho nacional para o Q&A da fábrica; sem DGX Station no continente. A velocidade é adequada quando o tamanho do modelo corresponde à tarefa; manter 8 bits, usar 4 bits com quantização consciente do fornecedor com condições, nunca 3 bits	≈0.5 s por recorte para um modelo de visão 8B no Thor estimate ; kit de desenvolvimento Thor RMB 59,104 na China; DGX Station $123–140k, 139 tokens/s no total para 32 utilizadores num modelo 743B

Questão	Julgamento	Números-chave (fontes e confiança nos capítulos)
④ **Pode o modelo de topo mais recente atuar como professor ao longo de dezenas de rondas e ser destilado em alunos locais, incluindo para fabrico flexível e braços robóticos?**	Inspeção: pré-etiquetagem e triagem; aceitar automaticamente as etiquetas do professor apenas para classes com "≥299 defeitos, zero falhas" num conjunto dourado. Robôs: os professores continuam a ser demonstrações humanas, RL em robôs reais e simulação, e os alunos aproximam-se do grau de produção apenas em estações estruturadas	Nenhum modelo de topo de novembro tem uma pontuação de defeitos industriais; alunos 97.8% em conectores treinados, 29–67% em nunca vistos, 37–59% em transportadores em movimento; modelos de fronteira a acionar articulações diretamente 0–5.5%
⑤ **Quanto custa e qual é o ROI por região?**	A revisão com IA amortiza-se no prazo de um ano na China continental, em Taiwan e nos EUA; alimentação com cobot 1.1–1.5 anos; Sudeste Asiático 2–6 anos; um braço de retirada numa linha SMT nunca se amortiza	Caixa de revisão RMB 50k por linha: Shenzhen 0.65 anos, Taiwan 0.53, EUA 0.46, Vietname 2.2 **estimate** ; preço aceitável de complemento de terceiros ≈RMB 10–30k por linha de pagamento único ou RMB 5–10k por ano
⑥ **A concorrência já é suficientemente forte, e é melhor começar agora ou esperar?**	A procura não está saturada, mas a oferta está lotada; a IA dos OEM limita a janela de retrofit a cerca de cinco anos. Entrar agora com um produto restrito através de parceiros de canal, sem lutar de frente com os fornecedores de plataformas; esperar pelos dados de produção de 2027–2028 antes de apostar na IA incorporada	≈26–66 mil AOIs inline sem IA no continente, a encolher 8–12% ao ano **estimate** ; o fornecedor líder de visão com IA teve receitas em 2025 de RMB 1.086 mil milhões e um prejuízo líquido de RMB 991 milhões; a receita 3C do comparável cotado mais próximo caiu 79.9% em termos homólogos entre janeiro e maio de 2026

FIG 01 **Onde a IA se encaixa hoje no chão de fábrica** ***A revisão IA de falsos alarmes AOI e a alimentação de máquinas por cobot são maduras e se pagam rápido; um braço de extração dedicado e um LLM como inspetor principal não valem a pena.***



Fonte: Retorno do modelo do autor (capítulo 4); o eixo de maturidade é um julgamento do autor.

O que construir e o que não construir

	Recomendação
Construir	Revisão com IA e classificação de defeitos sobre imagens de AOI, para EMS de média dimensão com AOIs antigos ou de marcas mistas e sem atualização de IA do fabricante do AOI; inspeção que o AOI de SMT cobre mal (furos passantes e soldadura manual, revestimento conforme, cabos, conectores e FPCs, aspeto final e parafusos); células de carga e descarga com cobot onde cada peça é manipulada (fixações de teste, tabuleiros, embalamento), com a inspeção como complemento; um conjunto dourado e modo sombra numa linha real como base de cada alegação
Não construir	Uma plataforma geral de "IA mais braço robótico"; um LLM multimodal como inspetor principal; um braço robótico numa linha SMT apenas para retirar placas NG (um desviador ou buffer de NG faz isso); humanoides e robôs como serviço; modelos de fronteira dentro de uma malha de controlo ou de uma função de segurança; pilhas de memória GPU nas estações para executar modelos "de tamanho real"; vendas diretas a fabs, infraestruturas críticas de Taiwan ou fábricas de eletrónica de defesa dos EUA
Com quem trabalhar	Fabricantes e distribuidores de AOI (interfaces de canal e de imagem; os distribuidores trouxeram 46.68% da receita de 2025 da Sinic-Tek); fabricantes de cobots (braços certificados e funções de segurança); empresas de EMS com 10–40 linhas (expansão progressiva e depósitos); integradores locais na Tailândia e noutros locais (para cumprir as regras de ligação local da BOI)

A linha hoje

01 O que as fábricas de eletrónica executam hoje e o que é tecnicamente viável

A linha de volume padrão: vários AOIs, raio-X por amostragem, revisão central, 13–22 segundos por placa

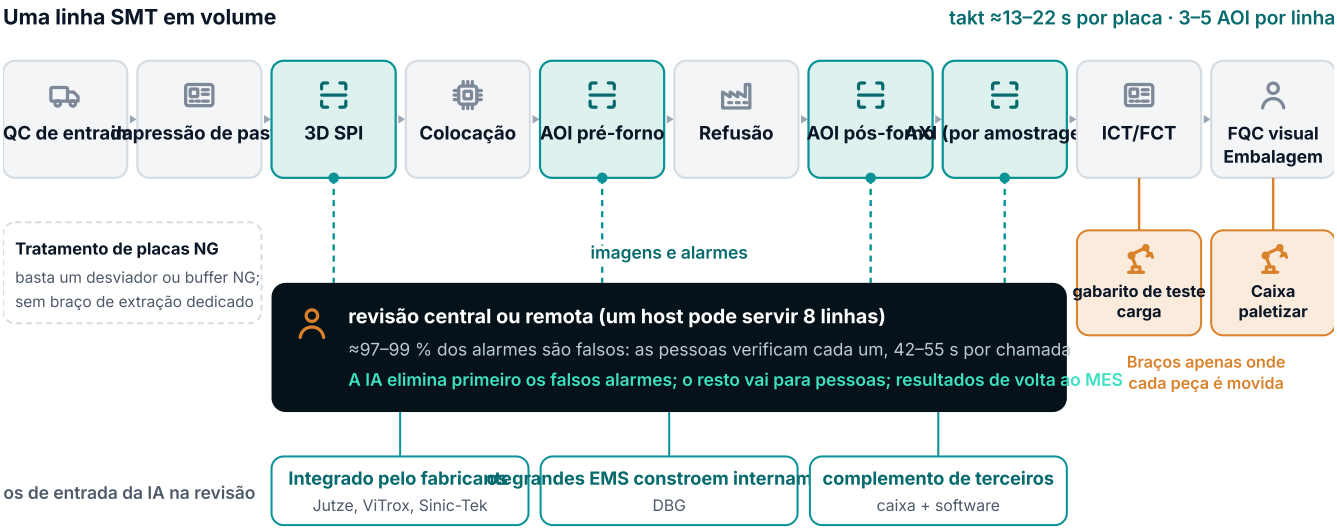
Os documentos primários de ODMs cotados mostram como é realmente uma linha de volume de 2025–2026. O prospeto de Hong Kong da Longcheer (dados até 30 de setembro de 2025) lista as suas verificações de qualidade SMT como "inspeção por raio-X, inspeção de pasta de solda (SPI), inspeção ótica automatizada (AOI), teste funcional de PCBA e inspeção visual"; o seu fluxo de processo passa por AOI antes do reflow, AOI depois do reflow, raio-X por amostragem, placas NG para uma estação de reparação SMT e, depois, uma verificação visual final. As suas fábricas de Huizhou, Nanchang e Vietname têm 34, 24 e 8 linhas SMT e 67, 60 e 13 linhas de montagem final (prospeto H-share da Longcheer) single source, documento primário. A fábrica de portáteis de 2023 da Huaqin planeou 18 linhas SMT com um orçamento de AOI de RMB 30.6 milhões, cerca de RMB 1.7 milhões por linha, o que, aos preços nacionais de AOI 3D, dá três a cinco AOIs por linha para inspeção antes e depois do reflow de placas de face dupla (prospeto A-share da Huaqin) documento primário + estimate. Uma linha de volume tem, portanto, normalmente dois a quatro pontos de revisão, não um.

As máquinas já são altamente automatizadas: a Huaqin refere 96% de automação nas suas linhas SMT (prospeto H-share da Huaqin) single source. As pessoas restantes estão na revisão, reparação, inspeção visual, teste e montagem final, e nenhuma empresa publica quantos revisores tem por linha (capítulo 7).

Três estimativas independentes concordam num takt de **13–22 segundos por placa** (um processo ambiental de 2026 para cinco linhas de Hebei (EIA do projeto SMT de Hebei), as linhas de portáteis da Huaqin e um caso da iFactory), ou 7–11 segundos por face para placas de face dupla estimate. A parte difícil de uma linha de elevada mistura não é o volume, mas a introdução de novos produtos e a troca de referência, e cada troca é outro arranque a frio para um modelo de IA.

Do lado do software, as estações de revisão dos próprios fabricantes de AOI já estão centralizadas: o software de estação de reparação da Sinic-Tek permite que "um host controle o equipamento de oito linhas SMT" (prospeto da Sinic-Tek) single source, primário, 2022, e depois de a Jabil ter implementado o controlo remoto de AOI ViTrox V-ONE em 2026, um engenheiro pôde gerir 5–8 AOIs remotamente contra 1–2 localmente (AtlasPCB) vendor claim. Um produto de revisão com IA de terceiros liga-se exatamente a esta estação central, através da exportação de imagens e resultados do AOI (IPC-CFX, IPC-HERMES-9852 ou um formato próprio do fornecedor). Os próprios AOIs estão a adicionar IA rapidamente: a Yamaha diz que o seu julgamento com IA demora cerca de um segundo por chamada, ajuda os inspetores após 30 imagens NG e atinge qualidade "muito alta" após 1,000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Uma linha SMT em volume: as pessoas ficam atrás do AOI** ***A mão de obra se concentra nos postos de revisão após a inspeção óptica automática. É aí que a IA entra.***



Fonte: Documentos públicos e material de fornecedores (Huaqin, DBG, Sinic-Tek e outros); compilado pelo autor.

Etapa	Produtos e tecnologia comuns	Fornecedores representativos	Até onde a IA chegou
Após a impressão	SPI 3D (luz estruturada, moiré, alguma triangulação a laser)	Koh Young, Sinic-Tek, TRI	Julgamento de pasta com IA dos OEM em produção; a Jutze diz que alguns modelos são “totalmente IA”
Após a colocação e o reflow	AOI 2D/3D com LEDs multiângulo e multicolor e luz estruturada, 8–20 µm por píxel, câmaras de 8–12 MP	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek e outros	Revisão com IA dos OEM (Jutze em produção, ViTrox V510Ai, Sinic-Tek com a Tencent); as EMS de topo constroem as suas (DBG); complementos de terceiros
Juntas inferiores e ocultas	Raio-X (AXI) por amostragem ou inline	—	A Hikvision alega que o seu modelo Guanlan avalia mal os vazios de raio-X menos de 1% das vezes vendor claim
Revisão e reparação	Estação de revisão central (um host, oito linhas), revisão remota, estação de reparação	Software dos fabricantes de AOI, ViTrox V-ONE	A IA fecha os falsos alarmes, as pessoas tratam do resto
Inspeção visual, IQC, FQC	Maioritariamente manual; algumas máquinas de aspeto com IA e carga por robô	SmartMore e outros	Máquinas de aspeto em produção; sem dados públicos de efetivos para IQC/FQC
Manipulação e alimentação de máquinas	Braços de seis eixos e cobots, SCARA, AMRs, orientação por visão 2D/3D	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Perceção aprendida mais movimento programado em produção; VLAs e humanoides ainda em fase piloto
Dados e sistemas	MES, IPC-CFX, IPC-HERMES-9852	—	Resultados escritos de volta no MES; o suporte em equipamento mais antigo é desigual

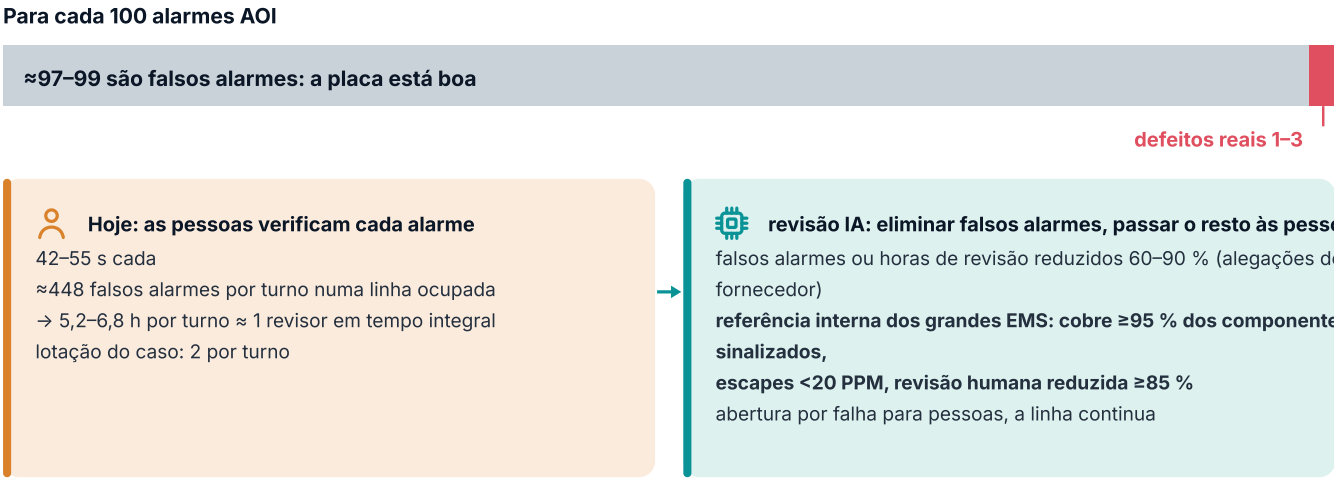
Fontes da tabela: os documentos acima e (Koh Young, relatório anual 2025 da Jutze, relatório anual 2025 da Sinic-Tek, ViTrox, Sohu sobre Hengtianyi, Xinhua sobre Hikvision).

As pessoas sentam-se atrás do AOI, porque a maioria dos alarmes é falsa

A cadeia de inspeção numa linha SMT/PCBA comum funciona assim: SPI 3D após a impressão para pasta em falta, insuficiente ou em excesso, pontes e desvios; AOI após a colocação (opcional) e após o reflow para peças em falta, deslocadas, rodadas, invertidas ou em lápide e volume de solda; raio-X quando há BGAs ou outras peças de terminação inferior; depois teste em circuito ou por sonda voadora, teste funcional e inspeção visual manual. Cada AOI tem um posto de revisão e uma estação de reparação atrás de si (prospeto da Sinic-Tek, JLCPCB) verified. O AOI pós-reflow é o mais comum e o de melhor valor, mas vê apenas forma e área; juntas, vazios e molhagem sob um BGA exigem raio-X ou teste elétrico (JLCPCB).

A mão de obra está na revisão porque a maioria dos alarmes é falsa. Uma EMS do Reino Unido registou **54,019 falsos alarmes contra 544 defeitos reais em 917 números de peça, cerca de 99% dos alarmes** (arXiv 2505.03576) single source, preprint; na fábrica da Siemens em Rastatt, **97.5% das placas que o AOI rejeitou estavam boas** (estudo de caso da Siemens) vendor claim. Uma linha ocupada vê cerca de 448 falsos alarmes por turno, a 42–55 segundos cada, 5.2–6.8 horas por turno, cerca de um revisor a tempo inteiro (iFactory) vendor claim; as horas são uma estimate.

FIG 03 **Cerca de 97–99 % dos alarmes AOI são falsos** ***Hoje uma pessoa confirma cada chamada; a revisão IA elimina primeiro os falsos e passa o resto às pessoas.***



Fonte: Estudo de caso Siemens, iFactory (alegação do fornecedor), relatório anual DBG 2025; horas de revisão estimadas pelo autor.

A partir de "horas de revisão por turno = placas × falsos alarmes por placa × segundos por chamada", um revisor consegue cobrir cerca de 0.7–1 linha de volume com um AOI baseado em regras. Quando a IA remove 60–90% dos falsos alarmes, só o tempo de julgamento permitiria 2–14 linhas por pessoa, mas as deslocações, o manuseamento do buffer de NG e as passagens para reparação limitam a cobertura local a duas a quatro linhas adjacentes; só a revisão central ou remota se aproxima do teto teórico estimate; os segundos por chamada vêm de uma fonte. Como os AOIs comuns digitalizam 60–80 cm² por segundo (guia de AOI 2026 da Yunheng), uma face de 20×15 cm demora 4–5 segundos; a IA só tem de processar as poucas regiões sinalizadas dentro de um takt, não voltar a digitalizar a placa.

O que já funciona em produção: classificação supervisionada sobre as próprias imagens do AOI

O padrão comprovado é a classificação supervisionada com aprendizagem profunda sobre as imagens do AOI que uma fábrica já possui. O complemento da Fujitsu no AOI da ADLINK em Taiwan reduziu os falsos alarmes em **mais de 90%** (caso Fujitsu); a OKI usa IA desde julho de 2026 para reduzir a confirmação manual após o AOI em cerca de **80%** no seu próprio negócio de EMS (Circuits Assembly); a Siemens refere 60% menos análise manual, €160,287 poupados no primeiro ano e uma amortização de oito meses (estudo de caso da Siemens) todas vendor

claims. Há duas coisas comuns a estes casos: quase nenhum publica uma taxa de fuga, e todos mantêm pessoas nos alarmes restantes. O único caso que dá um valor de fuga (0.8% para 0.3%, contra um contratual de 0.5%) é internamente inconsistente (iFactory). A IA reduz as horas de revisão, não o próprio trabalho de revisão.

A revisão com IA passou dos complementos de terceiros para as funcionalidades dos OEM e os sistemas internos, o que muda aquilo contra o que um recém-chegado é medido. O relatório anual de 2025 da Jutze lista a sua "atualização de AOI 3D" como em produção em volume e diz que "usa algoritmos de IA para remover a revisão manual na extremidade de inspeção" (relatório anual 2025 da Jutze); a Sinic-Tek desenvolveu um "sistema de revisão manual assistida por IA multimodal" desde 2019 e está a codesenvolver um AOI 3D de topo e base com a Tencent, visando 99.5%+ de cobertura de defeitos e 98% de precisão de reconhecimento (relatório anual 2025 da Sinic-Tek) vendor claims em documentos primários. As grandes empresas de EMS constroem as suas. O relatório anual de 2025 da DBG lista o "julgamento com IA para inspeção visual" como um projeto concluído que "cobre mais de 95% dos componentes sinalizados pelo AOI", com uma "taxa de fuga abaixo de 20 PPM" e "mais de 85% menos revisão manual", totalmente rastreável (relatório anual 2025 da DBG) single source, primário. Esse é o único valor de fábrica de eletrónica que encontrei que dá cobertura, fugas e redução em conjunto. **Um produto de terceiros vendido a EMS de nível 1 tem de igualar os 95% de cobertura, 20 PPM de fugas e 85% de redução da DBG**; abaixo do nível 1, o benchmark são os 50–85% de redução que os fornecedores anunciam.

A deteção de anomalias treinada apenas em peças boas é rápida (o EfficientAD-S demora 2.2 ms por imagem (EfficientAD, WACV 2024)), mas é fraca em anomalias "lógicas" como peças em falta ou erradas, e os melhores métodos ficam em média abaixo de 60% de AU-PRO no mais difícil MVTec AD 2 (MVTec AD 2) single source. Peças em falta e erradas são melhor verificadas contra o CAD e a BOM.

O teto a qualquer custo: camadas de máquinas, muito poucas pessoas, mas não zero

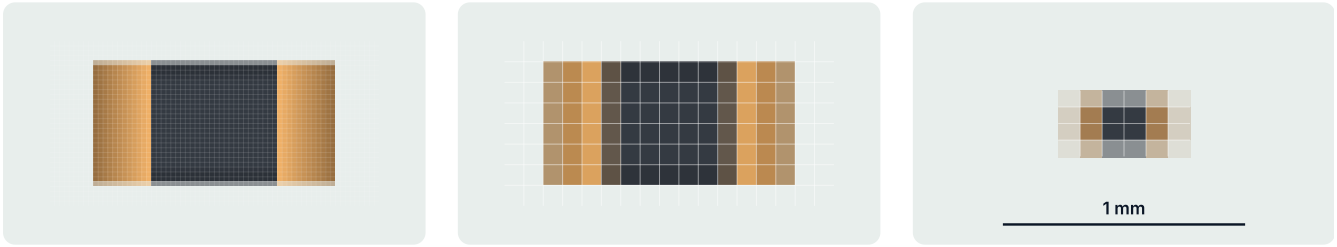
Não encontrei nenhuma fábrica de eletrónica que tenha removido totalmente a revisão manual e atingido zero fugas. A prática para produtos IPC-A-610 Classe 3 (alta fiabilidade) continua a ter inspetores treinados a decidir os casos-limite que os limiares automáticos não conseguem classificar, e um montador de PCB chama à "inspeção automatizada pode substituir a assinatura humana" um "equívoco comum" (PCBCart) single source. A linha de base humana também precisa de correção: citando um estudo da Sandia de 2012, um único inspetor visual apanha, na melhor das hipóteses, cerca de **80%** dos defeitos, e dois em série cerca de 96% (Instrumental) single source, citado por um fornecedor. A alegação defensável é "IA mais um pequeno ciclo de revisão humana supera a inspeção totalmente manual", não "a IA substitui a inspeção".

Sem limite de orçamento, o teto é SPI 3D, AOI 3D, raio-X, revisão com IA e estações de aspeto com IA empilhados, com pessoas apenas para os casos-limite, novos produtos e trocas, novos tipos de defeito e assinatura do cliente. As fábricas de topo referem 85–92% menos inspeção manual (I-Connect007, 2021, WiAdvance) vendor claims; a minha estimativa do teto sem orçamento é mão de obra de revisão a 8–15% da atual estimate. A montagem final continua intensiva em mão de obra: a automatização dos botões do iPhone 16 foi abandonada por causa de taxas de defeito elevadas (9to5Mac) single source.

Uma câmara geral de alta resolução não é um AOI

Os AOIs de produção captam imagem ladrilho a ladrilho a 8–20 µm por píxel, com câmaras de 8–12 MP, iluminação programável multiângulo multicolor e 3D por luz estruturada (Koh Young, Saki) verified. O setor notou há muito que uma peça 01005 cobre apenas cerca de 8×16 píxeis a 25 µm por píxel, "possivelmente insuficiente" para um julgamento fiável (Circuits Assembly, 2007). Pela mesma aritmética estimate, uma câmara 4K sobre um painel de 300 mm dá cerca de **78 µm por píxel**: uma peça imperial 0402 tem cerca de 13×6 píxeis, suficientes para a presença e um desvio grosseiro, mas não para a forma da junta, e uma 01005 tem cerca de 5×3 píxeis, efetivamente invisível. Atingir 15 µm por píxel sobre um painel inteiro exigiria cerca de 330 megapíxeis, e é por isso que os AOIs movem uma câmara sobre a placa. Quando alguém disser "0402", verifique qual deles quer dizer: o 0402 métrico é o 01005 imperial, muito além de qualquer câmara de grande angular.

FIG 04 **Uma câmera de uso geral não é um AOI** ***Uma câmera 4K sobre um painel de 300 mm dá ≈78 µm por pixel: um 0402 encolhe para ≈13×6 pixels e um 01005 desaparece.***



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

Câmera 4K sobre um painel de 300 mm: ≈78 µm/px
O 0402 encolhe para ≈13×6 px: só dá para presença e desvio grosseiro

Mesma câmera sobre um 01005
≈5×3 px: efetivamente invisível

Fonte: Circuits Assembly; cálculo do autor para um painel de 300 mm e um sensor 4K. As grades de pixels são ilustrativas.

Nas linhas que já têm AOIs, o melhor lugar para a IA é por cima das imagens do AOI. Só as linhas sem AOI, ou as estações de fim de linha, justificam uma estação de câmara fechada com iluminação controlada; uma câmara aberta de alta resolução sobre a linha é a mais exposta à luz ambiente e às sombras (CSDN).

Montagem e manipulação: perceção aprendida mais movimento programado está em produção

O relatório de 2026 da IFR diz que "a visão artificial para inspeção e orientação de robôs é madura e de grau industrial", enquanto "a IA física de uso geral e a aprendizagem a partir de simulação e experiência ainda não são de grau industrial" (sumário executivo do IFR World Robotics 2026) single source, primário. A eletrónica é o maior comprador de robôs (142,235 unidades instaladas em todo o mundo em 2025, 59% de todas as instalações na China (IFR, comunicado IFR China)), mas a China tem apenas 166 robôs por 10,000 trabalhadores industriais contra 307 nos EUA (IFR, abril de 2026). O padrão em produção é perceção aprendida mais movimento determinista; o movimento aprendido de ponta a ponta está perto do grau de produção apenas em estações estruturadas (capítulos 3 e 6).

O que cada região pode usar

A China continental depende de modelos nacionais e pesos locais; em Taiwan, a identidade do cliente decide; nos EUA, os direitos aduaneiros e a certificação.

Elemento	China continental	Taiwan (fábricas privadas)	Sudeste Asiático	Estados Unidos
Claude (qualquer canal)	● Não disponível	● Condicional: o fornecedor não pode ser detido ou controlado maioritariamente por pessoas ou entidades da China continental, Hong Kong ou Macau	● Idem	● Idem
Nuvem GPT e Gemini	● Não disponível	● Disponível	● Disponível (listas não verificadas país a país)	● Disponível
Modelos de nuvem chineses (Qwen, DeepSeek, Doubao, GLM, Kimi)	● Disponível (modelos registados; oferecer um serviço exige registo)	● Legal, mas o alojamento no continente raramente passa a análise de segurança do cliente; proibido no setor público, nas empresas estatais e nas infraestruturas críticas	● Legal (nós internacionais)	● Legal, análise rigorosa

Elemento	China continental	Taiwan (fábricas privadas)	Sudeste Asiático	Estados Unidos
Pesos abertos locais de origem chinesa (Qwen, GLM, DeepSeek)	● Primeira escolha; a implantação privada na fábrica não exige registo de modelo ou algoritmo (leitura de escritório de advogados)	● Disponível (setor público restrito)	● Disponível	● Condicional: análise de origem
Pesos abertos locais de origem não chinesa (ex.: Gemma 4)	● Disponível	● Disponível	● Disponível	● Disponível
Modelos de visão clássicos não generativos	● Disponível (fora das regras de IA generativa)	● Disponível	● Disponível	● Disponível
Jetson Thor/T5000/T4000, DGX Spark	● Disponível: a NVIDIA classifica-os 5A992.c, sem necessidade de licença dos EUA para uma fábrica civil comum; o Spark tem um SKU China; preços de canal cerca de 40–60% acima da tabela dos EUA	● Disponível	● Disponível	● Disponível
RTX 5090, RTX PRO 6000, DGX Station	● Não disponível: os dois primeiros são 4A090.a, o DGX Station 5A992.z; alternativas legais são 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	● Disponível	● Disponível, mas compradores com sede na China ou detidos, em última instância, pela China contam como chineses; a Malásia exige licenças para exportação e transbordo	● Disponível
Mac Studio M5 Ultra 512 GB	● Disponível (loja Apple China)	● Disponível	● Presumivelmente disponível	● Disponível
Huawei Ascend e outros aceleradores chineses	● Disponível	● Risco elevado	● Risco elevado	● Não utilizável na prática
Braços fixos, AOIs e câmaras industriais fabricados no continente	● Disponível	● Importável; desde 1 de dezembro de 2025 as autoridades podem agir diretamente sobre marcas do continente, e agências, empresas estatais, certas fundações e fornecedores de infraestruturas críticas podem ser restringidos ou proibidos de as usar	● Disponível	● Legal, mas o direito combinado é de cerca de 40% sobre braços, 37.5% sobre instrumentos do tipo AOI e câmaras, 25% sobre computadores de borda
AMRs e manipuladores móveis fabricados no estrangeiro	● Disponível	● Disponível	● Disponível	● Novos modelos precisam de aprovação condicional do Departamento de Guerra dos EUA
Câmaras Hikvision e Dahua	● Disponível	● Uso privado permitido, setor público proibido	● Disponível	● Novas autorizações FCC bloqueadas; proibidas para contratantes federais

Elemento	China continental	Taiwan (fábricas privadas)	Sudeste Asiático	Estados Unidos
Segurança e certificação de robôs	● GB 11291.1-2011 e GB 11291.2-2013 obrigatórias, em revisão desde abril de 2026; certificação CR voluntária; GB/T 39404-2026 cobre a cibersegurança do controlador	● Vedações por defeito; a colaboração sem vedações tem de cumprir a CNS 14490 ou a ISO 10218 com avaliação escrita	● Por cliente e destino de exportação, confirmado por escrito antes do envio	● O equipamento elétrico tem de ser "aprovado" (na prática uma marca NRTL ou avaliação de campo); a UL 1740 exige que as células de robô cumpram a ISO 10218-2 desde 11 de agosto de 2025
Fabs e fábricas de produtos controlados	—	● Os contratos de compra da TSMC exigem SEMI E187; equipamento do continente a ser removido das fabs de 2 nm	—	● Os dados técnicos ITAR incluem "fotografias"; política de negação para a China; os fornecedores controlados pelo continente não podem, na prática, servir fábricas de eletrónica de defesa

Fontes da tabela: (regiões suportadas da Anthropic, países suportados da OpenAI, disponibilidade do Gemini, medidas provisórias da CAC sobre IA generativa, dados de classificação de exportação da NVIDIA, eCFR 15 CFR 774, NVIDIA China DGX Spark, Apple China, Covington sobre as novas regras BIS, Greenberg Traurig, junho de 2026, Fulcrum, Tom's Hardware sobre orientação Ascend, lista de importação BOFT de Taiwan, histórico da Lei de Gestão da Cibersegurança de Taiwan, regras de análise de Taiwan para produtos que ponham em risco a cibersegurança nacional, Taipei Times, USITC HTS capítulo 99, Latham & Watkins, IPVM, plataforma de normas nacionais da China, GB 11291.2, atualização Intertek UL 1740, eCFR 22 CFR 126.1).

China continental: uma pilha completa, mas nem os modelos de fronteira fechados nem as máquinas individuais mais fortes podem entrar. Os modelos de nuvem nacionais registados, os pesos abertos locais (Qwen sob Apache-2.0, GLM e DeepSeek sob MIT), os modelos de visão clássicos e os AOIs, câmaras e cobots nacionais são todos utilizáveis. As próprias classificações de exportação da NVIDIA listam o kit Jetson AGX Thor, os módulos T5000/T4000 e o DGX Spark como 5A992.c, pelo que uma EMS ou fábrica de PCBA comum não precisa de licença dos EUA, enquanto o RTX 5090 e o RTX PRO 6000 (4A090.a) e o DGX Station (5A992.z, TPP 58,484) não podem ser comprados no continente (classificações de exportação da NVIDIA, eCFR) [primário; a leitura legal é minha e não é aconselhamento jurídico]. A RS China lista o kit Thor a RMB 59,104 com imposto, cerca de 1.4 vezes o preço de tabela dos EUA (RS China).

Quatro escritórios de advogados concordam que uma implantação privada na fábrica, para uso próprio da fábrica, não exige registo de modelo ou algoritmo, enquanto um serviço alojado multi-inquilino fica numa zona cinzenta (Covington, FPF, JunHe) verified: leituras consistentes, sem decisão oficial. A GB 11291.1-2011 e a GB 11291.2-2013 continuam a ser a base obrigatória de segurança de robôs e estão em revisão desde 28 de abril de 2026; a GB/T 39404-2026 sobre cibersegurança do controlador entrou em vigor a 1 de novembro de 2026 (plano de revisão 20262542, GB/T 39404-2026) single source, primário. Até a nova GB 11291 chegar (2028–2029 estimate), projete para a ISO 10218:2025 e mantenha os ficheiros de conformidade com a GB 11291.

Taiwan: as fábricas privadas podem usar quase tudo; a identidade do cliente decide. As APIs de fronteira, os pesos abertos e o equipamento do continente são todos legais (data.gov.tw). Mas, desde que a Lei de Gestão da Cibersegurança alterada entrou em vigor a **1 de dezembro de 2025**, as autoridades podem agir diretamente sobre produtos que sejam "marcas do continente" e restringi-los em órgãos governamentais, empresas estatais e fornecedores de infraestruturas críticas (base de dados de Leis e Regulamentos, regras de análise) verified; a lista de fornecedores de infraestruturas críticas é confidencial; e a TSMC escreveu a SEMI E187 nos seus contratos de compra (CNA) e, segundo notícias, está a remover equipamento do continente das suas fabs de 2 nm (TrendForce) single source. As EMS privadas, a montagem e as pequenas fábricas de PCBA estão abertas; as fabs, os serviços de parques e os clientes ligados ao governo estão fechados na prática. Um fornecedor controlado pelo continente deve projetar para a E187, entregar uma lista de materiais de software, oferecer atualizações offline e nunca ligar-se remotamente a partir do continente.

Sudeste Asiático: as menores restrições; os problemas são a amortização lenta e as casas-mãe chinesas. Tudo é utilizável, mas a orientação dos EUA de 31 de maio de 2026 exige uma licença para exportar artigos de computação avançada a entidades com sede na China, ou detidas, em última instância, pela China, mesmo quando o destinatário está num terceiro país (Greenberg Traurig) single source, e a Malásia tem as suas próprias licenças para chips avançados (Fulcrum). A procura está a mover-se para aqui: a TrendForce esperava que a Tailândia, o Vietname e a Malásia se aproximassem dos 50% da capacidade SMT de ODM de servidores em 2026 (TrendForce) single source, uma previsão de 2023.

Estados Unidos: a pilha completa, ao preço de direitos aduaneiros, certificação e dados controlados. Sob o capítulo 99 da pauta aduaneira dos EUA, um braço robótico fabricado no continente paga cerca de 40% de direito combinado, os instrumentos do tipo AOI e as câmaras cerca de 37.5% e os computadores de borda 25%; fabricado na Malásia ou na Indonésia cerca de 10–12.5%, no Vietname, Tailândia ou Filipinas 12.5–15%, em Taiwan 10%, com os computadores a 0% fora do continente (USITC HTS capítulo 99, Great Lakes Customs Law) pauta primária + estimate. O aviso de 10 de novembro da USTR prorrogou 178 antigas exclusões da Secção 301 até 9 de janeiro de 2027, nenhuma delas cobrindo robôs, instrumentos de inspeção, câmaras ou computadores (Federal Register 2026-22925) verified.

A OSHA exige que o equipamento elétrico do local de trabalho seja "aprovado" (na prática uma marca NRTL, ou uma avaliação de campo de cerca de $3,000 a $20,000+ por célula) e, desde 11 de agosto de 2025, a UL 1740 coloca células de robô inteiras sob a ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). A eletrónica de defesa está fechada na prática aos fornecedores controlados pelo continente: o ITAR trata as fotografias de produção como dados técnicos e aplica uma política de negação à China (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Uma célula com destino aos EUA deve usar um braço fixo, câmaras sem restrições e modelos locais ou de nuvem dos EUA, com montagem final no Sudeste Asiático quando isso altere genuinamente a origem.

A qualquer orçamento, o teto difere por região, mas em nenhum lado chega a zero pessoas estimate. O continente pode empilhar SPI 3D, AOI 3D, raio-X, revisão com IA e uma equipa interna de IA nacionais para reduzir a mão de obra de revisão a 8–15%, mas, sem modelos de fronteira fechados ou DGX Station, o Q&A por LLM à escala da fábrica tem de correr sobre pesos abertos locais e computação nacional (um modelo de tamanho real para muitos utilizadores precisa de, pelo menos, dois servidores Ascend de oito placas). Taiwan e o Sudeste Asiático podem tecnicamente usar o melhor de tudo, incluindo APIs de fronteira e DGX Station, limitados pela identidade do cliente, pelas marcas do continente e pelas licenças de casas-mãe chinesas. Os EUA não têm limites técnicos, ao preço de direitos aduaneiros, certificação NRTL e separação de pessoal para dados controlados. Em todo o lado, as pessoas permanecem para os casos-limite, novos produtos e trocas, novos tipos de defeito e assinatura do cliente.

Um produto vendido entre regiões precisa, portanto, de duas pilhas de modelos: uma pilha do continente sobre modelos nacionais registados ou pesos abertos locais, e uma pilha do estrangeiro sobre APIs de fronteira mais um modelo local de origem não chinesa. A política da Anthropic em vigor a 12 de novembro de 2026 julga por propriedade e controlo, pelo que um fornecedor controlado pelo continente não pode construir sobre o Claude nem através de uma subsidiária no estrangeiro, de uma plataforma de nuvem ou de um revendedor (Anthropic) single source, primário.

No local

02 Executar a IA dentro da fábrica: salas limpas, isolamento de rede, hardware e quantização

Numa fábrica isolada da rede, a restrição é o processo, não a computação

A inspeção por visão totalmente offline é rotineira; as restrições são de procedimento. O guia de 2024 do MIIT da China sobre segurança de sistemas de controlo industrial pede zoneamento, firewalls ou díodos de dados, manutenção remota auditada e limitada no tempo, suportes digitalizados e patches testados (texto do guia) single source; os clientes de semicondutores acrescentam a SEMI E187 e a E188, que a TSMC escreveu nos seus contratos de compra em 2023 (SEMI, TXOne) verified. Os resultados voltam para o MES através de IPC-CFX e a identidade da placa viaja por IPC-HERMES-9852 (norma Hermes). A armadilha comum é o licenciamento que assume uma rede: o modo offline da Roboflow guarda os pesos em cache numa licença de 30 dias (Roboflow). Um produto offline precisa de ativação offline, atualizações assinadas em suportes digitalizados, uma lista de materiais de software, controlo USB, registos de auditoria locais e um adaptador CFX.

FIG 05 **Uma implantação isolada (air gap)** ***Uma caixinha na linha decide; a sala de máquinas responde perguntas; as falhas recaem nas pessoas e os dados nunca saem da fábrica.***



Fonte: Síntese do autor (capítulo 2).

Os modelos de fronteira fechados não entram, na prática, na sala de servidores de uma fábrica. O Gemini só é oferecido isolado da rede através do Google Distributed Cloud para governos e grandes empresas (Google Cloud); o Foundry Local da Microsoft abre modelos grandes apenas a "clientes qualificados" e não nomeia nenhum modelo GPT (Microsoft); o Claude só é oferecido em regiões de nuvem governamentais, sem implantação no local (Claude for government); nenhum dos três é oferecido na China continental. O equipamento da fábrica tem de ser projetado em torno de pesos abertos; os modelos de fronteira podem, no máximo, etiquetar ou explicar imagens na sede, offline, quando o cliente permite que imagens anonimizadas saiam.

O hardware da sala de servidores deve ser tratado como de grau de escritório: o Mac Studio está classificado para 10–35 °C e o DGX Spark para 5–30 °C, com relatos de encerramentos térmicos sob inferência contínua (Apple, fó-

rum NVIDIA) verified. A linha recebe caixas industriais sem ventoinha, como um PC industrial Jetson T4000 classificado para −25 a 60 °C (AAEON) vendor claim. Como uma hora de paragem custa a uma grande fábrica $36,000 a $2.3 milhões (Siemens True Cost of Downtime 2024) [inquérito de fornecedor], cada passo de IA na linha tem de **falhar aberto**: em qualquer falha, recua para o processo existente de AOI mais pessoas.

Zonas limpas e ESD: uma sala SMT comum é uma área protegida contra ESD, não uma sala limpa classificada

O que limita onde a computação pode ficar é a estática e o fluxo de ar, não a classe de limpeza. A maioria das áreas de montagem SMT "não é formalmente controlada em partículas", e as fábricas de classe mundial "podem atingir ou aproximar-se da ISO 8" (painel de especialistas da Circuitnet) single source. A PCBA médica e os módulos de câmara ficam normalmente em ISO 7, o die nu e o alinhamento ótico em ISO 5–6 local (PCBSync) single source; os construtores de salas limpas citam ISO 7–8 para SMT em geral (Angstrom Technology) vendor claim.

O requisito universal é o controlo de ESD: a ANSI/ESD S20.20-2021 exige condutores isolados abaixo de 35 V e mantém os isoladores de campo forte a pelo menos 30 cm dos dispositivos sensíveis (ANSI, ESDA, Desco) verified. As fabs colocam a computação na sub-fab ou no centro de dados, em racks de refrigeração líquida que cumprem a SEMI S2/S8/F47 (Prodrive) vendor claim.

A entrada de equipamento numa área limpa é testada quanto à emissão de partículas sob a ISO 14644-14; os robôs têm esses certificados (Fraunhofer IPA), mas não encontrei nenhum PC industrial, caixa de IA de borda ou servidor GPU com um. A CE, a FCC e a IEC 60068 mostram que uma caixa tolera o seu ambiente, não que pode entrar numa área protegida contra ESD; o coordenador de ESD do cliente verifica isso. Os servidores GPU nunca vão para a linha: um DGX Station consumiu 1,292 W na GPU em teste (StorageReview).

FIG 06 **As zonas limpas e ESD decidem onde o computador pode ficar** ***Uma nave SMT comum é uma área protegida ESD, não uma sala limpa. Servidores GPU e NPU ficam na sala de máquinas.***



Fonte: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, normas SEMI e material de fornecedores; compilado pelo autor.

Zona	Ambiente típico	Computação na linha	Servidores GPU/NPU
Sala SMT/PCBA comum (consumo, industrial)	Área protegida contra ESD sob a S20.20 ou IEC 61340-5-1, normalmente não uma sala limpa classificada (≈equivalente ISO 8); 40–60% HR	Caixa sem ventoinha numa caixa metálica ligada à terra (Jetson T4000/Thor, caixa NPU chinesa) junto ao AOI ou ao posto de revisão; monitores e caixas de plástico fora da área ou a ≥30 cm das placas	Sala de servidores ou rack de TI com ar condicionado

Zona	Ambiente típico	Computação na linha	Servidores GPU/NPU
Médica, módulos de câmara, eletrónica EV (sala limpa ISO 7–8)	Sala limpa mais proteção ESD	Sem ventoinha, selada IP65/66, lisa e fácil de limpar; pedir dados ISO 14644-14 ou um certificado IPA	Fora da sala limpa, ligada em rede
Die nu, ligação por fio, estações óticas (ISO 5–6 local)	Miniambiente	Apenas câmaras e luzes no interior; caixa de computação fora	Sala de servidores
Sala limpa de fab (ISO 5 e mais limpa)	Planos de ESD e atração eletrostática específicos de fab, equipamento para SEMI E78; cibersegurança para SEMI E187/E188	Apenas controladores integrados na ferramenta e qualificados para fab; sem PCs complementares com ventoinhas	Racks de refrigeração líquida na sub-fab ou centro de dados da fab

A tabela é a minha síntese das fontes acima estimate. Para um produto, isso significa uma caixa de linha sem ventoinha com menos de 100 W, uma caixa metálica ligada à terra de proteção, entrada de 12–24 V CC, funcionamento em ampla gama de temperatura com o fluxo de ar real do armário e um relatório de emissão de partículas ISO 14644-14 pronto para clientes em ISO 7 e mais limpa. Qualquer projeto em que “um servidor GPU serve quatro linhas” coloca o servidor na sala de servidores e deixa apenas uma caixa pequena ou uma câmara na linha.

Limites de velocidade por uso: a precisão é mais difícil do que a velocidade

As fábricas e os fornecedores quase nunca publicam metas de latência ou disponibilidade da IA. Os limiares abaixo derivam do takt, do tempo de revisão humana, das taxas de controlo de robôs e da investigação sobre interação estimate:

Uso	Limite de velocidade	Limite de precisão	Disponibilidade
Inspeção inline de cada peça	Dentro de ½–1 takt: 240–1,000 recortes num takt de 10–30 s, 8–100 recortes por segundo	Fugas não piores do que o “AOI mais pessoas” de hoje (exemplo contratual <0.5%)	Falhar aberto, nunca parar a linha; meta ≥99.9% (sem SLA público)
Segunda opinião inline apenas sobre recortes sinalizados pelo AOI	Sinalizações (normalmente ≤5–20 por placa) dentro de ½ takt: 0.5–3 s por recorte	Idem; treinado nas próprias imagens da linha	Recua para a revisão humana
Revisão com IA dos falsos alarmes	2–5 s por chamada, mais de dez vezes mais rápido do que os 42–55 s de uma pessoa	Fugas não piores do que um revisor humano	A degradação acrescenta pessoas, nunca para a linha
Controlo de robô (VLA)	100–300 ms por bloco de ação com fragmentação em tempo real, saída de 10–50 Hz	≥99% de sucesso de tarefa para superar o movimento programado	Sistema de segurança independente do modelo
Planeamento de robô e escolha de competências	1–3 s quando necessário em cada ciclo; ≤1 min se apenas em exceções ou trocas	A saída tem de ser código verificável ou um ID de competência	Recua para o programa fixo
Q&A (SOPs, consultas MES)	Primeiro token idealmente ≤1–2 s, no máximo 10 s; ≥15–20 tokens/s por utilizador; 5–20 utilizadores concorrentes	Citações precisas; fiel em contexto longo	Grau de TI de escritório (≈99–99.5%)
Análise e relatórios de causa raiz	Minutos são aceitáveis (lote)	Máxima; fiel em contexto longo	Grau de TI de escritório

A Physical Intelligence mediu 97 ms de inferência para um VLA de classe 3B (cerca de 108–139 ms com rede e pré-processamento), concluiu que a fragmentação em tempo real mantém o débito inalterado até 200 ms de atraso extra e completou tarefas finas como acender um fósforo ou ligar um cabo de rede em mais de 300 ms (Physical Intelligence) [investigação de fornecedor]; a investigação sobre interação coloca o limite para manter uma linha de pensamento em um segundo e para a atenção em dez (NN/g). A precisão é a parte difícil: o Qwen3-VL-8B aberto obtém em média apenas 56.8 zero-shot em cinco conjuntos de dados de anomalias industriais (IndusAgent) single source. Nenhum hardware resgata um modelo de visão-linguagem que não tenha sido afinado nas próprias imagens da linha.

FIG 07 **Limites de velocidade por uso** ***A inspeção em linha é medida em milissegundos, o Q&A em segundos, a análise em minutos. As pessoas levam 42–55 s por chamada.***



Fonte: Síntese do autor (capítulos 1–2).

Hardware: Thor para a estação, um Mac Studio ou aparelho nacional para o Q&A da fábrica, e sem DGX Station no continente

Dispositivo	Memória e largura de banda	Preço (novembro de 2026)	Para que serve	China continental
Kit de desenvolvimento Jetson AGX Thor / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 no lançamento → $5,499 (julho de 2026); T5000 $4,999, T4000 $2,999; kit de desenvolvimento RS China RMB 59,104 com imposto	Modelos de visão 2–8B a reverificar recortes; VLAs de classe 3B em tempo real	● Sim (5A992.c, sem licença para fábricas civis)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; o ASUS GX10 China do mesmo chip RMB 32,999 no JD (out. 2025)	MoE 8–35B multiusuário	● Sim (5A992.c, modelo China, canais autorizados)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (jun. 2026) → ≈$16,005 listagem média no eBay em novembro	27–35B FP8 multiusuário; MoE de classe 120B a 4 bits	● Não (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB unificada, 1.2 TB/s; 480 W máx. contínuos	Desde $22,299 nos EUA; ≈RMB 173–186k estimados na China	DeepSeek 671B a 4 bits para 1–3 utilizadores	● Sim

Dispositivo	Memória e largura de banda	Preço (novembro de 2026)	Para que serve	China continental
DGX Station GB300	748 GB (252 GB HBM3e a 7.1 TB/s + 496 GB LPDDR5X a 396 GB/s); 1,292 W medidos na GPU	$123–140k preço de rua em novembro (≈RMB 830–940k antes de direitos); $95–117k em agosto	Alta concorrência em MoE 120–235B que cabe na HBM; apenas alguns utilizadores em modelos 433–743B	● Não (5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	RMB 13,300 de um comerciante JD	Revisão em lote de baixo volume; poucos modelos suportados	● Sim (nacional)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Não obtido; DeepSeek completo precisa de dois em W8A8, quatro em BF16	Q&A e análise multiusuário à escala da fábrica	● Sim (nacional)
RTX 5090 D V2 (modelo China)	24 GB	RMB 16,499 no lançamento (ago. 2025), a subir desde então com os preços da memória	Modelos de visão 8B, VLAs pequenos	● Sim

Fontes: (NVIDIA China Thor, lançamento NVIDIA Thor, CNX sobre subidas de preço, dados de classificação de exportação da NVIDIA, RS China, Implicator, IThome sobre ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, especificações Apple, TechRadar, IThome sobre preços M5 Ultra na China, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, matriz de suporte vLLM-Ascend, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). Os preços do Mac Studio de 512 GB no continente são extrapolados dos prémios de configuração dos EUA e são de baixa confiança estima-te .

FIG 08 **Computação local: memória, preço e o que se pode comprar legalmente na China continental** ***As classificações de exportação decidem a lista curta antes dos benchmarks.***



Fonte: Classificações de exportação da NVIDIA, Apple, RS China, JD.com e preços de revisão (novembro de 2026).

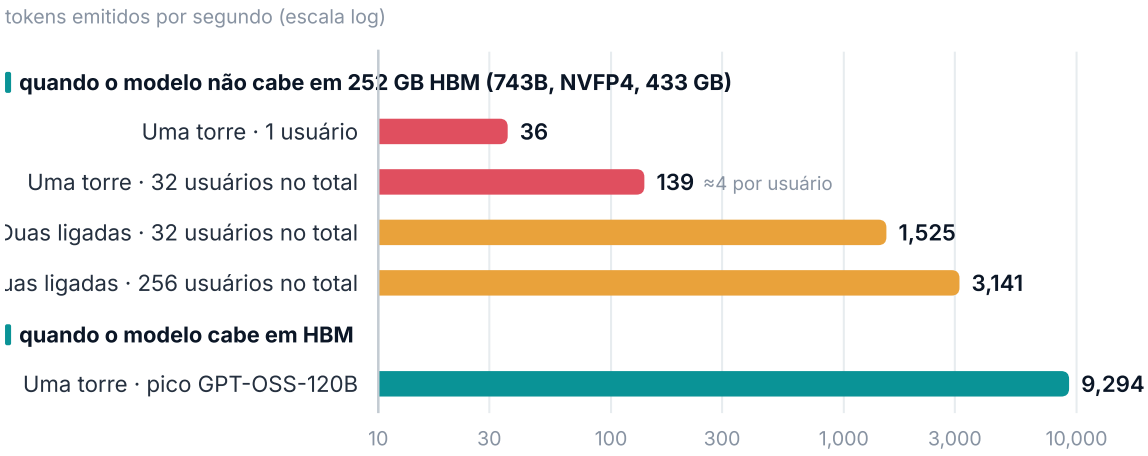
O Jetson Thor torna a implantação ao nível da estação viável. Os testes TensorRT Edge-LLM da NVIDIA mostram o Qwen3-VL-8B (NVFP4) a pré-encher 2,048 tokens em 104.5 ms e a descodificar 44.4 tokens/s no Thor (TensorRT Edge-LLM) vendor claim ; para um recorte de AOI e uma resposta de 20 tokens, isso é cerca de 0.5 s com o modelo 8B e 0.2 s com o 2B, ou 8.5–21 recortes por segundo em lote estimate . O comprimento da saída, não a contagem de parâmetros, define a latência, pelo que a descodificação restrita a um token de classe ou JSON curto faz parte do projeto. O GR00T N1.7 (3B) corre a 25.1 Hz no Thor (Jetson AI Lab) vendor claim . Um Orin NX demora cerca

de 1.8 s por recorte e um Atlas 300I Duo tem um p95 acima de 23 s, servindo apenas para trabalho em lote (CSDN) single source. No continente, orçamente os produtos NVIDIA a 40–60% acima do preço de tabela dos EUA estimate.

Um Mac Studio de 512 GB torna possível um modelo de tamanho real numa só máquina, para poucos utilizadores. O M5 Ultra (agosto de 2026) oferece até 512 GB a 1.2 TB/s desde cerca de $22,299 (Apple, TechRadar); o DeepSeek 671B a 4 bits (≈380–404 GB) descodifica 17–20 tokens/s num M3 Ultra (Pinggy) verified e um estimado de 25–29 no M5 Ultra. Os prompts longos são o ponto fraco: um prompt de recuperação de 10,000 tokens demora cerca de 17–70 s estimate, além do limite de interação de 10 segundos.

O DGX Station é a máquina individual mais forte, mas não está disponível no continente e custa cinco a seis vezes o preço de um Mac. O GB300 em distribuição tem **748 GB**, dos quais apenas 252 GB são HBM rápida (Wccftech, Tom's Hardware, ago. 2026) verified. Testes independentes mostram cerca de 9,300 tokens/s de saída no GPT-OSS-120B, que cabe na HBM, mas um modelo 743B que transborda para LPDDR5X deu a 32 utilizadores apenas 139 tokens/s no total, cerca de 4 cada; duas unidades ligadas chegaram a 1,525 (StorageReview, ago. 2026, StorageReview, out. 2026) single source. Os preços de rua subiram para $123–140k em novembro (CyberPowerPC, monitors.com). No continente, as opções realistas para o mesmo trabalho são o Mac Studio, servidores CPU+GPU estilo KTransformers (KTransformers) [teste de fornecedor] ou servidores Ascend nacionais.

FIG 09 **Uma grande estação de trabalho não é um assistente compartilhado** ***Quando o modelo não cabe na memória de alta largura de banda, 32 usuários simultâneos obtêm cerca de 4 tokens por segundo cada um.***



Fonte: StorageReview (ago. e out. 2026); ServeTheHome. Testes independentes de fonte única.

A escassez de memória moveu as mesmas peças 20–100% num ano (o DGX Spark passou de $3,999 para $6,950 (TechRadar)), pelo que as cotações de equipamento devem ser indexadas em vez de fixas.

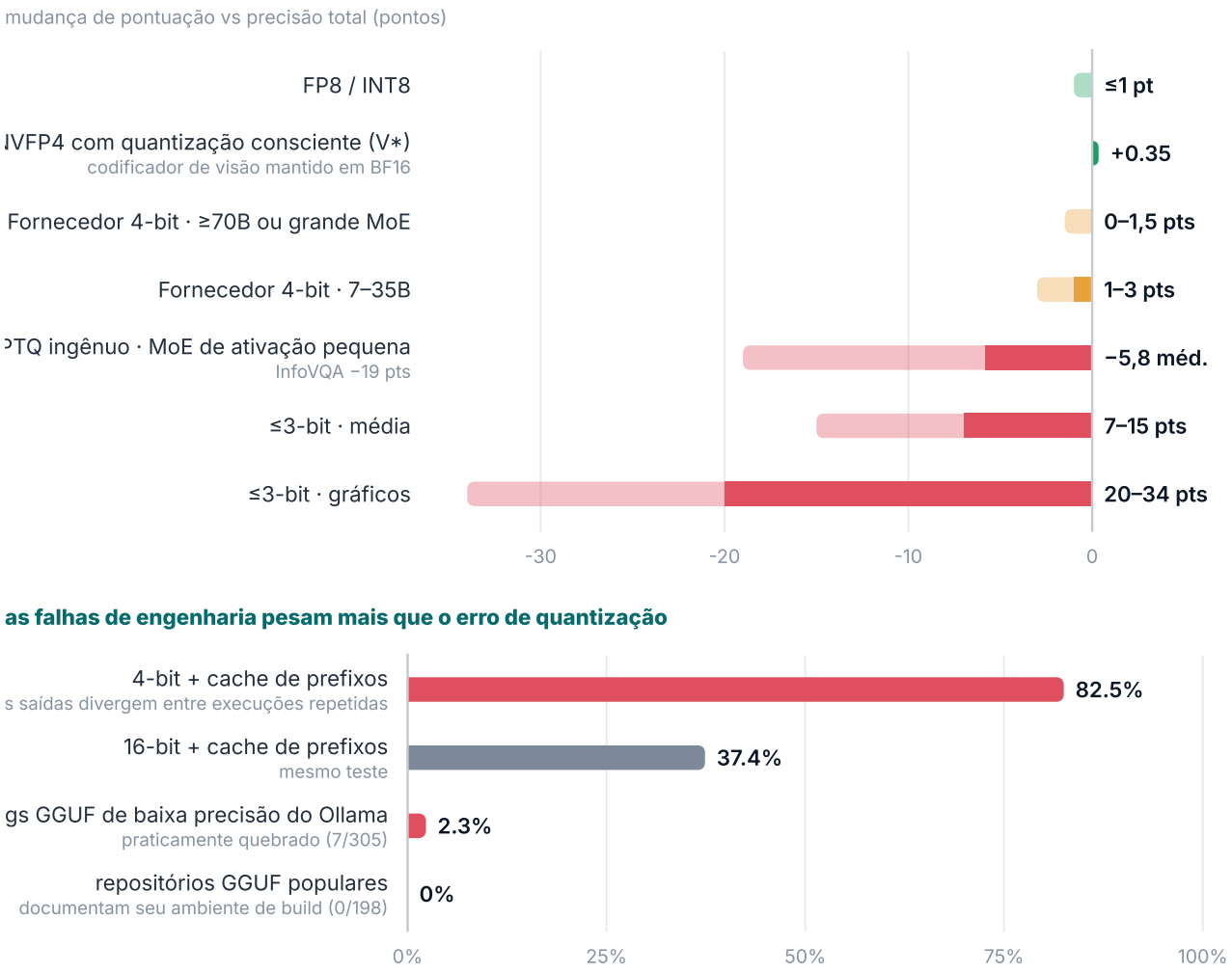
Os aceleradores nacionais executam modelos de tamanho real; a manutenção do software é o gargalo. O Ascend 910B tem 64 GB de HBM, mas o vLLM-Ascend marca o DeepSeek V3/R1 como "já não mantido" e suporta apenas alguns modelos em placas mais pequenas (vLLM-Ascend) single source, primário, e cada novo modelo demora um a dois meses a adaptar (Caijing/Anue); envie uma versão de modelo congelada e validada. Servir o DeepSeek 671B completo a 20–100 utilizadores precisa de, pelo menos, dois servidores Atlas 800I A2 de oito placas (China Telecom Cloud) verified, mas os concursos públicos para aparelhos LLM na China totalizaram apenas RMB 27.5 milhões entre janeiro e setembro de 2025 (C114) single source: as fábricas pagam por integração e serviço, não por uma caixa de computação.

Quantização: 8 bits é seguro, 4 bits depende da receita, 3 bits ou menos está fora

8 bits (FP8 ou INT8) é essencialmente sem perdas para modelos de visão-linguagem (Red Hat, Hugging Face) verified. A 4 bits, a perda depende do tamanho e da receita: a compilação AWQ do Qwen 72B perde menos de um

ponto no MMMU, enquanto o 7B perde 2.8 (Qwen AWQ) verified, e as tarefas com muitos detalhes caem ainda mais (o GPTQ ingénuo a 4 bits custa ao Qwen3-VL-30B-A3B 19 pontos no InfoVQA (VEQ)). A 3 bits ou menos, a perda média é de 7–15 pontos e os gráficos perdem 20–34 (VLMQ) verified; além de 64k tokens, o 4 bits perde até 59%, pior em entrada não inglesa (arXiv 2505.20276). Mantenha o codificador de visão em FP16/BF16.

FIG 10 **Quantização: 8-bit é seguro, 4-bit depende da receita, 3-bit e menos é descartado*Erros de engenharia na pilha de serviço custam mais precisão que a própria quantização.***



Fonte: Ficha do modelo NVIDIA Nemotron, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 e outros.

Dois avisos importam mais. Primeiro, **até 11 de novembro de 2026, nenhum estudo ou cartão de modelo mediu a quantização em imagens de defeitos de PCB/SMT**, pelo que revalide no conjunto dourado. Segundo, os erros da pilha de serviço são muitas vezes maiores do que o erro de quantização: o mesmo modelo pontuou cerca de 30 pontos abaixo sob Ollama do que sob a implementação de referência (SiliconBench); 7 das 305 etiquetas oficiais de modelos de código do Ollama estavam essencialmente quebradas embora os ficheiros estivessem intactos (arXiv 2609.05881); nenhum dos 198 repositórios GGUF populares registou a sua compilação (arXiv 2610.19498); e, com reutilização de cache de prefixo, 82.5% dos episódios divergiram a 4 bits contra 37.4% a 16 bits (arXiv 2609.04748) single sources. O 4 bits com quantização consciente do fornecedor está perto de sem perdas em tarefas de detalhe (NVIDIA Nemotron 3.5) vendor claim, mas os VLAs a 4 bits perdem 1.5–4.3 pontos sob perturbação e correm mais devagar no Thor (LoopQ) single source. Gira um modelo quantizado como um novo calibre: fixe os hashes dos pesos, registe a compilação, reinicie a cache de prefixo e revalide após qualquer mudança de precisão, motor ou modelo.

Precisão	≥70B denso ou ≥100B MoE	7–35B	≤4B ou tarefas com muitos detalhes	Veredito
FP8 / INT8	Seguro (≤1 ponto)	Seguro	Seguro; o padrão para inspeção	Usar
4-bit AWQ/GPTQ (receita de fornecedor, codificador de visão FP16)	0–1.5 pontos	1–3 pontos; GPTQ ingénuo em MoE de pequena ativação −5.8 em média, −19 no InfoVQA	2–3 pontos; validar primeiro as tarefas de detalhe	Condicional
NVFP4 (receita de fornecedor)	≈0.3–1 ponto abaixo do FP8 (dados de fornecedor)	Compilações conscientes de quantização perto de sem perdas em V*, OCRBench	Sem dados públicos	Condicional (preferir compilações conscientes de quantização)
GGUF Q4_K_M	Não testado	Alta variância em tarefas de domínio	Não para texto pequeno ou defeitos	Não para julgamento de detalhe
GGUF Q5_K_M e acima	Presumivelmente seguro	Perto de sem perdas em dois testes	O piso para detalhe no llama.cpp	Usar
≤3 bits (incluindo modelos 1T espremidos a 1.8–2 bits numa máquina)	Não para decisões	Não	Não	Não

Que configuração para que trabalho

Configuração	Inspeção inline	Revisão com IA dos falsos alarmes	Planeamento e controlo de robô	Q&A e análise
Detetores dedicados ou modelos de anomalia (classe YOLO, RF-DETR, PatchCore), INT8/FP16, qualquer caixa de borda NPU ou GPU	● Sim: o juiz principal, 2–12 ms por recorte	● Sim (classificador afinado na linha)	○ Apenas perceção	—
Modelos de visão 2–4B, NVFP4/AWQ/FP8 com codificador de visão FP16, em Thor/T4000 ou DGX Spark	● Condicional: reverificar recortes sinalizados pelo AOI a 0.2–0.3 s cada	● Condicional, abaixo de 1 s após afinação	● Espinha dorsal de VLA (VLA 2–3B a ≈40–130 ms por bloco)	○ Fraco
Modelos de visão 7–9B (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 ou 4 bits de fornecedor, em Thor, Spark ou uma placa de 24–32 GB	● Condicional: ≈0.5 s por recorte sinalizado	● Sim: 0.5–1 s por chamada, melhor valor	● Planeamento e escolha de competências (1–3 s), não na malha de controlo	● Q&A básico de SOP para um ou poucos utilizadores
MoE 26–35B (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● Não (demasiado pesado)	● Sim, como explicador ou para casos difíceis	● Planeamento	● Sim (8 utilizadores num Spark, ≈0.1 s primeiro token)

Configuração	Inspeção inline	Revisão com IA dos falsos alarmes	Planeamento e controlo de robô	Q&A e análise
MoE 106–241B a 4 bits num Mac Studio, RTX PRO 6000 ou Spark	● Não	● Não	● Planeamento offline	● Sim, para um ou poucos utilizadores
DeepSeek 671B ou Kimi K2.x 1T a 4 bits num Mac de 512 GB, servidor KTransformers ou DGX Station	● Não	● Não	● Apenas geração offline de programas	● Condicional: 1–3 utilizadores e prompts curtos; prompts longos demoram 16–170 s
Qualquer coisa a ≤3 bits; GGUF Q4_K_M para trabalho de detalhe com modelos pequenos; Ollama para muitos utilizadores	● Não	● Não	● Não	● Não

Fontes: os dados TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research e KTransformers acima, e (testes DGX Spark, testes RTX PRO 6000, Hardware Corner sobre Atlas, Ultralytics YOLO26, RF-DETR) **testes de comunidade e fornecedores, na maioria single source**.

Assim, às duas preocupações habituais: **não tem de ser lento** se o modelo corresponder ao trabalho (modelos de visão pequenos e médios em Thor ou Spark cumprem todos os orçamentos inline e de revisão; os modelos 671B–1T ficam fora da linha), e **a quantização não tem de o tornar errado** com FP8/INT8 por defeito, ou 4 bits de receita de fornecedor acima de 7B validados no conjunto dourado. Os riscos maiores são a precisão industrial zero-shot de cerca de 57 pontos dos modelos gerais e os erros da pilha de serviço. Dê aos pedidos de inspeção tamanhos de lote fixos e uma fila própria (a temperatura 0 não é bit-reprodutível sob lotes (documentação vLLM)), mantenha um reserva quente N+1 e faça cada passo na linha falhar aberto.

Professor e aluno

03 Professor e aluno: os modelos de fronteira podem pré-etiquetar dados de inspeção; para os braços robóticos, os verdadeiros professores continuam a ser pessoas

Candidatos a professor em novembro de 2026: a visão genérica duplicou, faltam as pontuações de defeitos industriais

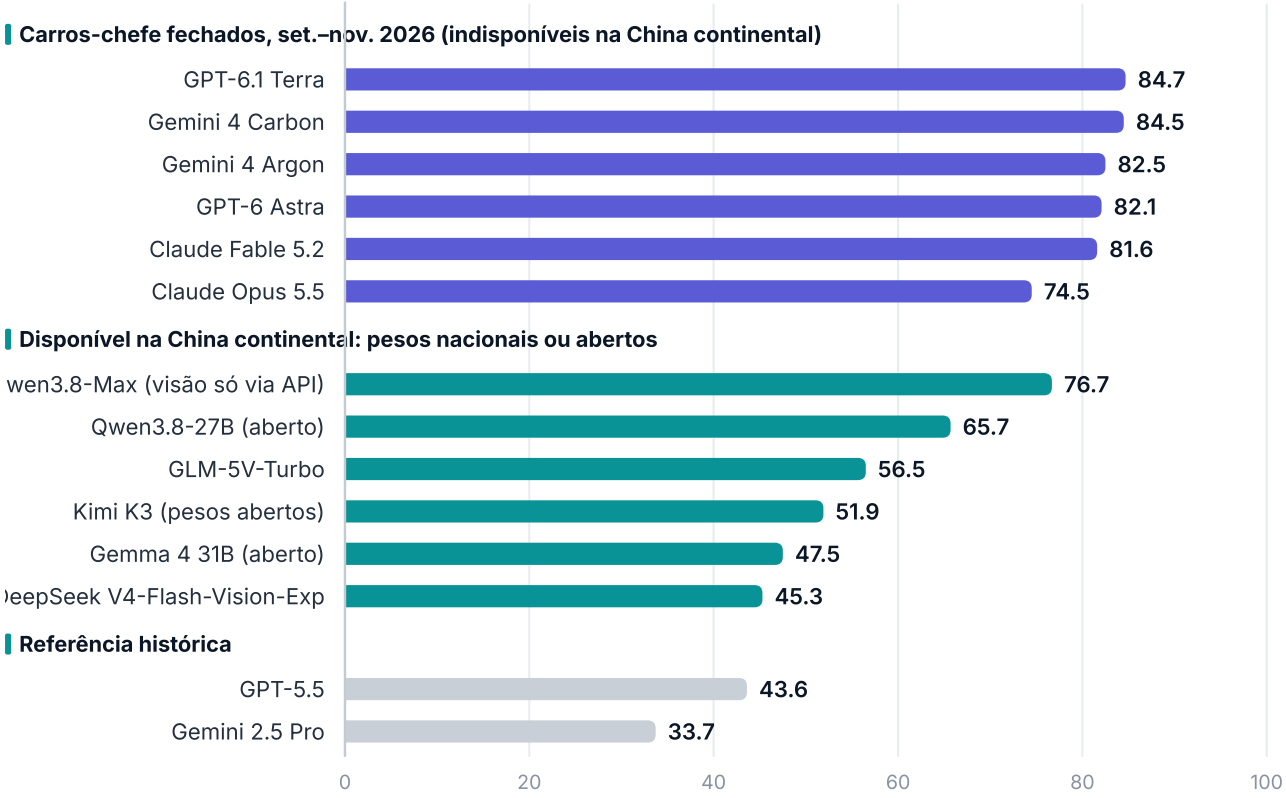
Todos os modelos de topo fechados que poderiam atuar como professor em novembro de 2026 aceitam imagens: Claude Fable 5.2 (disponibilidade geral a 9 de novembro, $8/$40 por milhão de tokens) e Opus 5.5 ($4/$20) (Anthropic, visão geral de modelos Claude) verified; GPT-6 Astra ($10/$50) e GPT-6.1 Terra ($1/$5) (preços OpenAI); e Gemini 4 Argon e Carbon, abertos por fases (Google Carbon). Utilizáveis na China continental estão o Qwen3.8-Max (visão via API) e modelos abertos como Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp e Gemma 4 (blog Roboflow, Hugging Face).

A lacuna principal tem de ser dita com clareza: nenhum destes modelos tem uma pontuação publicada em MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB ou qualquer conjunto de dados de PCB/SMT. O MMAD continua sem tabela de classificação (MMAD GitHub), e os dois artigos de anomalias industriais mais recentes (agosto–setembro de 2026) ainda usam o Gemini 2.5 Pro e o Claude Sonnet 4 como as suas bases gerais mais recentes (InspectorGPT, Anomaly-LR). Uma pesquisa de texto completo no arXiv a 11 de novembro encontrou zero resultados para "Fable 5.2" ou "Gemini 4 Carbon" neste contexto (API arXiv) single source, ausência de evidência. Os substitutos mais próximos:

- O modelo de fronteira mais recente com uma pontuação MMAD é o GPT-5.5: 69.95% em "defeito ou não" e 76.20% em média em sete tarefas num subconjunto de 500 imagens, contra 95.24% e 86.65% de peritos humanos (AnomalyClaw) single source; o GPT-4o pontuou 68.63% e 74.92% em 2024 (MMAD, ICLR 2025), cerca de 1.3 pontos de progresso em dois anos.
- No LDU-Bench, imagens reais de revisão SEM de litografia IC e o único benchmark de eletrónica com modelos de 2026, a triagem bom-ou-mau é razoável (macro F1 0.932 para o GPT-5.4, 0.884 para o GLM-5V-Turbo, 0.876 para o Claude Opus 4.6, 0.504 para o Gemma 4 31B), mas classificar 11 morfologias de defeito chega apenas a 0.20–0.41 (LDU-Bench) single source.
- Nas avaliações de visão independentes da Roboflow (atualizadas a 10 de novembro de 2026), o mAP@50 de deteção de objetos é 84.5 para o Gemini 4 Carbon, 84.7 para o GPT-6.1 Terra e 81.6 para o Fable 5.2, contra 33.7, 43.6 e 38.6 para o Gemini 2.5 Pro, GPT-5.5 e Claude Opus 4.8 (Roboflow Vision Evals) single source, avaliador independente, conjuntos de dados privados.

FIG 11 **Candidatos a professor: as pontuações genéricas dobraram em uma geração*Mas nenhum dos carros-chefe 2026 tem pontuação publicada em um banco de defeitos industriais.***

detecção genérica Roboflow mAP@50



Bancos de defeitos industriais (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): zero pontuações publicadas para esses carros-c
as pontuações de detecção genéricas dobraram em uma geração, mas isso não é julgamento de defeito; teste no seu próprio conjunto de referência

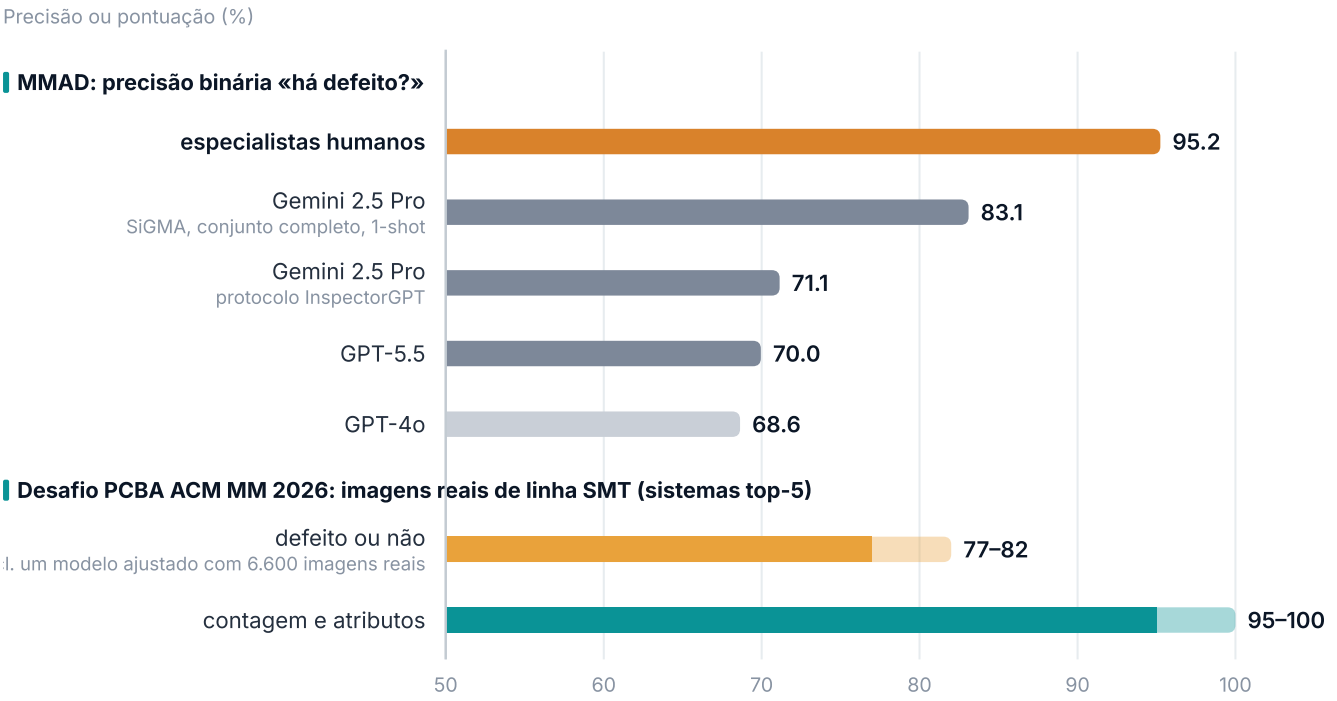
Fonte: Roboflow Vision Evals; notas de versão de modelos; páginas de MMAD, MVTec AD 2 e outros bancos (em 11 nov. 2026).

Candidato a professor (nov. 2026)	Preço por M de tokens (entrada/saída)	mAP@50 / contagem de deteção Roboflow	Benchmarks de defeitos industriais	Utilizável na China continental
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Nenhum	Não
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Nenhum	Não
GPT-6 Astra	$10 / $50	82.1 / 80.2	Nenhum	Não
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Nenhum	Não
Gemini 4 Carbon	$2 / $10 (promocional)	84.5 / 90.1	Nenhum	Não
Qwen3.8-Max (visão via API apenas)	$2 / $6	76.7 / 81.1	Nenhum	Sim
Qwen3.8-27B (aberto)	Local	65.7 / 64.9	Nenhum	Sim
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	Triagem LDU-Bench 0.884	Sim
Gemma 4 31B (aberto)	Local	47.5 / 51.4	Triagem LDU-Bench 0.504	Sim (pesos locais)

Candidato a professor (nov. 2026)	Preço por M de tokens (entrada/saída)	mAP@50 / contagem de deteção Roboflow	Benchmarks de defeitos industriais	Utilizável na China continental
Referência: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD binário 69.95% / 68.63%	● Não
Peritos humanos	—	—	MMAD binário 95.24%	—

Os artigos mais recentes ainda testam a geração anterior, mas apontam na mesma direção. No conjunto MMAD completo, os peritos humanos pontuam 95.2% em defeito-ou-não e o Gemini 2.5 Pro 83.1% (SiGMA), mas o mesmo modelo pontuou apenas 71.14% noutro artigo (InspectorGPT): o protocolo de teste move a pontuação tanto como uma geração de modelo. No desafio VAND 4.0 da CVPR 2026, os modelos de visão-linguagem prontos a usar atingiram um pico de 61.88 AP contra 90.84 dos métodos supervisionados, e os organizadores concluíram que "atualmente não podem substituir classificadores afinados" (VAND 4.0) single source. No desafio PCBA da ACM MM 2026, os cinco primeiros pontuaram 76.92–82.47 em defeito-ou-não, enquanto o tipo de componente e a contagem de pinos chegaram a 95–99.5 (arXiv 2609.21276) single source: reconhecer é muito mais fácil do que julgar.

FIG 12 **«Há defeito?» continua difícil para modelos gerais** ***O melhor modelo geral fica mais de 12 pontos atrás dos especialistas humanos no MMAD; sistemas ajustados se saem melhor em imagens reais de linha.***



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

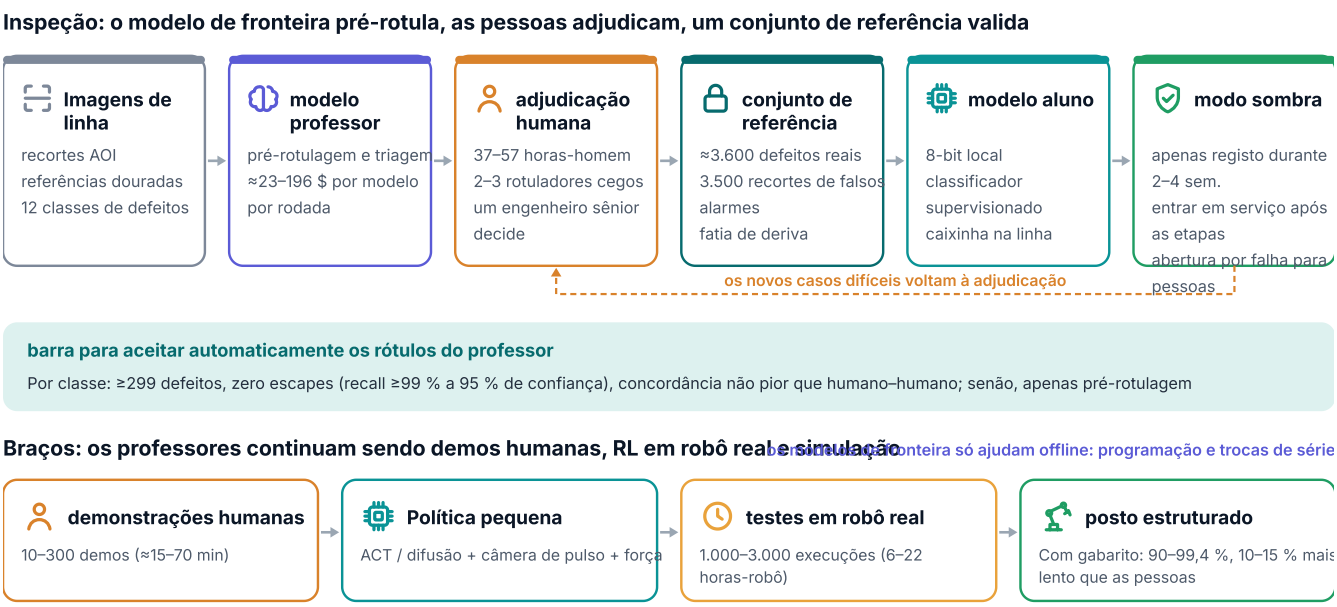
Seguem-se três conclusões. Ver e contar duplicaram numa geração, mas julgar a anormalidade avança devagar, depende fortemente do protocolo e traz enviesamentos sistemáticos; o quanto os novos modelos ganham só pode ser medido nas próprias imagens de uma fábrica. Dentro da sua distribuição de treino, os especialistas 7B afinados ainda batem os modelos gerais (Anomaly-LR, UniPCB). E os dados reais de linha fazem cair as pontuações: em painéis eletrónicos montados, o AUROC ao nível da imagem caiu de 97–98 no MVTec AD para 31.5–78.6 (arXiv 2606.01992) single source. Para as empresas do continente, o professor mais forte disponível, o Qwen3.8-Max, fica cerca de 8 pontos abaixo da fronteira global.

Executar dezenas ou centenas de rondas: o dinheiro não é o problema, a qualidade das etiquetas é

O dinheiro não é o problema. Aos preços de tabela de novembro de 2026, 1,000 imagens de um megapíxel custam cerca de $3 (GPT-6.1 Terra) a $32 (GPT-6 Astra) por ronda, e cem rondas sobre 100,000 imagens cerca de $32,000–135,000 antes de descontos por volume estimate (docs de visão Claude, guia de visão OpenAI, Roboflow Vision Evals). Para etiquetas simples, uma pessoa custa mais ou menos o mesmo que o Terra; a vantagem do professor é a velocidade e a escala, não o preço. Os NDAs dos clientes normalmente mantêm as imagens de produto dentro da fábrica, pelo que o conjunto dourado é etiquetado no local.

O obstáculo real é a qualidade, e mais rondas não corrigem erros sistemáticos. A Roboflow executa cada tarefa três vezes e as pontuações movem-se apenas ±0.1–1.4 pontos (Roboflow), pelo que um modelo repete sobretudo os seus erros e a votação por maioria não o salva. As rondas ajudam apenas quando cada uma acrescenta informação: uma ou duas boas imagens de referência acrescentam cerca de 3–4 pontos, mais não acrescentam nada (AD-Copilot); pôr um modelo a argumentar contra uma boa referência ao longo de vários turnos acrescenta 6.23 pontos de AUROC entre domínios para o GPT-5.5 e 2.95 no MMAD, atingindo o pico após cerca de três chamadas (AnomalyClaw) single sources. "Um aluno treinado em etiquetas de professor aproxima-se das etiquetas humanas" só foi demonstrado em imagens naturais: um ResNet-101 treinado em etiquetas Qwen3-VL de COCO e Objects365 atingiu 95.7–99.4% do resultado com etiquetas humanas, e os autores dizem que os dados de domínio não foram testados (arXiv 2602.20972) single source.

FIG 13 **Dois pipelines professor–aluno** ***Para inspeção, um modelo de fronteira pré-rotula e as pessoas adjudicam; para braços robóticos, os verdadeiros professores continuam sendo as demonstrações humanas.***



Fonte: Síntese do autor (capítulo 3); custos e tamanhos de amostra são estimativas.

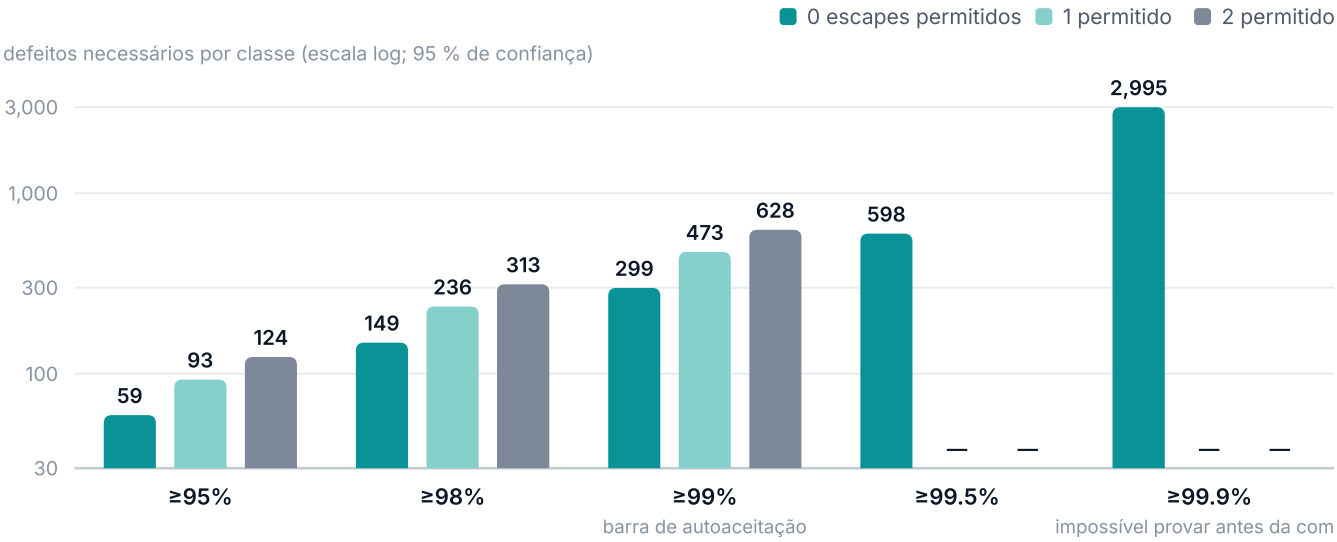
Um protocolo viável estimate:

1. **Recolher o conjunto dourado do arquivo de AOI da própria fábrica e das placas NG retidas**, estratificado em 12 classes de defeito verdadeiras (em falta, deslocada, lápide, polaridade, peça errada, ponte/curto, solda insuficiente, solda em excesso, bola de solda, terminal levantado/aberto, peça danificada, material estranho) e quatro causas de falso alarme (sombra, fluxo, serigrafia, tolerância), mais uma "fatia de deriva" de pelo menos 300 defeitos de uma segunda linha ou com iluminação, cor de placa ou números de peça diferentes.

2. **Etiquetar às cegas com dois ou três inspetores formados em IPC-A-610**, com um engenheiro sénior a decidir e raio-X, teste elétrico ou cortes transversais a resolver as amostras ambíguas. Aceitar uma definição de classe apenas se a concordância entre inspetores atingir κ ≥0.75 (≥0.90 para classes críticas); a concordância humana em 11 morfologias de defeito de eletrónica no LDU-Bench foi apenas 0.727, pelo que algumas classes precisarão de fusão ou redefinição. A concordância de atributos estilo AIAG (eficácia ≥90%, omissão ≤2%, falso alarme ≤5%, κ >0.75; >0.90 para libertação) é um mínimo sensato (Fabrico, iFactory).
3. **Dimensioná-lo com limites binomiais exatos (Clopper–Pearson) unilaterais a 95%** (intervalos de confiança binomiais, regra de três:

Sensibilidade a provar	0 omissões permitidas	1 permitida	2 permitidas
≥95%	59 defeitos	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **Qual deve ser o tamanho do conjunto de referência*Provar 99 % de recall com zero escapes exige 299 defeitos por classe; 99,9 % não pode ser provado antes da compra.***



Fonte: Intervalos de confiança binomiais exatos (95 %).

Sessenta defeitos por classe bastam para um ecrã passa/falha; 300 são necessários para decidir a aceitação automática. O conjunto completo, cerca de 3,600 defeitos mais 3,500 recortes de falso alarme, demora cerca de 37–57 horas-homem a etiquetar e arbitrar, e executar um candidato a professor sobre ele custa $23–196; o custo real são duas a seis semanas de recolha, definidas pela classe mais rara estimate. Envie recortes de AOI de componente único, não placas inteiras, com o alvo de pelo menos 42–48 píxeis de largura (arXiv 2610.09313).

1. **Aceitar automaticamente as etiquetas do professor apenas para uma classe com zero omissões em ≥299 defeitos dourados**, concordância modelo-verdade não pior do que a concordância humano-humano, e pelo menos 50% dos falsos alarmes resolvidos. Classes críticas como curto-circuitos, peças em falta, polaridade e peças erradas precisam de 99.5–99.9% (598–2,995 defeitos); caso contrário, tudo o que a IA não conseguir resolver com alta confiança vai para uma pessoa. Uma taxa de fuga abaixo de 0.1% dos defeitos não pode ser provada antes da compra (2,995 defeitos com zero omissões; 0.01% exigiria 29,956), pelo que a linha mantém

amostragem humana das chamadas resolvidas pela IA, ICT/FCT a jusante como rede de segurança e SPC sobre as fugas. As etiquetas verificadas são depois destiladas num detetor ou classificador convencional e revalidadas em cada novo produto.

Os termos e as regras de dados também decidem o professor. A Anthropic proíbe usar os seus serviços para treinar modelos concorrentes (termos comerciais Anthropic); a OpenAI permite que as saídas treinem classificadores internos que não sejam vendidos a terceiros (termos comerciais OpenAI); a DeepSeek permite explicitamente a destilação (termos DeepSeek) [termos primários]. Nenhuma define "modelo concorrente", pelo que o uso comercial precisa de revisão legal, e as regras de dados dos clientes normalmente mantêm as imagens no local. O padrão mais seguro é um professor de pesos abertos executado localmente, com modelos de fronteira de nuvem apenas quando o cliente o autoriza por escrito.

Alunos de inspeção: pequenos e rápidos; os gargalos são a ótica, os dados e a validação

O aluno não é o gargalo: o YOLO26n demora 3.80 ms por imagem em INT8 num Orin Nano Super de $249–399 (Ultralytics) [teste de fornecedor], a NVIDIA refere 98.51% em três classes de defeito de PCB após adaptar um modelo de fundação em cerca de 600 imagens (NVIDIA) vendor claim, e os defeitos sintéticos acrescentam apenas 0.9–2.5 pontos de mAP (arXiv 2605.04635). As restrições são a ótica, os dados de linha, a validação e as licenças: o YOLO da Ultralytics é AGPL-3.0 ou de licença empresarial (Ultralytics), pelo que os produtos fechados precisam de uma licença ou de um detetor Apache-2.0 como o RF-DETR.

Alunos robóticos: as estações estruturadas estão perto do grau de produção; o trabalho flexível e dinâmico não

Em novembro de 2026 há três tipos de professor de robô, e apenas um pode tornar-se um aluno local: VLAs abertos e afináveis (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Os modelos de robô fechados (π0.6/π0.7 através de uma API beta, Gemini Robotics 2 para testadores de confiança) não podem ser implantados no local (The AI Insider, DeepMind). Os modelos gerais de fronteira são lentos e fracos em manipulação fina: no teste da Artificial Analysis pontuaram 46–69 com 28–45% de sucesso de agarramento e cerca de 30 minutos por tarefa, "demasiado lentos e caros para qualquer uso de produção" (AA-Robotics); o GPT-6 Astra conseguiu uma ranhura de precisão apenas 2 vezes em 20 (RoboCurve); e acionar articulações diretamente dá 0–5.5% no LIBERO a 0.2–0.4 Hz contra os cerca de 83 Hz necessários (Anthropic) verified em três fontes. Os 64.9% do Fable 5.2 no benchmark simulado RoboDojo são uma execução interna (Anthropic) vendor claim.

Estudos de terceiros desde outubro concordam: cerca de 33–44% no LIBERO-10 para os três agentes líderes (arXiv 2610.30051); agentes a reclamar falsamente sucesso em 24–39% das subtarefas (arXiv 2611.08209); e 104 de 108 tarefas quase estáticas com robôs reais para o Opus 5.5, mas a 280 segundos cada e sem inserção ou montagem (AMP1) single sources. Os modelos de fronteira pertencem ao offline: escrever programas, orquestrar tarefas e redigir critérios de sucesso.

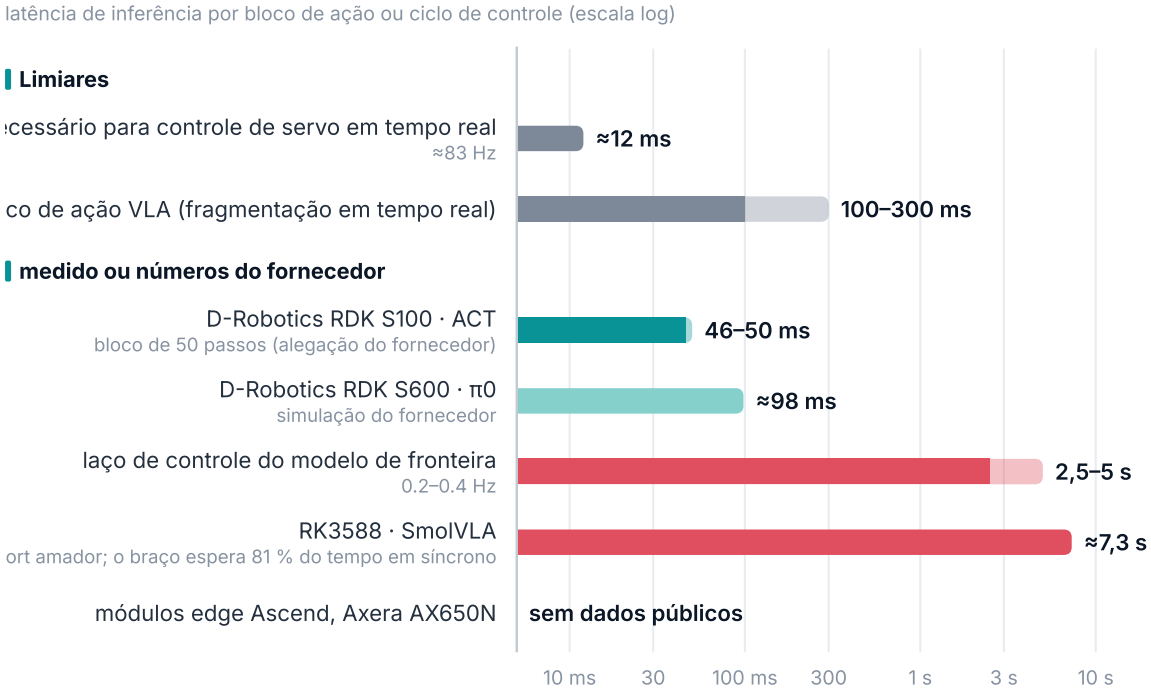
A qualidade do aluno tem uma margem clara: boa dentro da distribuição de treino, a colapsar fora dela. O π0.7 "muitas vezes excede 90%" em tarefas vistas e chega a 60–80% em nunca vistas (artigo π0.7) [investigação de fornecedor]; o ContactDP atinge 97.8% em conectores treinados mas 28.6–66.7% em nunca vistos (ContactDP); pequenas mudanças de câmara ou pose fazem cair 95% abaixo de 30% no LIBERO-Plus (LIBERO-Plus, CVPR 2026); o agarramento aprendido de transportadores em movimento atinge um pico de 58.89% (ForeTime-VLA) single sources, um trabalho que o rastreamento por visão clássica já faz em produção há anos.

A evidência perto da produção vem toda de estações estruturadas, demonstrações humanas e correção em robô real. A Neurometka manteve uma máquina de estados e movimento pré-ensinado e aprendeu apenas o agarramento, a inserção de cabos e a soldadura, cada um com menos de 20 minutos de dados: 322 de 324 em produção, 12.8% mais lenta do que uma pessoa (Neurometka) single source, fornecedor. A Fraunhofer atingiu 92.8% em quatro conectores com 68 minutos de demonstrações (Fraunhofer); o HIL-SERL de Berkeley teve sucesso 100 vezes em

100 na inserção de RAM, SSD e USB após 1–2.5 horas de RL em robô real (HIL-SERL); a AgiBot executou 18–20 segundos por passo dentro de um takt de 20 segundos numa linha de teste de tablets (IThome) vendor claim.

A destilação compra velocidade, não generalização: o π0.5 corre a 0.70 Hz no Jetson Orin em bruto e a 10.30 Hz após agendamento assíncrono (EagleVLA); a destilação de camadas leva-o de cerca de 3 Hz para cerca de 10 Hz (Shallow-π) verified. Um aluno herda apenas o que o professor cobriu; cada novo produto continua a precisar de dados de tarefa.

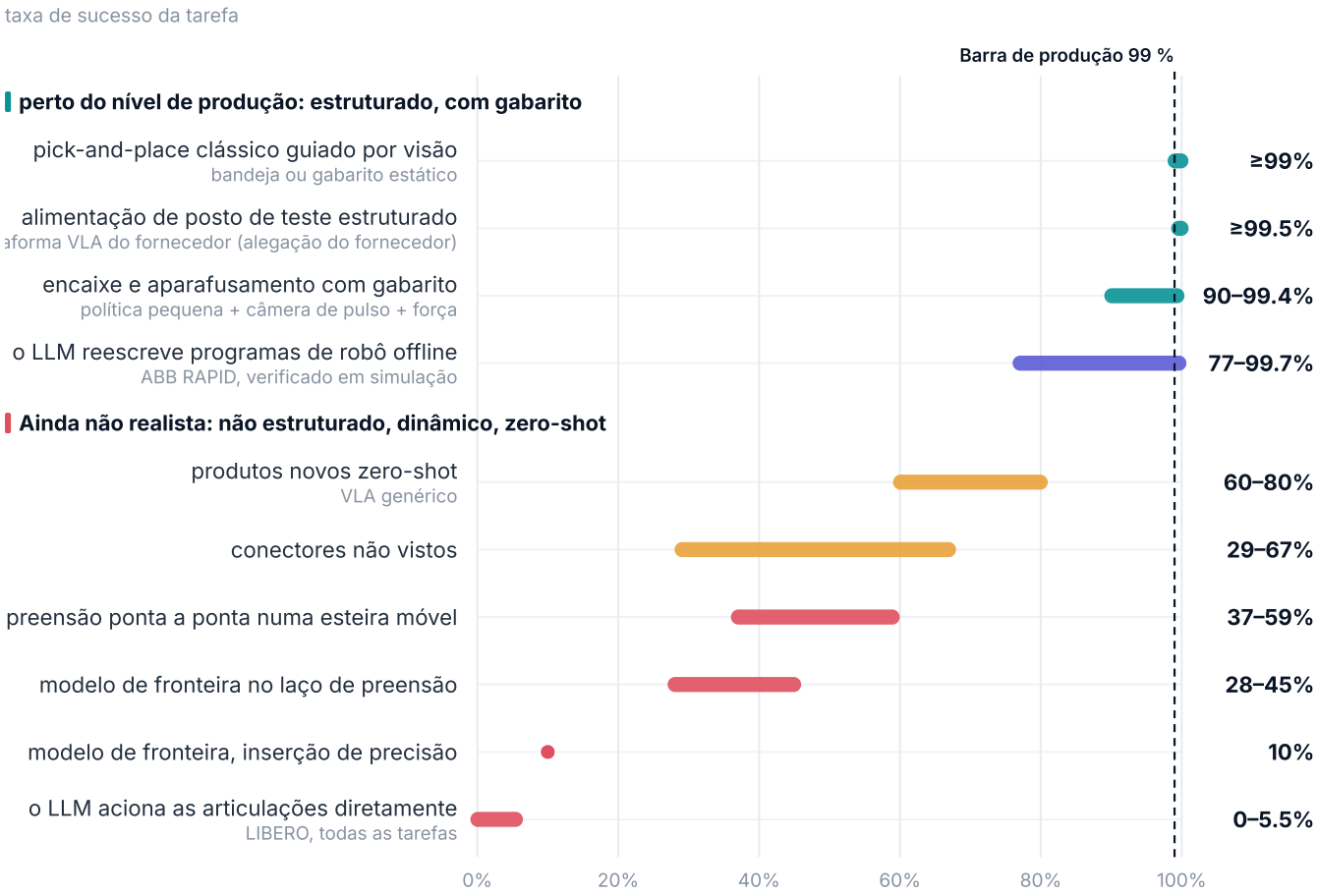
FIG 15 **Latência de política edge** ***Uma política ACT pequena roda em ≈50 ms numa placa edge; um VLA portado para um chip amador leva 7,3 s.***



Fonte: Documentação e fórum da D-Robotics (números do fornecedor); ports comunitários do GitHub; pesquisa da Anthropic.

Nos chips de borda chineses, a única política aprendida medida é uma política ACT no RDK S100 da D-Robotics a 46–50 ms por bloco de 50 passos, sem taxa de sucesso publicada (docs D-Robotics RDK S) vendor claim; um porte amador do SmolVLA num RK3588 demora cerca de 7.3 s por bloco (GitHub). A quantização não compra velocidade de graça: nenhuma compilação π0.5 a 4 bits foi mais rápida do que a 16 bits com tamanho de lote 1 (LoopQ). As demonstrações são caras (a Reuters referiu cerca de 300 tentativas por movimento utilizável para um teleoperador novato (Notebookcheck)), e a validação custa mais: provar uma taxa de falha ≤0.1% com 95% de confiança demora cerca de 3,000 sucessos consecutivos, 17–22 horas-robô estimate. Nenhuma política de manipulação aprendida tem certificação de segurança.

FIG 16 **Alunos robô: perto do nível de produção só quando o posto é estruturado** ***Postos com gabarito alcançam 90–99 %+; tarefas não estruturadas, dinâmicas e zero-shot ainda não são realistas.***



Fonte: Artigos e material de fornecedores citados no capítulo 3; alguns números são alegações do fornecedor.

Tarefa de eletrónica	Aluno viável	Sucesso esperado	Velocidade vs uma pessoa	Custo por troca	Veredito 2026–2028
Pick-and-place de elevada mistura, kitting, alimentação de máquinas (tabuleiro ou fixação estática)	Perceção aprendida mais movimento programado ou planeado	≥99% com orientação por visão clássica	Igual às estações convencionais	Horas	Viável já
Inserção de conector fixado, FFC e cabo, aparafusamento, soldadura	Máquina de estados + política pequena BC/difusão/ACT + câmara de pulso + força/binário (idealmente toque) + correção humana	90–99.4%	10–15% mais lenta	10–300 demos (≈15–70 min) + 1,000–3,000 execuções de validação (6–22 horas-robô)	Viável, intensivo em I&D meses para a primeira estação
Alimentação de estação de teste estruturada e triagem OK/NG	Plataforma VLA de fornecedor + gémeo digital + RL em robô real	≥99.5% vendor claim, conflituoso	18–20 s por passo, dentro de um takt de 20 s	≤4 h de retreino + 15 min de calibração	Viável numa plataforma de fornecedor
Programação de troca	O LLM escreve ou edita programas de robô offline, verificados em simulação	77–99.7% em reescritas ABB RAPID	Sem efeito no takt	Minutos a horas por programa	Viável como ferramenta de engenharia

Tarefa de eletrónica	Aluno viável	Sucesso esperado	Velocidade vs uma pessoa	Custo por troca	Veredito 2026–2028
Agarramento aprendido de ponta a ponta em transportadores em movimento	Aluno VLA	37–59%; ≤37% em tapetes rápidos	—	≥458 demos por configuração	**Não antes de 2028; usar rastreamento clássico mais perceção aprendida**
Novos produtos sem novos dados (zero-shot)	VLA genérico	60–80%; conectores nunca vistos 29–67%	—	—	**Não realista**
LLM de fronteira na malha a acionar o braço	Agentes GPT-6, Claude, Gemini	Agarramento 28–45%; inserção de precisão 10%	≈30 min ou 280 s por tarefa vs ≈20 s	—	**Não realista**
Cablagens de encaminhamento livre e montagem flexível de vários passos	—	Sem evidência de produção em 2026	—	—	**Não antes de 2028**

Fontes: os estudos acima e (Universidade de Passau sobre reescrita RAPID, GraspVLA, Komax) **na maioria single source e dados de fornecedores** . Em 2026–2028, o valor de um modelo grande para robôs é como prior pré-treinado e programador offline, não como algo que "ensina" um aluno. Todos os casos perto da produção que encontrei (Neuromeka, AgiBot na Longcheer, Fraunhofer, ContactDP) dependem de demonstrações humanas, deteção de força ou toque, simulação ou gémeos digitais e RL supervisionada. Nenhum depende de etiquetas de um modelo de fronteira.

Custos e retorno

04 Custos e retorno: a revisão com IA amortiza-se em meses, a alimentação de máquinas em anos, os braços de retirada nunca

Custo da mão de obra: os EUA são cerca de 5.5 vezes Shenzhen, Taiwan cerca de 1.5 vezes, a maior parte do Sudeste Asiático abaixo de metade

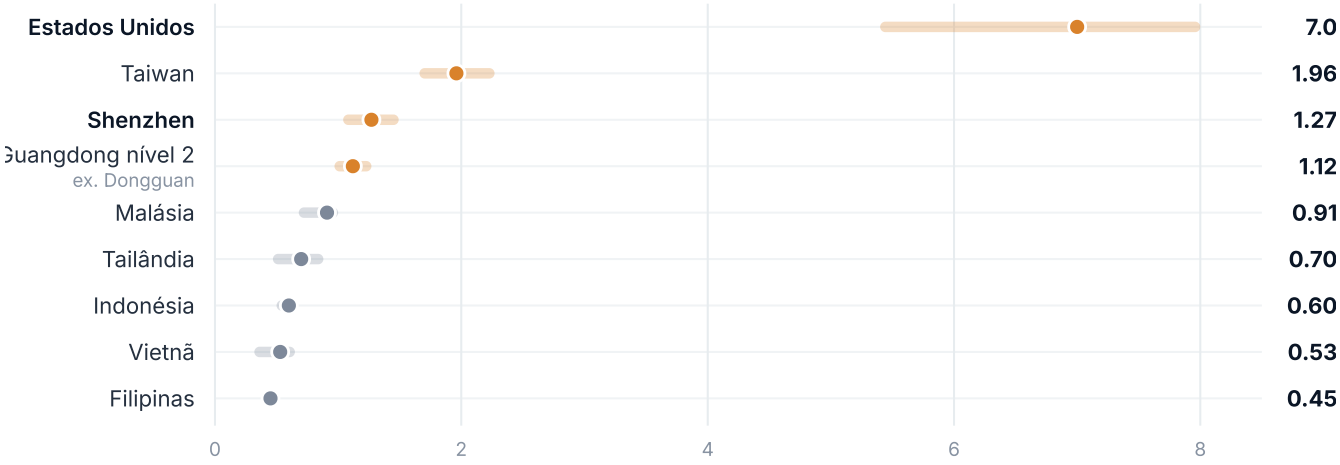
O retorno de substituir mão de obra começa pelo custo dessa mão de obra. A orientação salarial oficial de Shenzhen coloca a mediana de 2024 para um inspetor de qualidade industrial em **RMB 74,074** por ano; os salários são 86.92% do custo total da mão de obra na indústria eletrónica, pelo que o custo totalmente carregado é de cerca de RMB 85,000. O inquérito de 2025 da JETRO coloca o custo anual de um trabalhador fabril chinês em $12,667, cerca de RMB 85,100, dentro de 0.2% (orientação salarial de Shenzhen, JETRO 2025) verified .

Região	Custo anual totalmente carregado de um inspetor ou operador (baixo / base / alto)	Base
China continental, Shenzhen	RMB 70k / **85k** / 100k ($10.4k / **12.7k** / 14.9k)	Mediana oficial × 1.15; corresponde à JETRO verified
Áreas salariais de nível 2 de Guangdong (Dongguan etc.)	RMB 65k / **75k** / 85k	Mediana de Guangdong 2023 × 1.15 mais dois anos de aumentos estimate
Delta do Rio Yangtzé (Suzhou, Kunshan, Wuxi)	Como Shenzhen	Sem base para mão de obra de inspeção mais barata; o salário mínimo de topo de Jiangsu está cerca de 5.6% acima do de Shenzhen estimate
Taiwan	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	Inquérito salarial ocupacional do Ministério do Trabalho mais ≈1.18 encargos legais; JETRO como extremo alto estimate
Malásia	$6.8k / **9.1k** / 10.0k	JETRO e salários mínimos estimate
Tailândia	$4.7k / **7.0k** / 8.8k	Idem
Indonésia	$5.0k / **6.0k** / 6.3k	Idem
Vietname	$3.2k / **5.3k** / 6.5k	Idem
Filipinas	$4.0k / **4.5k** / 4.9k	Idem
Estados Unidos	$54k / **70k** / 80k	Mediana de inspetor BLS maio 2025 $48,570 ÷ salários a 65.5% da remuneração estimate

Fontes: (JETRO 2025, orientação salarial de Guangdong, gabinete de RH de Wuxi, inquérito salarial ocupacional de Taiwan, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **O que custa por ano um trabalhador de inspeção ou movimentação** ***Plenamente carregado, os EUA são ≈5,5× Shenzhen e Taiwan ≈1,5×.***

custo anual carregado por trabalhador de CQ ou movimentação (10k $; ponto = base, barra = baixo–alto)



Fonte: Orientação salarial oficial de Shenzhen, Ministério do Trabalho de Taiwan, JETRO, BLS; estimativas do autor.

Três correções estabelecem a base. Primeira, Shenzhen subiu o piso de contribuição para pensões de RMB 4,775 para 5,521 por mês em julho de 2026 (Shenzhen Bendibao) single source, pelo que um inspetor de Shenzhen em 2026 custa mais provavelmente RMB 80–90k. Segunda, um posto não é uma pessoa: dois turnos de 12 horas precisam de cerca de 2.0 pessoas (seis dias por semana) a 2.4 (todo o ano) no Delta do Rio das Pérolas, cerca de 2.7–3.1 no Sudeste Asiático sob limites de horas extraordinárias, e 4.2–4.5 para 24×7 nos EUA e em Taiwan estimate. A revisão com IA normalmente permite a um revisor cobrir duas a quatro linhas, pelo que a poupança realista é de **0.5–1.5 pessoas a tempo inteiro por linha, não uma pessoa por linha por turno**. Terceira, os compradores em Taiwan, Vietname e Malásia não procuram apenas custo: no inquérito da JETRO, 61.1% das empresas em Taiwan, 48.2% no Vietname e 44.9% na Malásia disseram que contratar se tornou mais difícil nos últimos dois anos, contra 18.6% na China continental (JETRO 2025) single source.

Preços de equipamento: quatro bases que não devem ser alvo de média

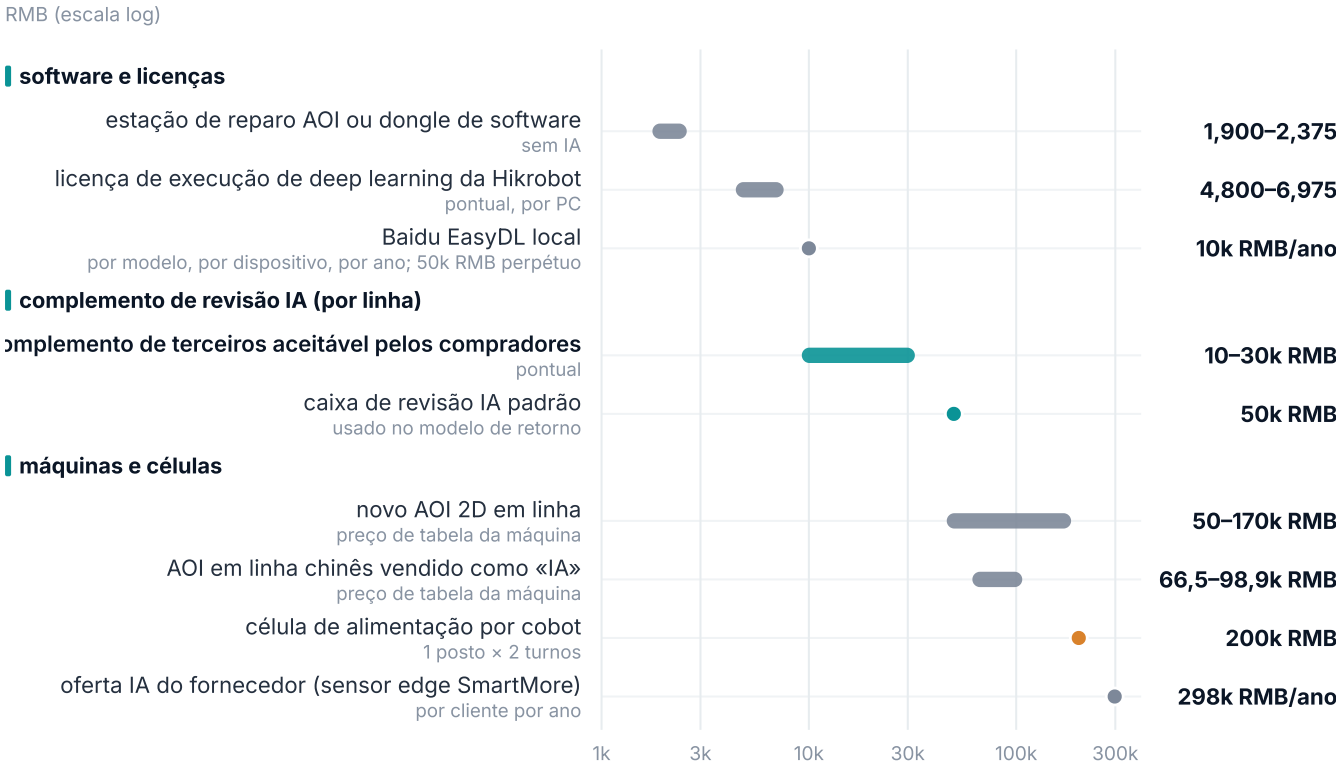
Os preços de AOI divergem sobretudo porque medem coisas diferentes. As listagens online (novembro de 2026, novas, usadas e aluguer misturados) rondam RMB 17–80k para 2D offline ou de secretária, RMB 50–170k para 2D inline e RMB 80–475k para 3D inline, até RMB 700k para 3D coreano (1688 AOI inline, 1688 AOI offline, 1688 AOI 3D) single source, preços de tabela. Um concurso de novembro de 2025 limitou um AOI 3D mais um SPI 3D a RMB 800k (aviso de concurso). O prospeto da Sinic-Tek implica RMB 428–512k por AOI 3D em 2019–2022, cerca de RMB 340–470k pelo seu relatório de 2025 (prospeto da Sinic-Tek, relatório anual 2025 da Sinic-Tek) primário + estimate. Um fornecedor no estrangeiro estima $3,000–20,000 para 2D de secretária, $30–70k 2D inline, $50–90k 3D de entrada e $100–200k+ 3D de gama alta, com a mesma máquina inline a $30–50k na China (Averroes) [estimativa de fornecedor de um vendedor de software de IA concorrente].

A revisão com IA em si não tem preço público em RMB; os fornecedores fazem cotações caso a caso após uma avaliação gratuita no local (Sohu sobre Hengtianyi), e as pesquisas em listagens de comércio eletrónico, concursos, artigos de fornecedores e meios taiwaneses não encontraram nenhuma cotação pública (1688, 51SMT). As âncoras: cerca de €27k por linha implícitos no caso da Siemens estimate; $21,924 de tabela para o módulo de rede de revisão remota de um distribuidor dos EUA (Manncorp) [preço de tabela de fornecedor]; cerca de RMB 298k por cliente por ano para os sensores de IA de borda da SmartMore em 2025 (36Kr) single source. Um AOI 2D inline nacional vendido como "IA" tem tabela de RMB 66.5k (1688 AOI inline), e um buffer de NG para reinspeção manual cerca de RMB 50k (1688 revisão AOI). Assim, o teto para um complemento de revisão de terceiros é de cerca de RMB

10–30k por linha, ou uma taxa anual semelhante; acima disso, o comprador preferiria substituir a máquina por um AOI nacional com IA integrada estimate.

Âncora	Preço (RMB)	O que compra	Fonte
Estação de reparação de AOI ou dongle de software (TRI, Zhenhuaxing antigos)	≈1,900–2,375	Chave de licença, sem IA	1688 [preço de tabela]
Runtime geral de visão por aprendizagem profunda (dongle Hikrobot VisionMaster)	≈4,800–6,975 pagamento único por PC	A plataforma; o cliente constrói e mantém os modelos	1688 [preço de tabela]
Implantação no local Baidu EasyDL	10,000 por modelo por dispositivo por ano, ou 50,000 perpétuo	Serviço de inferência empacotado	docs Baidu AI [página de preços de 2022]
AOI inline nacional vendido como "IA"	66,500–98,900 por máquina	Substituição de máquina	1688 [preço de tabela]
Sensores de IA de borda SmartMore	≈298,000 por cliente por ano	Solução de fornecedor em várias máquinas	36Kr single source

FIG 18 **A escada de preços** ***De alguns milhares de RMB de licença de software a 200–300k RMB por uma célula robótica ou uma oferta de IA do fornecedor.***



Fonte: Anúncios 1688, documentação Baidu AI (tabela 2022), 36Kr, uma entrevista do setor (não verificada).

Uma taxa anual estilo Baidu, com um servidor de inferência para uma a quatro linhas e um a quatro modelos cada, implica uma subscrição de RMB 2.5–40k por linha por ano, realisticamente cerca de RMB 5–10k; a 20% de uma licença perpétua, a subscrição e a compra empatam no quinto ano, uma estrutura que vale a pena copiar estimate. Os compradores estão habituados a taxas de software quase nulas: toda a receita "outros" da Sinic-Tek (peças, serviço, aluguer, acessórios) foi 7.65% das vendas de 2025, cerca de RMB 2,300–3,100 por máquina instalada por ano (relatório anual 2025 da Sinic-Tek) primário + estimate. Uma subscrição de IA de RMB 5–10k por linha por ano é

duas a cinco vezes o que os compradores gastam agora em software de AOI, pelo que tem de ser justificada pelas horas de revisão que liberta (cerca de RMB 75–90k por pessoa a tempo inteiro por ano), não por comparação com as taxas de software existentes.

Para os braços robóticos, o preço médio dos cobots na China caiu para cerca de RMB 58,600 em 2025 (GGII) estimate, e a média de seis eixos da Dobot caiu 19.1% nesse ano, à medida que modelos mais baratos ganharam quota (Huxiu) verified, embora os preços de tabela dos mesmos modelos tenham subido 5–15% em 2026 (Sina Finance). Nos EUA, um FANUC CRX-10iA tinha tabela de $51,434 (Vention) e compilações colocam um UR7e em $30–48k (cobotcost) estimate. O braço é apenas 40–50% de uma célula dos EUA (Standard Bots); na China, câmaras, garras, segurança e engenharia chegam facilmente a duas a quatro vezes um braço de RMB 40k, pelo que orçamente o custo absoluto de engenharia em vez de um múltiplo estimate.

Retorno: revisão em meses, alimentação em anos, retirada nunca

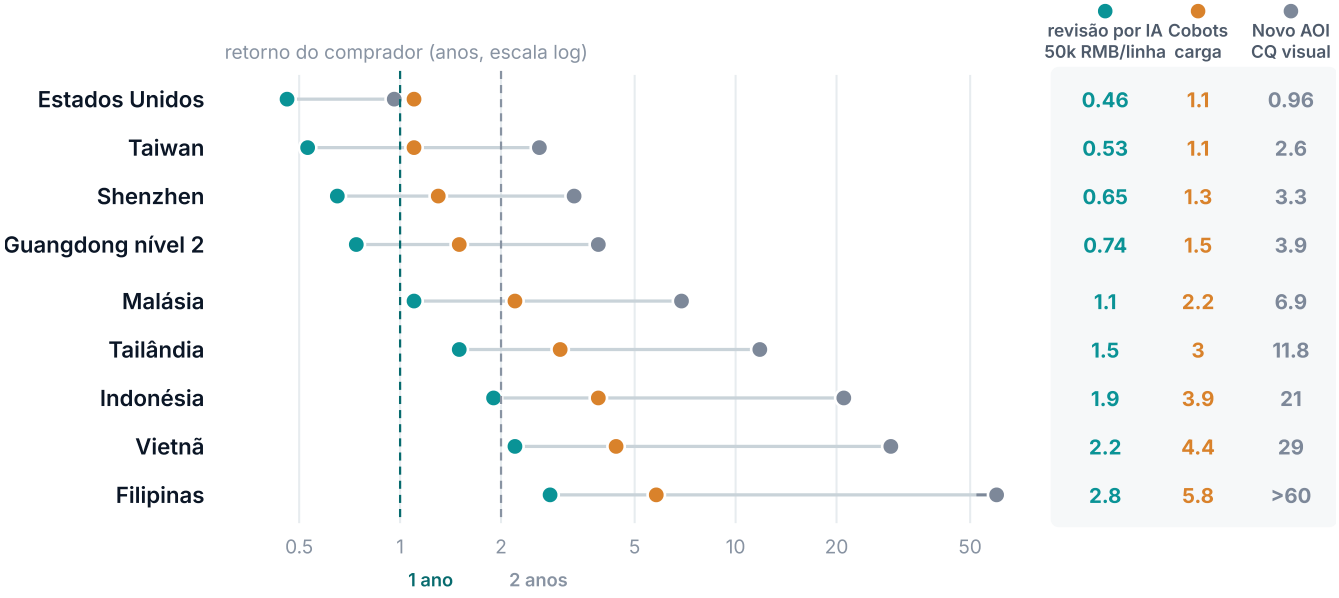
Retorno simples = investimento ÷ (pessoas poupadas × custo de mão de obra carregado − custo anual de funcionamento), para uma linha de dois turnos sem subsídios nem direitos aduaneiros. O equipamento de Taiwan tem preço de 1.3 vezes a China, Vietname, Malásia e Tailândia 1.15 vezes, Indonésia e Filipinas 1.2 vezes, e os EUA a preços do estrangeiro estimate.

Cenário	Investimento (China / EUA)	Pessoas poupadas	Custo anual de funcionamento
(a) Caixa de revisão com IA ou caixa de câmara com IA	RMB 50k / $30k por linha	1.0 por linha	15%
(b) Novo AOI 2D inline a substituir verificações visuais manuais (fica 0.5 revisor por turno)	RMB 200k / $60k	1.0	12%
(c) Célula de carga com cobot (uma estação × dois turnos)	RMB 200k / $140k	2.0	8%
(d) Célula integrada de deteção-e-retirada (inspeção 1.0 + retirada 0.07)	RMB 300k / $200k	1.07	8%
(e) Um braço robótico adicionado apenas para retirar placas NG	RMB 150k / $100k	0.07	8%

Região	(a) Revisão com IA	(b) Novo AOI	(c) Célula de cobot	(d) Deteção-e-retirada	(e) Braço de retirada
Shenzhen	0.65 anos (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	Nunca
Guangdong nível 2	0.74	3.9	1.5	5.3	Nunca
Taiwan	0.53	2.6	1.1	3.6	Nunca
Malásia	1.1	6.9	2.2	9.2	Nunca
Tailândia	1.5	11.8	3.0	15	Nunca
Indonésia	1.9	21	3.9	25	Nunca
Vietname	2.2 (1.6–4.5)	29	4.4	34	Nunca
Filipinas	2.8	>60 (custo de funcionamento ≈ poupança)	5.8	≈100	Nunca

Região	(a) Revisão com IA	(b) Novo AOI	(c) Célula de cobot	(d) Deteção-e-retirada	(e) Braço de retirada
Estados Unidos	0.46	**0.96**	**1.1**	3.4	Nunca

FIG 19 **Retorno por região** ***A revisão IA se paga em meses nos EUA, Taiwan e Shenzhen e em 1–3 anos no Sudeste Asiático; a alimentação por cobot, em anos.***



Um braço robótico só para extrair placas NG nunca se paga, em lugar nenhum
Com rendimento de primeira passagem acima de 90 %, a extração ocupa só 0,003–0,14 FTE por turno

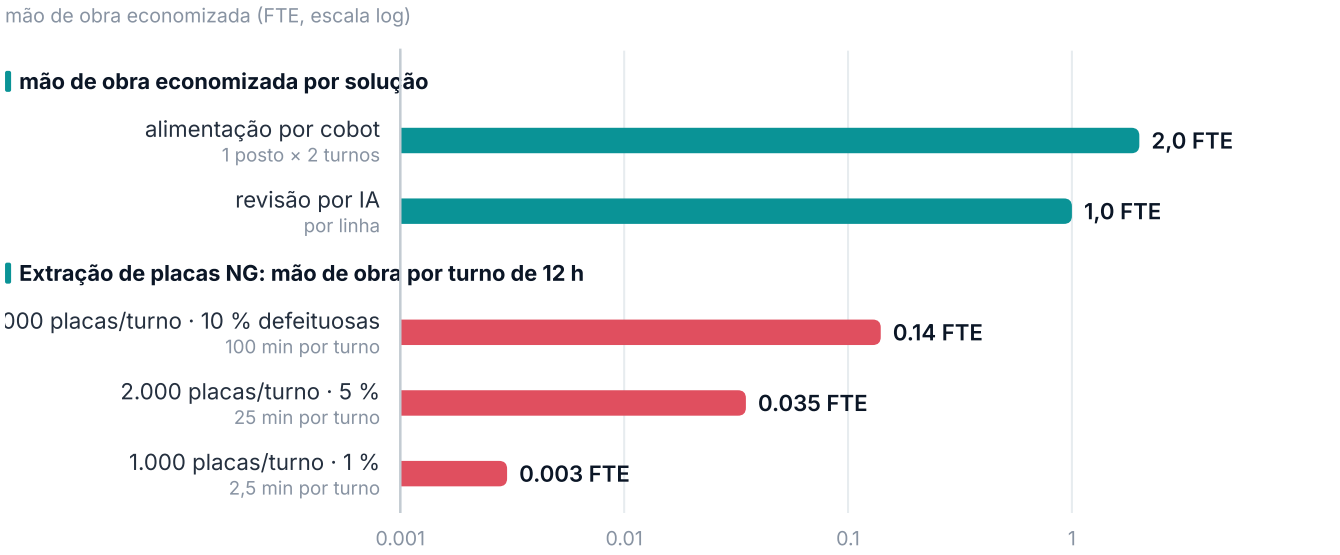
Fonte: Modelo do autor: caixa de revisão IA de 50k RMB por linha; célula de alimentação por cobot a 200k RMB ou 140k $, dois turnos.

Todos os valores são estimativas a partir dos dados de entrada acima. Sensibilidades: reduza as pessoas poupadas para metade e cada retorno duplica sensivelmente; execute um turno e (b)–(d) duplicam; a depreciação de 100% no primeiro ano nos EUA encurta o retorno em cerca de 20%, enquanto os direitos aduaneiros sobre equipamento chinês, japonês ou coreano o alongam em cerca de 10–40%; um subsídio local de 15% na China encurta-o em cerca de 15%. Não encontrei nenhum inquérito independente sobre os limiares de retorno dos compradores; os fornecedores dizem que o AOI se amortiza "normalmente em 9–18 meses" (Averroes) e o caso da Siemens reivindica oito meses vendor claims. Uso um e dois anos como linhas de filtragem. O modelo deixa de fora os ganhos de qualidade (garantia, devoluções) e os custos de rotatividade e formação, pelo que é conservador.

A parte contraintuitiva de "detetar e retirar" é que a mão de obra está na revisão, não em retirar as poucas placas NG reais da linha. A 15 segundos de manuseamento por placa NG estimate:

Placas por turno	Taxa de defeito	Tempo de retirada por turno de 12 horas	Pessoas
1,000	1%	2.5 min	0.003
2,000	5%	25 min	0.035
4,000	10%	100 min	0.14

FIG 20 **Por que extrair placas ruins não se paga** ***Acima de 90 % de rendimento de primeira passagem, a extração ocupa só 0,003–0,14 de pessoa por turno.***



Fonte: Estimativas do autor.

O caso base (2,000 placas por turno, 5% defeituosas, dois turnos) é 0.07 de uma pessoa, no valor de cerca de RMB 6,000 por ano em Shenzhen, menos do que o custo anual de funcionamento de cerca de RMB 12,000 do braço, pelo que nunca se amortiza em lado nenhum; mesmo o pior caso (4,000 placas por turno, 10% defeituosas) demora cerca de 13 anos em Shenzhen e 9 nos EUA estimate. Se a preocupação é parar a linha para retirar, um desviador de NG ou um transportador de buffer padrão resolve-o sem braço. As pessoas do setor dizem "acima de 90% de rendimento à primeira passagem, pare a linha e retire a placa à mão", e **isso está certo para a retirada, mas não para a inspeção**. O desenho de produto numa linha SMT é, portanto, "inspecionar mais marcar ou desviar", com braços robóticos reservados para os lugares onde cada peça é manipulada (carga de fixação de teste, manuseamento de tabuleiros e revistas, embalamento), onde a inspeção se torna um complemento quase gratuito e o cenário (c) se aplica.

Subsídios, impostos e direitos aduaneiros movem o retorno em 15–25%, mas não a ordenação

China continental: o apoio baixa sobretudo o custo de financiamento; máquinas individuais raramente se qualificam diretamente. Um subsídio nacional de empréstimo paga 1.5 pontos de juro por até dois anos em 14 cadeias prioritárias (Yicai) verified, cerca de RMB 9,000 numa máquina de RMB 300k; a recompensa de atualização de Guangdong de até 20% exige um projeto registado com pelo menos RMB 7 milhões de equipamento novo e exclui Shenzhen (regras de Guangdong), e Dongguan pagou efetivamente 4.42% em julho de 2026 (alocação de Dongguan) single source, primário. Um dispositivo de inspeção com IA é aprovado apenas como equipamento "com software de suporte" dentro do projeto registado do comprador, pelo que o caso base assume ausência de subsídio.

Taiwan: os incentivos vão para o comprador. Sob o artigo 10-1 do Estatuto para a Inovação Industrial, alterado em maio de 2025 para incluir IA, um comprador pode creditar 5% das despesas de 2025–2029 em nova maquinaria inteligente ou IA (incluindo inspeção baseada em CNN/DNN) contra o imposto sobre o rendimento desse ano, ou 3% ao ano durante três anos (Estatuto, regras de crédito) verified; as bolsas do lado do fornecedor estão fechadas ao capital do continente (MOEA) single source.

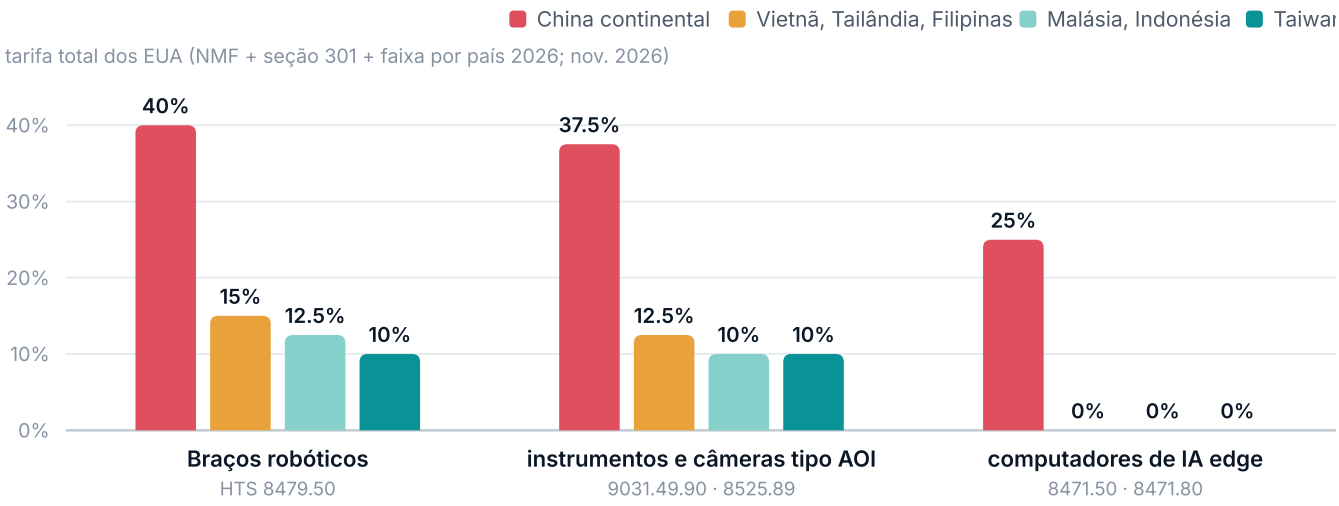
Sudeste Asiático: só a Malásia e a Tailândia recompensam a compra de automação. O subsídio de capital de automação de 200% da Malásia sobre os primeiros RM10 milhões vale cerca de 24% do capex para um comprador lucrativo (MIDA, PwC Malásia) verified; o BOI da Tailândia dá uma isenção fiscal de três anos até 50–100% de um investimento de atualização, o limite mais alto quando estão envolvidos fornecedores de automação tailandeses

(BOI) single source, primário, pelo que entrar na Tailândia funciona melhor com um integrador local. O Vietname, a Indonésia e as Filipinas recompensam a construção de fábricas em vez da compra de automação (Vietnam Law Magazine, PwC Filipinas).

Estados Unidos: a depreciação encurta o retorno; os direitos aduaneiros alongam-no. Os ativos comprados após 19 de janeiro de 2025 têm depreciação permanente de 100% no primeiro ano, cerca de 21% de escudo fiscal federal para um comprador lucrativo (IRS Notice 2026-11) single source, primário; a investigação da Secção 232 sobre robôs e maquinaria industrial terminou a 28 de outubro de 2026 com negociações em vez de novos direitos, a reavaliar dentro de 180 dias (Casa Branca, Mayer Brown) verified. A montagem final no Sudeste Asiático reduz a sobretaxa sobre braços e instrumentos do tipo AOI de cerca de 37.5–40% para 10–15% (capítulo 1).

Região	Instrumentos principais	Efeito no retorno do comprador estimate
China continental	Subsídio de empréstimo de 1.5 pontos por até dois anos; recompensas de atualização provinciais e municipais (Guangdong exige projetos registados de ≥RMB 7M; Dongguan pagou efetivamente 4.42%); medidas de Shenzhen 2026–2029 (taxas não públicas)	Cerca de 0–5% para uma máquina individual; 4–15% dentro de um grande projeto
Taiwan	Crédito §10-1 do Estatuto, 5% ou 3% × 3 anos	Cerca de 5% mais curto, se o comprador pagar imposto suficiente
Malásia	Subsídio de capital de automação de 200% (primeiros RM10M)	Cerca de 15–25% mais curto para compradores lucrativos
Tailândia	Isenção fiscal de três anos do BOI até 50–100% do investimento	Cerca de 15–25% mais curto para compradores lucrativos fora de um feriado fiscal
Vietname, Indonésia, Filipinas	Incentivos de imposto sobre o rendimento para construção de fábricas; depreciação extra de 20% do equipamento nas Filipinas	Quase nenhum para compras individuais de automação
Estados Unidos	Depreciação de 100% no primeiro ano; ≈37.5–40% de direito combinado sobre equipamento fabricado no continente	Depreciação ≈20% mais curta; direitos ≈10–40% mais longos

FIG 21 **Tarifa dos EUA sobre robôs, AOI e computadores edge** ***Braços fabricados na China continental pagam ≈40 %; enviados do Sudeste Asiático ou Taiwan, 10–15 %.***



Fonte: Capítulo 99 do Tarifário Harmonizado USITC; Federal Register; USTR (novembro de 2026).

Por região: na China continental, uma caixa de revisão com IA amortiza-se em meses e uma célula de cobot em 1.3–1.5 anos, enquanto um novo AOI comprado apenas para cortar mão de obra demora 3–4 anos e a deteção-e-retirada integrada 4.5–5.3, ambas além de um ecrã de dois anos. Taiwan faz ligeiramente melhor do que Shenzhen (0.53 anos para a revisão, 1.1 para uma célula de cobot), e contratar lá é mais difícil. Na Malásia e na Tailândia, uma caixa de IA demora 1–1.5 anos e uma célula de cobot 2–3, cerca de 15–25% mais rápido com incentivos fiscais locais para compradores lucrativos. Na Indonésia, no Vietname e nas Filipinas, só uma caixa de IA muito barata se amortiza em 2–3 anos; a procura aí vem mais dos requisitos de qualidade dos clientes e da escassez de mão de obra. Nos EUA, todas as opções exceto um braço de retirada puro se amortizam em 0.2–3.4 anos, com os direitos aduaneiros, a reavaliação da Secção 232 e a certificação NRTL como as principais variáveis.

Mercado e concorrência

05 Mercado, concorrência e como os compradores pagam

Dimensão do mercado: os números publicados diferem por múltiplos; só uma contagem ascendente é útil

Cada número de mercado tem de ser lido com o seu âmbito. O mercado de visão artificial da China foi de RMB 18.15 mil milhões em 2024 (Zhiyan) verified, com a eletrónica 3C como o maior comprador; as estimativas globais para a visão artificial diferem cerca de três vezes e para o AOI cerca de 6.5 vezes (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Ascendente a partir das receitas da Koh Young, TRI, Jutze, Sinic-Tek e ViTrox, o equipamento global de inspeção SMT é da ordem de $1–2 mil milhões estimate.

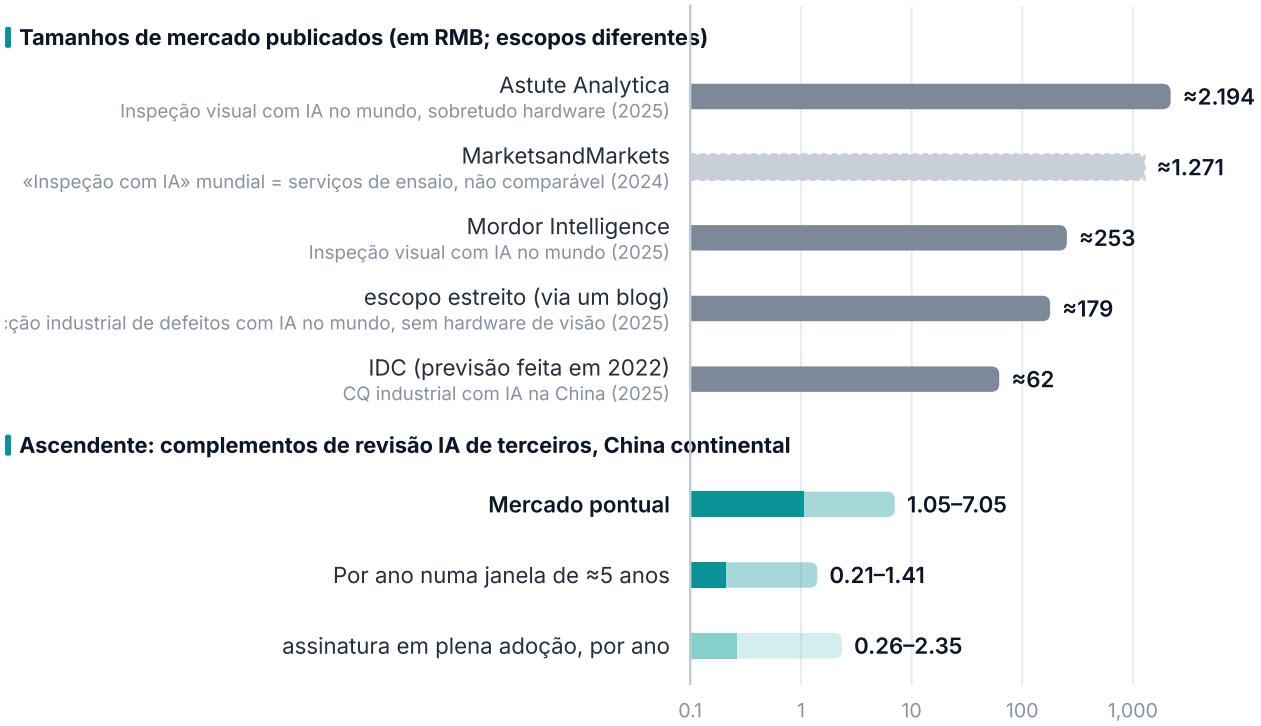
Não existe um mercado autónomo de "software de inspeção visual com IA". Os números publicados diferem em mais de dez vezes e todos incluem hardware ou serviços:

Âmbito	Publicador	Valor
"Deteção de defeitos industriais com IA" global (estreito, excluindo hardware de visão)	Empresa não nomeada, via um blog da iFactory	$2.66 mil milhões em 2025, 8.6% CAGR até 2035, eletrónica 34%
"Inspeção visual com IA" global	Mordor Intelligence	$3.76 mil milhões em 2025, $12.8 mil milhões em 2031, 22.67% CAGR
"Inspeção visual com IA" global (sobretudo hardware)	Astute Analytica	≈$32.66 mil milhões em 2025
"Inspeção com IA" global (serviços de teste e certificação; não comparável)	MarketsandMarkets	$18.92 mil milhões em 2024
"Inspeção de qualidade industrial com IA" da China	IDC, previsão feita em 2022	$958 milhões em 2025 (≈RMB 6.2 mil milhões), 91.5% em comunicações, eletrónica e duas outras indústrias

Fontes: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) **cada uma single source, na maioria resumos de relatórios pagos**.

FIG 22 **Os tamanhos de mercado diferem em mais de dez vezes*Apenas o escopo estreito e uma contagem ascendente são úteis para decisões.***

100 M RMB (escala log; USD a 6,7188)



Fonte: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; estimativa ascendente do autor.

Ascendente, só há uma contagem pública de linhas SMT: a New Venture Research estimou cerca de 73,231 linhas em todo o mundo em 2020, 47% delas na China (19% em 2000), cerca de 70% de gama alta e 30% de gama baixa (EMSNow, resumo NVR SMT2020) single source, 2020 ; a NVR pertence agora à IPC e os dados mais recentes são pagos (NVR). Crescendo 3–5% ao ano com uma quota da China de 40–47%, isso dá cerca de 87,000–98,000 linhas em todo o mundo em 2026 e 35,000–46,000 na China continental (base 40,000) estimate .

Entrada	Baixo	Base	Alto	Base
Linhas SMT da China continental (2026)	35,000	40,000	46,000	Base NVR 2020 × crescimento × quota
Quota sem revisão com IA	75%	80%	85%	Janela de retrofit, abaixo
Quota que um terceiro pode alcançar (não interna, não ligada à IA de OEM, pode pagar)	40%	50%	60%	Pressuposto: ≈30% de linhas de gama baixa mais algumas linhas de gama alta de marcas mistas, excluindo construções internas estilo DBG
Linhas aproveitáveis	**10,500**	**16,000**	**23,500**	Produto
Preço único por linha	RMB 10k	RMB 20k	RMB 30k	Escada de preços e teto (capítulo 4)
Mercado aproveitável único (base instalada)	**RMB 105M**	**RMB 320M**	**RMB 705M**	Produto
Por ano numa janela de ≈5 anos	RMB 21M	RMB 64M	RMB 141M	A janela encolhe 8–12% ao ano
Subscrição (RMB 2.5k / 5k / 10k por linha por ano, adoção total)	RMB 26M por ano	RMB 80M por ano	RMB 235M por ano	Âncora de taxa anual estilo Baidu

FIG 23 **Ascendente: 40.000 linhas → 16.000 atendíveis → ≈320 M RMB pontuais** ***O nicho de revisão IA de terceiros na China continental é real, mas pequeno.***



Um novo entrante ganhando 2–5 % das linhas atendíveis em 3 anos: ≈320–800 linhas, ≈2–5 M RMB por ano
Como assinatura (5k RMB por linha por ano): 1,6–4 M RMB recorrentes, só 0,3–2,3 % da previsão da IDC para o CQ industrial com IA na China

Fonte: NVR (base 2020); estimativas do autor.

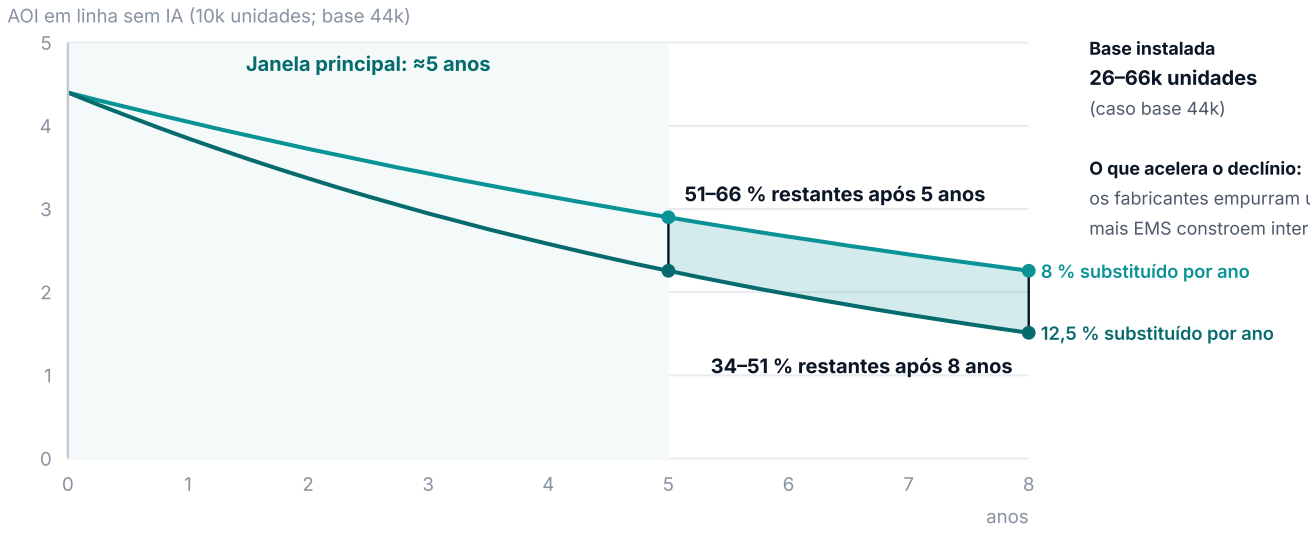
Tudo isto são estimativas; nem a quota nem o preço foram medidos publicamente. O nicho vale apenas 0.3–2.3% ao ano da previsão da IDC para toda a inspeção de qualidade industrial com IA da China em 2025 (cerca de RMB 6.2 mil milhões). Um recém-chegado que conquiste 2–5% das linhas aproveitáveis de base em três anos teria cerca de 320–800 linhas: RMB 6.4–16 milhões no total a RMB 20,000 por linha, cerca de RMB 2–5 milhões por ano, ou RMB 1.6–4 milhões por ano em receita recorrente a RMB 5,000 por linha estimate. Isso chega para uma unidade de negócio dentro de uma empresa, não para uma empresa autónoma de escala de capital de risco. Cada 1,000 linhas instaladas em Taiwan, no Sudeste Asiático ou nos EUA valem cerca de RMB 3–15 milhões de pagamento único a preços do continente; os preços são mais altos em Taiwan e nos EUA, mas os compradores aí usam mais frequentemente AOIs 3D de OEM com opções de IA estimate.

A janela de retrofit é de cerca de cinco anos

Nenhum fabricante de AOI publica a vida útil da máquina; as empresas de EMS depreciam o equipamento em 3–10 anos (relatório anual 2025 da DBG), e os AOIs mais antigos perdem atualizações de software e peças após o fim de vida (I-Connect007) single source. Uso 8–12 anos (base 10).

A 1.0–1.65 AOIs inline por linha, o continente tem cerca de 35,000–78,000 AOIs inline em serviço (base cerca de 55,000), consistente com os envios conhecidos (Circuits Assembly, relatório anual 2025 da Sinic-Tek). Nenhuma fonte mede quantos executam revisão com IA; assumindo que as máquinas nacionais a oferecem desde cerca de 2023 e que 40–60% das entregas recentes a usam, **cerca de 26,000–66,000 AOIs inline na China continental (base cerca de 44,000) não executam IA** estimate. Se cada máquina nova passar a incluir IA, esse parque encolhe 8–12.5% ao ano: 51–66% permanece após cinco anos e 34–51% após oito, mais depressa se os fabricantes de AOI empurrarem atualizações de IA para máquinas mais recentes. **A maior parte da oportunidade cai nos próximos cinco anos ou assim**, com as EMS de média dimensão a executar AOIs 2D nacionais mais antigos, marcas mistas e sem caminho de atualização de IA no seu núcleo.

FIG 24 **A janela de retrofit é de cerca de cinco anos** ***Os AOI em linha sem IA encolhem 8–12,5 % ao ano à medida que são substituídos ou atualizados.***



Fonte: Estimativas do autor; base instalada e participações não verificadas de forma independente.

Os concorrentes já são fortes: a cunha de revisão está cercada

A cunha de revisão com IA está a estreitar-se a partir de três direções. Os líderes de AOI e visão estão a integrar IA nas novas plataformas: a Jutze "adotou totalmente algoritmos de IA" (relações com investidores da Jutze, junho de 2026), a receita de IA industrial da OPT cresceu 96.77% em 2025 (relatório anual 2025 da OPT) single source, primário , e a ViTrox reivindica até 95% de precisão para revisão autónoma (ViTrox) vendor claim . Os grandes fabricantes constroem os seus (iThome). E as empresas de IA financiadas por capital de risco vendem em escala: os clientes de sensores de borda da SmartMore cresceram de 65 para 251 em dois anos (36Kr) single source , a Hikvision vende o seu modelo Guanlan para inspeção (Xinhua), e a integração "olho, cérebro e mão" já está em oferta da SmartMore, Hikrobot e Mech-Mind (36Kr, Securities Star).

FIG 25 **A cunha de revisão está cercada por três lados** ***O espaço para terceiros está em EMS médios e em nichos de inspeção mal cobertos pelo AOI.***



Fonte: Relatórios anuais de Jutze, Sinic-Tek e DBG; material público de Hikvision e SmartMore; compilado pelo autor.

O espaço deixado a um recém-chegado é estreito: EMS de média dimensão com AOIs 2D/3D mais antigos e sem atualização de IA do fabricante; linhas que misturam várias marcas de AOI (um fornecedor de retrofit já suporta Omron, VI, Koh Young e Mirtec (Sohu sobre Hengtianyi)); nichos que os AOIs de SMT cobrem mal, como juntas de furo passante e soldadas à mão, revestimento conforme, cabos, conectores e FPCs, e aspeto final e parafusos; e pequenas fábricas que não podem pagar um AOI de RMB 230–280k.

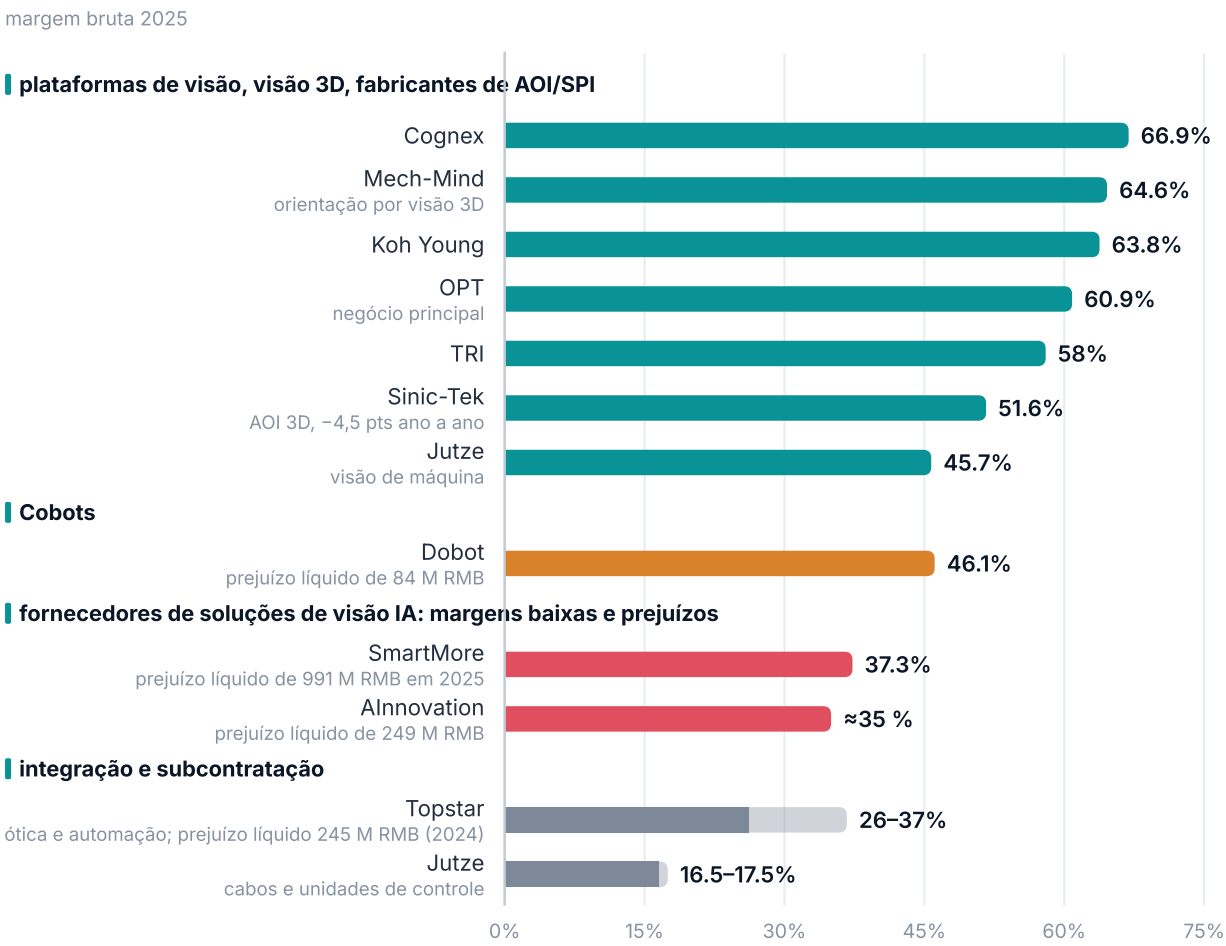
O comparável cotado mais próximo dá um aviso. A Weiyi Zhizao, que vende equipamento de inspeção com IA e robôs, viu a receita 3C cair 79.9% em termos homólogos entre janeiro e maio de 2026, reconheceu "concorrência intensificada e pressão de preços", adiou a sua listagem em Hong Kong a 25 de outubro de 2026 e divulgou 15 projetos terminados antes da conclusão, 12 deles projetos de robôs incorporados (prospeto da Weiyi Zhizao, aviso de adiamento) single source, primário . A AInnovation terminou 2025 com 29.8% das suas contas a receber provisionadas (relatório anual 2025 da AInnovation). A cunha de inspeção 3C está lotada, e o poder de preço por linha é fraco.

A margem está no hardware transformado em produto

Posição na cadeia	Empresas representativas e margem bruta de 2025	Lucro e cobranças
Plataformas de visão	Cognex 66.9%	Lucrativa; a desinvestir ≈$22M de receita de baixa margem
Equipamento AOI/SPI	Koh Young 63.8%; TRI 58%; negócio principal OPT 60.85%; AOI 3D Sinic-Tek 51.62% (−4.53 pontos homólogos); visão artificial Jutze 45.74%	Lucrativa; dias de recebíveis 139 (Jutze) a 247–274 (OPT)
Orientação por visão 3D	Mech-Mind 64.6%	—
Fornecedores de soluções de visão com IA	SmartMore 37.3%; AInnovation ≈35%	Com prejuízos: prejuízo líquido da SmartMore RMB 991M em 2025, AInnovation RMB 249M
Cobots	Dobot 46.1%	Prejuízo líquido da Dobot RMB 84M; receita da UR de $326M (2022) para ≈$232M (2025, previsão)
Integração de sistemas	Robótica e automação Topstar 26.2–36.7%	Prejuízo líquido RMB 245M (2024); dias de recebíveis 180–240
Hardware por contrato	Cabos e unidades de controlo Jutze 16.5–17.5%	—

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

FIG 26 **A margem está no hardware industrializado** ***Plataformas de visão e fabricantes de equipamentos obtêm 46–67 % de margem bruta; fornecedores de soluções de IA, 35–37 % e perdem dinheiro.***



Fonte: Relatórios anuais ou prospectos de 2025.

Um recém-chegado só mantém a margem ao transformar a inspeção num produto, equipamento ou software; a afinação projeto a projeto empurra as margens para cerca de 35% com prejuízos persistentes. A mecânica também funciona contra as pequenas empresas: a receita só é reconhecida após a aceitação no local, os adiantamentos são pequenos (Jutze, OPT), os pilotos são normalmente gratuitos (A3/Mindtrace), e os donos de marca especificam muitas vezes o equipamento de inspeção nas linhas dos seus fabricantes por contrato.

Como os compradores pagam: 30% na assinatura, 60% após aceitação, 10% retidos por um ano

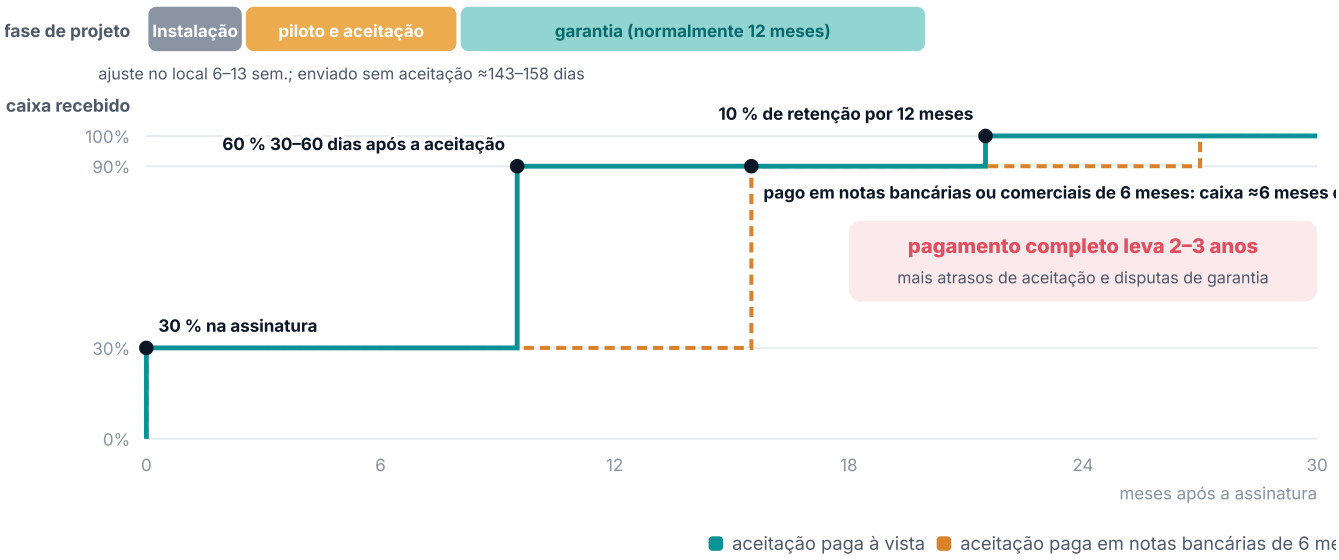
Os termos de pagamento que as fábricas de eletrónica do continente usam para equipamento de inspeção estão explicitados em 13 contratos de compradores nomeados divulgados na resposta ao inquérito de listagem da Sinic-Tek (resposta ao inquérito Sinic-Tek 8-1-1) single source: um documento primário, vários compradores independentes, 2020–2022 :

Comprador (ano)	Assinatura	Entrega	Aceitação	Retenção ou saldo
Uma unidade Hikvision (2021)	30% em 15 dias	—	65% em 30 dias da aceitação final	5% de retenção de garantia pagos 12 meses após a aceitação; parte paga em notas de aceitação bancária de 6+ meses
Nanchang Huaqin (2020–2021)	30% em 10 dias úteis	—	60% aos 60 dias após a aceitação	10% após uma garantia de um ano

Comprador (ano)	Assinatura	Entrega	Aceitação	Retenção ou saldo
Guangdong Yinuo (2021, AOI 3D)	30% em 10 dias úteis	—	60% em 30 dias	10% após três meses de funcionamento sem falhas
Xiamen Hongxin, Jingmen Hongyi (2022)	30% em 10 dias úteis	—	60% em 30 dias úteis da aceitação final, notas bancárias de 6 meses	10% após uma garantia de um ano
AKM Panyu, Suzhou (2021–2022)	30% em 7–10 dias úteis	—	60% um mês após aceitação e faturação, notas bancárias de 180 dias	10% após 360 dias
BYD Precision, Shenzhen, Huizhou (2020)	—	30% à chegada	70% após aceitação e faturação	Tudo em notas comerciais de 6 meses
Shenzhen H&T (2020)	—	—	90% após aceitação e faturação	10% após seis meses de uso sem falhas

Outros documentos primários mostram a mesma estrutura: os grandes projetos da Mech-Mind pagam 30% na assinatura, 30% antes da entrega, 30% após aceitação e 10% de retenção de garantia ao longo de 12–15 meses (prospeto da Mech-Mind); a GKG diz que a retenção no setor do equipamento é normalmente de 5–10% do contrato, paga um ano após a aceitação (resposta ao inquérito GKG) single sources, primário. A norma é, então, cerca de 30% na assinatura (20–30%, zero em algumas fábricas como a BYD), 60–65% após aceitação (pagos 30–60 dias depois, muitas vezes em notas bancárias ou comerciais de seis meses, o que atrasa o dinheiro utilizável em mais meio ano), e 5–10% de retenção por 3–12 meses (12 o mais comum) verified em vários contratos.

FIG 27 **Como os compradores pagam *30 % na assinatura, 60 % após a aceitação, 10 % retidos por um ano: o pagamento completo leva 2–3 anos.***



Fonte: Resposta de consulta da Sinic-Tek, Mech-Mind, GKG, JT e outros documentos; esquema em condições usuais, prazo estimado.

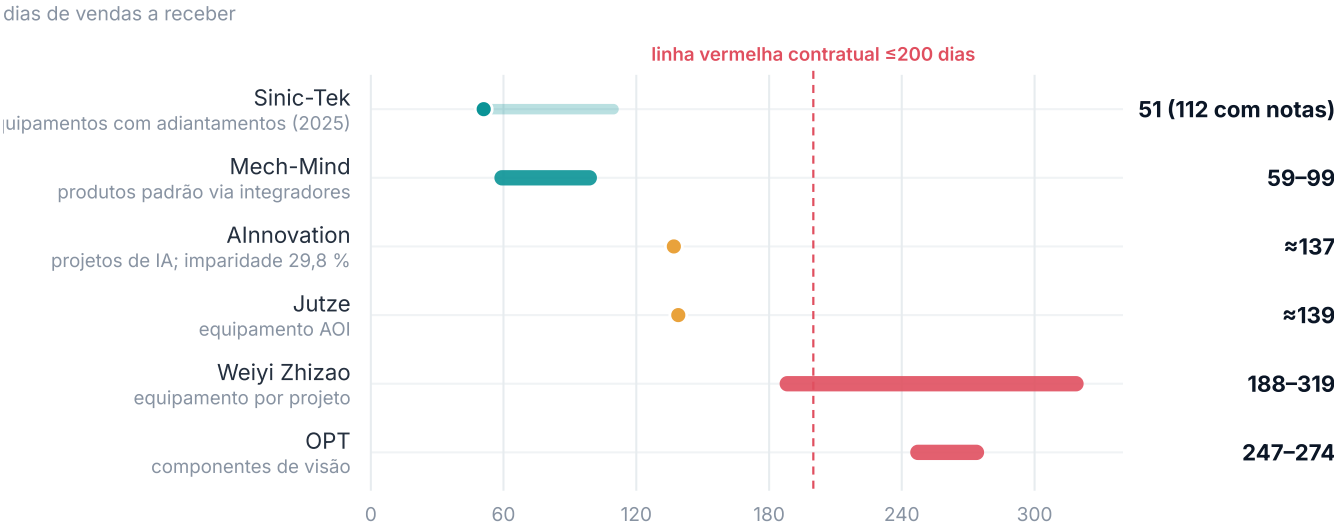
A aceitação e os ensaios são lentos e pagos pelo vendedor. Os bens enviados mas não aceites da Sinic-Tek demoraram cerca de 143–158 dias a rodar, porque os grandes clientes de EMS e de marca precisam de testes de integração de linha inteira e de um ensaio prolongado (resposta ao inquérito Sinic-Tek 8-1-2); os projetos de IA da Weiyi demoram 4–8 meses da avaliação à implantação (prospeto da Weiyi Zhizao) single sources, primário. Não há dados de conversão para pilotos de revisão com IA; a Sinic-Tek converteu apenas cerca de 18–19% das suas máquinas de

demonstração diretamente em vendas. Planeio um em três a um em cinco para pilotos gratuitos e cerca de 80% para pilotos pagos após a lista curta estimate.

Empresa	Modelo de negócio	Dias de vendas por cobrar
Sinic-Tek (2025)	Equipamento com depósitos	≈51 (≈112 incluindo notas)
Mech-Mind	Produtos padrão através de integradores	59–99
Jutze	Equipamento AOI	≈139
AInnovation (2025)	Projetos de IA	≈137 no fim do ano, com 29.8% de imparidade
Weiyi Zhizao	Equipamento por projeto	188–319
OPT	Componentes de visão	247–274

Fontes: os relatórios anuais e prospetos acima; contas a receber (incluindo ativos de contrato) ÷ receita × 365 primário + estimate.

FIG 28 **Dias de vendas a receber** ***Produtos padrão 51–99 dias; equipamentos por projeto 188–319 dias.***



Fonte: Documentos de empresa; contas a receber (incl. ativos contratuais) ÷ receita × 365.

As licenças de revisão com IA vendidas através de fabricantes de AOI ou integradores devem cobrar como os 60–100 dias da Mech-Mind; as vendas diretas por projeto como os 190–320 dias da Weiyi. Juntando como cronologia de caixa: piloto gratuito de um a três meses, qualificação num grande ODM até cerca de doze meses, construção, entrega e aceitação de três a sete meses; 30% chegam com a encomenda, 60% cerca de quatro a nove meses depois e os últimos 10% cerca de 16–21 meses depois. **Do primeiro piloto numa EMS grande até ao pagamento integral podem passar dois a três anos**; as EMS de média dimensão sem concursos formais podem encurtar a qualificação para semanas estimate. Os critérios de aceitação nos contratos divulgados cobrem apenas a integração de linha inteira, o débito (um projeto taiwanês exigia 100 unidades por hora) e meses de funcionamento sem falhas; nenhum especifica taxas de fuga ou de falso alarme (resposta ao inquérito Anda) single source, primário. Os critérios de aceitação por escrito para a revisão com IA têm de ser acordados antes do piloto.

Quatro padrões de falha

As falhas documentadas caem em quatro padrões: enviar antes de o produto ser fiável (Rethink (The Robot Report)); uma camada de software sobre o hardware de um líder que o ecossistema do líder substituiu depois (READY Robotics (The Robot Report)); ficar sem dinheiro no fundo de um ciclo (Dextrous Robotics (Robots & Startups)); e empresas de modelos a acabar em contratações de aquisição (Covariant (GeekWire)). As lições: transformar em

produto de forma restrita em vez de afinar projeto a projeto; deter capital de giro para receita baseada em aceitação e cobranças longas; fornecer serviço local para fábricas que funcionam 24 horas; e não escalar antes de o produto estar comprovado.

Timing: a inspeção não ganha nada em esperar; o hardware de braços fica mais barato

Os argumentos para esperar são fracos. O hardware não ficou mais barato em 2026: os fabricantes de robôs subiram os preços 5–15%, alguns produtos 25–50% (Sina Finance) single source, e os kits da NVIDIA subiram uma terceira vez (36Kr). Os VLAs de grau de produção não chegarão dentro da janela (IFR, Instrumental). Entretanto, cada substituição de AOI remove um alvo de retrofit e cada implantação de IA de OEM bloqueia os dados e o processo de uma linha; a capacidade do modelo está a comoditizar-se, e o que é escasso é o acesso a fábricas e a dados fabris etiquetados. O apoio local também tem janelas: as regras de 2026 de Shenzhen de "cidade pioneira em IA", com até RMB 2 milhões para projetos de demonstração de aplicação de IA e vales de modelo a cobrir 30% das compras, vigoram até **30 de junho de 2027** (Yicai, maio de 2026) termos single source.

O que ainda está a ficar mais barato depressa são os humanoides e as mãos hábeis (as mãos caíram de cerca de RMB 50,000 para menos de 10,000 em dois anos (Sina Finance)), e os cobots de entrada à medida que a mistura de produtos desce; isso argumenta por adiar a parte incorporada, não a parte de inspeção. **No geral: inspeção já, como produto restrito; células de robô já, através de parceiros e integração; IA incorporada mais tarde, com apenas uma estação de I&D hoje.**

IA incorporada

06 IA incorporada: primeiro muda como os robôs são ensinados, depois como se movem

O que está a mudar é a perceção e a programação; o movimento mantém-se determinista

Todos os líderes de braços robóticos comercializam agora "IA física", mas o que foi enviado cai em três camadas: perceção aprendida sem ensino (AI Masterless Vision da FANUC, a enviar em janeiro de 2027 (FANUC); seleção neural de agarramento da KUKA (KUKA)); modelos grandes que escrevem programas e planeiam tarefas (iiQWorks.Copilot da KUKA, que avisa que o código gerado por IA deve ser verificado num gémeo digital (KUKA); Yaskawa com Gemini Robotics-ER, onde "a IA generativa decide o que fazer" e o robô decide "como se mover" (Yaskawa)); e plataformas de computação e dados (AI Accelerator da UR com um Jetson Orin (Universal Robots)) vendor claims . Nenhum líder publica taxas de sucesso de produção para as suas funcionalidades de IA, e os líderes japoneses fazem parcerias com a Google, SoftBank, Microsoft e outros em vez de construir VLAs de fronteira (índice de notícias FANUC): a camada de política aprendida virá das empresas de modelos, os braços e controladores certificados dos fabricantes de braços.

FIG 29 **A pilha de IA corpórea** ***Modelos grandes planejam offline, políticas pequenas rodam na célula, movimento e segurança permanecem determinísticos.***

Camada	O que faz e até onde vai a evidência	Escala de tempo · quem a controla
Planejamento L4	**modelos de fronteira (offline)** escrever ou alterar programas de robô offline, trocas de série, explicar falhas reescreve programas ABB RAPID a 77–99,7 %, implantado após simulação laço de controle a só 0,2–0,4 Hz: fica fora do tempo real	**segundos a minutos** gigantes de modelos
Habilidades / políti L3	**ACT, políticas de difusão, VLAs pequenos** postos estruturados com gabarito 90–99,4 % chips edge: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**blocos de ação 100–300 ms** fabricantes de chips · fabrican
controle de movime L2	**Controlador de robô** trajetórias, servo, controle de força, detecção de colisão código determinístico; as políticas aprendidas só definem metas	**milissegundos (tempo real exige >83 Hz)** fabricantes de robôs
Segurança L1	**PLC de segurança, scanners, limites de força e velocidade** certificado conforme ISO 10218-1/-2:2025 e regras locais Sem ML em funções de segurança: avaliação por organismo notificado da UE a partir de 20 jan. 2027	**tempo real estrito** robô certificado + integrador

As pequenas empresas ganham na integração e no hardware de detecção: colocar percepção aprendida e políticas pequenas em células robótic

Fonte: Síntese do autor (capítulos 3 e 6); ISO 10218:2025; Regulamento de Máquinas da UE.

A que distância do trabalho não estruturado, flexível e de pequenos lotes: estações estruturadas em 1–3 anos, montagem flexível 5+

Os números reais de fábrica começaram a aparecer, mas são relatados por fornecedores, estreitos e de estações estruturadas. Os oito humanoides da AgiBot na linha de teste final de tablets da Longcheer registaram 64,828 tarefas a 99.99% em mais de 64 horas (AgiBot, Yicai Global) [fornecedor e cliente], mas isto é trabalho de alimentação

que um braço convencional com visão consegue fazer. A UBTech diz que o seu Walker S2 trabalha a 30–50% da eficiência humana (Irish Times/FT) single source. Os 11 meses do Figure 02 na BMW Spartanburg produziram números de volume (1,250+ horas, 90,000+ peças), enquanto os seus números de ciclo, sucesso e intervenção são metas, não resultados (Figure, BMW Group) verified, apenas numeradores. A Nucleus diz que cerca de 40% do seu vídeo de fábrica é teleoperado (Interesting Engineering) vendor claim, e uma investigação da Reuters citou cerca de 20% da produção de um trabalhador para alguns humanoides chineses (Notebookcheck) single source. Nenhuma implantação publica taxa de sucesso, tempo de ciclo e taxa de intervenção em conjunto, e nenhuma foi auditada, mas são exatamente esses três que um cálculo de retorno precisa.

Implantação	Taxa de sucesso	Takt	Intervenções ou teleoperação	Tempo de atividade ou intervalo de falha	Execução de volume
Figure 02 @ BMW Spartanburg	Apenas meta (>99% por turno)	Apenas meta (84 s)	Apenas meta (0 por turno)	Não divulgado (topo do antebraço como ponto de falha)	1,250 h
Agility Digit @ GXO	Não divulgado	Não divulgado	Não divulgado	Não divulgado	100,000 caixas
Apptronik Apollo @ Mercedes-Benz	Não divulgado	Não divulgado	Não divulgado	Não divulgado	Não divulgado
Agile ONE @ Audi	Não divulgado	Não divulgado	Não divulgado	Não divulgado	Não divulgado
Nucleus (fábrica não nomeada)	Não divulgado	Não divulgado	≈40% teleoperado (método não divulgado)	Não divulgado	≈2 h de vídeo
AgiBot G2 @ Longcheer	99.99% de 64,828 tarefas	18–20 s por passo	Não divulgado	Uma falha de comunicação, ≈2 min para recuperar	64+ h
Célula híbrida Neuromeka	322/324	12.8% mais lenta do que uma pessoa	Não divulgado	Não divulgado	324 ciclos
Physical Intelligence π0.6 num local de cliente	—	—	96.4% autónomo ao longo de um turno	Não divulgado	Um turno

Fontes: este capítulo e o capítulo 3 na maioria vendor claims.

FIG 30 **O que as implantações robóticas em fábrica realmente divulgam** ***A maioria divulga apenas o volume; nenhuma divulga juntas taxa de sucesso, takt e taxa de intervenção.***

	taxa de sucesso	Takt	Intervenções / teleoperação	Disponibilidade / falhas	Volume
Figure 02 @ BMW	só objetivo >99 %/turno	só objetivo 84 s	só objetivo 0/turno	n/d	1.250 h
Agility Digit @ GXO	n/d	n/d	n/d	n/d	100.000 contentores
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
Nucleus (fábrica não nomeada)	n/d	n/d	≈40 % teleoperado*	n/d	≈2 h de vídeo
AgiBot G2 @ Longcheer	99.99%	18–20 s por etapa	n/d	uma falha, 2 min	64+ h
célula híbrida Neuromeka	322/324	12,8 % mais lento	n/d	n/d	324 ciclos
π0.6 @ site do cliente	—	—	96,4 % autônomo	n/d	1 turno

* método não divulgado

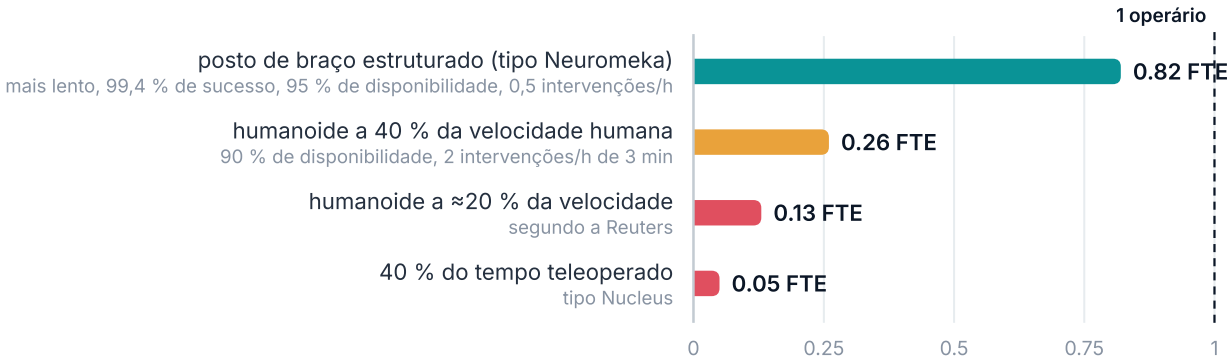
Divulgado (sobretudo alegações do fornecedor) só metas n/d

Fonte: Material público de BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence e outros.

A teleoperação importa mais para a economia do que a taxa de sucesso. Pessoas líquidas substituídas por robô por turno ≈ (ciclo humano ÷ ciclo do robô) × tempo de atividade × sucesso, menos o tempo de intervenção e a teleoperação estimate, minha fórmula. Uma estação estruturada estilo Neuromeka (12.8% mais lenta, 99.4% de sucesso, 95% de tempo de atividade, 0.5 intervenções de 2 minutos por hora) substitui cerca de 0.82 pessoas; um humanoide a 40% da velocidade humana, 90% de tempo de atividade e duas intervenções de 3 minutos por hora cerca de 0.26; aos 20% da Reuters, cerca de 0.13; e se, como a Nucleus, um operador o teleoperar 40% do tempo, cerca de 0.05, quase nada. A cerca de RMB 600k por humanoide e RMB 85k por pessoa por ano no continente, os últimos três pagam-se ao longo de décadas; só a estação de braço estruturada tem um retorno realista (capítulo 4) estimate.

FIG 31 **Mão de obra líquida substituída por robô** ***Uma célula de braço estruturado substitui ≈0,82 de trabalhador; um humanoide teleoperado 40 % do tempo, ≈0,05.***

mão de obra líquida substituída por robô e turno



Fonte: Estimativas baseadas em fórmulas do autor.

Fora das estações estruturadas, os modelos gerais ainda pontuam 40–90% em tarefas novas: o Gemini Robotics 2 atinge 74.2% em pick-and-place de laboratório, 78.9% em kitting de ferramentas e 44% em atar um saco (eesel/DeepMind); no RoboChallenge Table30, 30 tarefas com robôs reais, o π0.5 pontua 43.7% e o melhor modelo aberto 50.3% (RoboChallenge, QbitAI).

As previsões externas divergem, mas nenhuma apoia o uso generalizado em dois anos. A IFR diz que a IA física geral "ainda não é de grau industrial" e que a maioria dos humanoides "ainda são protótipos ou pilotos" (IFR); a Inte-

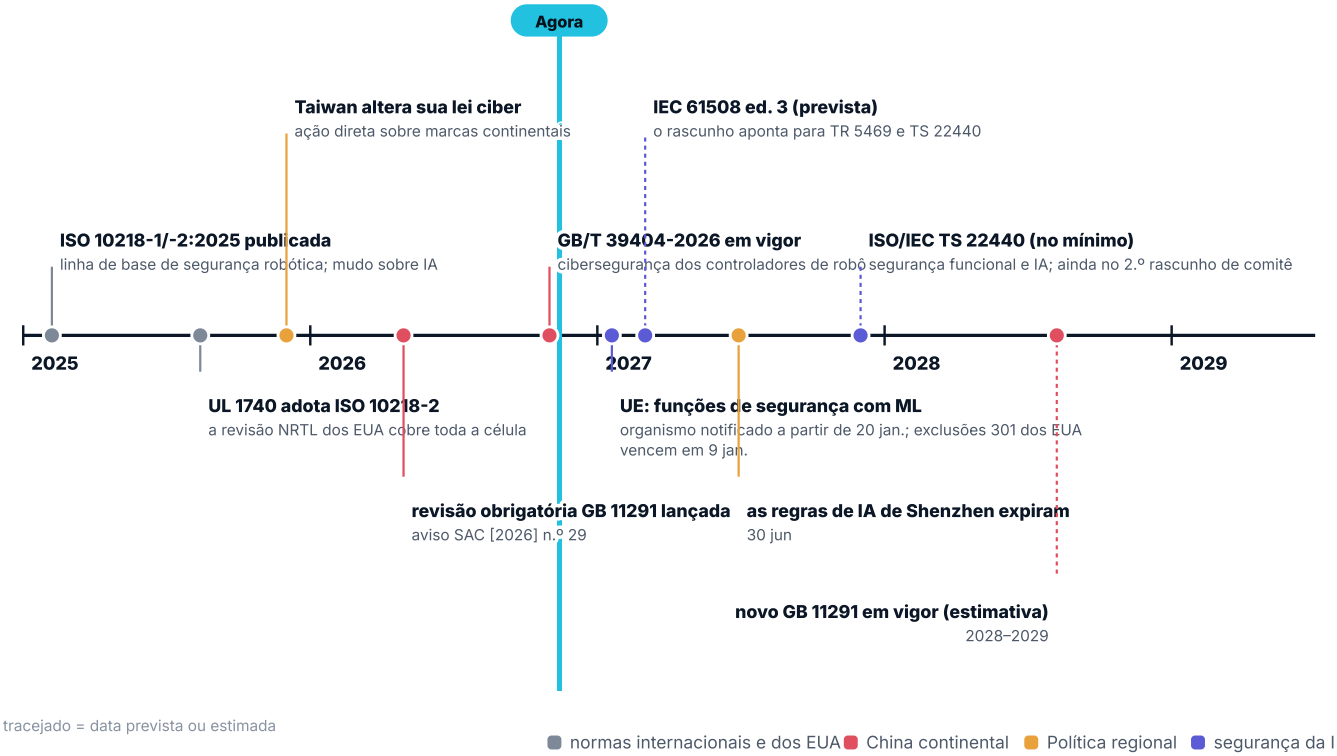
ract Analysis estima que apenas cerca de 10% dos humanoides produzidos em 2025 entraram em uso real, com o ponto de viragem comercial "depois de 2032" (automation.com); a UBS vê um "momento EV" dos humanoides como improvável antes de 2030 (Yicai Global); os informadores chineses estimam mais dois a três anos antes do uso à escala fabril (Gasgoo) single sources. O meu julgamento estimate: em fábricas de eletrónica e automóveis bem dotadas, com o fornecedor no local, as **tarefas de estação estruturada** (alimentação, triagem OK/NG, inserção fixada) chegarão a ≥99% perto da velocidade humana em cerca de 1–3 anos; o **kitting de peças mistas** com VLAs gerais em cerca de 3–5; as **cablagens de encaminhamento livre e a montagem flexível de vários passos** em mais de 5. O retorno de humanoides não funciona hoje: um Walker S2 a cerca de RMB 600k (10jqka) single source, a 30–50% da eficiência humana em dois turnos, equivale a cerca de 0.6–1.0 trabalhadores, um retorno simples de 7–12 anos aos custos de mão de obra do continente antes de integração, manutenção e paragens estimate.

Onde está bloqueado

Elo	Evidência em 2026
Dados	O setor estima que são precisas pelo menos 10 milhões de horas de dados reais de alta qualidade e apenas cerca de 500,000 tinham sido recolhidas até ao início de 2026 (Gasgoo); a Reuters refere cerca de 500,000 horas contra cerca de 100 milhões necessárias (Notebookcheck); um teleoperador novato precisa de cerca de 300 tentativas por movimento utilizável; as fábricas não partilham dados de processo; as políticas treinadas em simulação caem acentuadamente em robôs reais
Fiabilidade	Estações estruturadas 98–99.99% (vendor claims, falha não definida); modelos gerais 40–90% em tarefas novas; conectores nunca vistos 29–67%
Velocidade	Humanoides a 30–50% da velocidade humana (UBTech), alguns a cerca de 20% (Reuters); agentes de fronteira cerca de 280 s por tarefa quase estática com robô real; malhas de controlo LLM a 0.2–0.4 Hz contra os cerca de 83 Hz necessários (Anthropic)
Custo	Walker S2 com média de cerca de RMB 600k; retorno de 7–12 anos estimate
Certificação de segurança	A ISO 10218-1/-2:2025 torna a segurança funcional mais explícita e absorve os requisitos colaborativos, mas nada diz sobre IA (The Robot Report); a ISO/IEC TS 22440 (segurança funcional e IA) era ainda um segundo rascunho de comité em novembro de 2026 (ISS, ISO Open Data); a orientação IEC 61508-3 desaconselha IA acima de SIL 1 (Analog Devices); sob o Regulamento de Máquinas da UE, as funções de segurança com aprendizagem automática precisam de avaliação por organismo notificado a partir de 20 de janeiro de 2027 (EUR-Lex)
Mãos e ferramentas	As mãos hábeis gerem cerca de 30 N (≈3 kg) por dedo contra cerca de 10 kg de uma mão humana (Gasgoo); as mudanças de ferramenta continuam manuais
Integração	Leituras erradas de sensores, desencontros de protocolo e ligações MES no local; as taxas de intervenção quase nunca são divulgadas

A implicação de segurança é direta: as normas nem proíbem as políticas aprendidas nem as reconhecem. A ISO 10218-1/-2:2025 nada diz sobre IA; a ISO/IEC TS 22440 (segurança funcional e IA) ainda está em rascunho de comité e é improvável antes do final de 2027 ou 2028 estimate; a ISO/IEC TR 5469:2024 é apenas orientação, recomendando alternativas não IA, restrições de supervisão e votação redundante; a IEC 61508 edição 3 é esperada no início de 2027 (ISO Open Data, BSI, 61508 Association) verified. Nenhuma norma publicada oferece ainda uma via aceite para certificar a perceção por aprendizagem automática como função de segurança. As regras regionais também se movem: a China está a rever a GB 11291 (em vigor cerca de 2028–2029) (plano de revisão 20262543); a UL 1740 trouxe a ISO 10218-2 para a análise NRTL dos EUA desde agosto de 2025; a UE exige avaliação por organismo notificado das funções de segurança com aprendizagem automática a partir de 20 de janeiro de 2027; e nenhum humanoide está certificado para uso fabril sem vedações enquanto a ISO 25785-1 ainda está em rascunho (i-SCOOP). A arquitetura defensável: a aprendizagem automática toma apenas decisões de processo ("que peça", "esta placa está boa"), enquanto uma camada de segurança determinista certificada (normalmente PL d / Categoria 3) e um PLC de segurança decidem a velocidade, a força e as paragens. Trate qualquer alegação de "IA de câmara para segurança sem vedações" como não comprovada até um certificado PL d ou SIL 2 nomear o componente de aprendizagem automática.

FIG 32 **Cronologia de normas e regulamentação** ***Uma via de certificação para IA dentro de funções de segurança não se assentará antes de 2027–2029.***



Fonte: ISO, IEC, UL, EUR-Lex, SAC e bases de dados regulatórias nacionais.

Que camada é a oportunidade: os modelos pertencem aos gigantes, as caixas de borda aos fabricantes de chips, as pequenas empresas à integração e ao hardware de deteção

A camada de modelos precisa de capital e dados e pertence aos gigantes (a AgiBot treinou o GO-2 em "dezenas de milhares de horas" (iFeng Tech); os modelos mais fortes são fechados ou só de nuvem). A camada de caixas de borda pertence aos fabricantes de chips: o Jetson Thor é o padrão de facto, e no continente as placas RDK da D-Robotics (D-Robotics) e os coprocessadores Rockchip (CSDN) competem, na maioria sem resultados de VLA publicados.

As pequenas empresas têm duas aberturas. Uma é o hardware de deteção que compra fiabilidade diretamente: o toque eleva o sucesso de inserção do ContactDP de 58.7% para 97.8%, e os sensores de força/binário e as câmaras de pulso fazem trabalho semelhante; os kits de recolha de dados também se vendem (10jqka), embora as mãos, os sensores e os braços estejam todos a cair de preço. A outra é a camada de integração entre os fabricantes de braços e as empresas de modelos, detendo os dados de tarefa de cada estação, o envelope de segurança e a ligação MES, que os fabricantes de braços não vendem hoje como produto. Os concorrentes caem em quatro grupos: líderes de braços (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot e outros), empresas de visão mais braço (Mech-Mind, Hikrobot, SmartMore), empresas de IA incorporada (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) e entrantes de chips automóveis.

Aceitação do comprador: nenhum número sem denominador

Como nenhum fornecedor publica taxas de intervenção, intervalos de falha ou contagens de sucesso de longo prazo, estes pertencem aos termos de aceitação do contrato. Estatisticamente, 3,000 ciclos de produção consecutivos com zero falhas provam uma taxa de falha de, no máximo, 0.1% com 95% de confiança; 10 falhas em 3,000 provam apenas ≤0.57%; provar ≤1%, ≤0.5%, ≤0.1% e ≤0.01% com zero falhas demora 299, 598, 2,995 e 29,956 ciclos estimate, binomial exato . Por esse padrão, os 322/324 da Neuromeka têm um intervalo de 95% bilateral de 97.8–

99.9% e não provam ≥99%; as cerca de 6 falhas da AgiBot em 64,828 tarefas implicam 99.98–100%, mas apenas se "falha" for definida como o comprador a define estimate. Estes são os requisitos de aceitação que eu colocaria a qualquer célula de braço robótico ou incorporada; os limiares baseiam-se nas próprias metas da Figure na BMW e nos casos da Neuromeka e da AgiBot, e são julgamento, não uma norma estimate:

Item	Exigir	Limiar de aprovação sugerido
Definições	Definições assinadas de ciclo, sucesso, falha (sucesso parcial, re-agarramento, queda, fuga de qualidade), intervenção (toque, teleoperação remota, reset, paragem de emergência) e como a autonomia é calculada	As definições vêm antes de qualquer métrica; nenhuma percentagem sem fórmula
Intervenções e teleoperação	Por hora-robô, por tipo e duração; tempo médio entre intervenções; minutos de teleoperação como quota do tempo de execução; operadores por robô	Estações estruturadas ≤0.5 intervenções por hora-robô (≥2 h de intervalo); zero teleoperação durante a aceitação
Sucesso de longo prazo	≥3,000 ciclos de produção consecutivos em ≥3 turnos, ≥2 números de peça e variação normal de iluminação e abastecimento; falhas classificadas; limite inferior exato	Limite inferior ≥99.5% para estações inline que param a linha, ≥99.0% offline; fugas de qualidade reportadas separadamente
Tempo de ciclo	P50/P90/P99 por peça incluindo recuperação, contra ≥30 ciclos humanos cronometrados na mesma estação	≤1.15 × P50 humano, ou dentro do takt da linha
Tempo de atividade e falhas	Horas de execução, registo de paragens, falhas por componente, tempo de troca de peças	≥95% de tempo de atividade em ≥2 semanas; MTTR ≤20 min com peças no local
Recuperação	Injeção de falhas roteirizada: quedas, oclusão, mudança de iluminação, paragem do transportador, perda de MES ou rede, perda de energia, paragem de emergência	≥90% de recuperação autónoma, de volta à produção em 2 min
Troca	Horas, demonstrações, competência do operador e custo de recolha de dados por novo número de peça ou fixação	Estações estruturadas: ≤1 turno e ≤100 demos por novo número de peça
Segurança	Braço e célula para ISO 10218-1/-2:2025, avaliação de risco da célula, regras locais (GB 11291 no continente; CNS 14490 mais avaliação escrita em Taiwan; ANSI/A3 R15.06-2025 e UL 1740 nos EUA)	Nenhuma função de segurança implementada por aprendizagem automática; uma camada de segurança certificada (normalmente PL d / Cat 3) limita a velocidade e a força; sem robôs com pernas sem vedações até a ISO 25785-1 ser publicada e avaliada
Computação e comportamento offline	Hardware de inferência, latência P99, comportamento sem rede, versionamento de modelo, revalidação após cada atualização	A inferência crítica para o takt não depende da nuvem; re-aceitação total após qualquer mudança de modelo, quantização ou motor
Evidência	Vídeo não editado e registos brutos para toda a janela de aceitação; testemunhados pela equipa de qualidade do cliente ou por terceiros	Sem reels de destaques nem resumos só do fornecedor
Termos comerciais	Marcos de pagamento ligados a métricas de aceitação; penalidades de disponibilidade quando faturado por hora-robô; regras de propriedade de dados e acesso remoto	Pagamento final e retenção libertados apenas após a aceitação de 3,000 ciclos e ≥4 semanas de funcionamento estável
Economia	Pessoas líquidas substituídas = rácio de velocidade × tempo de atividade × sucesso − tempo de intervenção − quota de teleoperação	Rejeitar qualquer proposta cuja substituição líquida não suporte o retorno reivindicado

Fontes: (Figure, Neuromeka, Yicai Global, i-SCOOP); os limiares são as minhas sugestões.

Questões em aberto

07 O que o registo público não consegue responder e como um piloto resolve isso

As questões que as fontes públicas não conseguem responder caem em cinco grupos: por quanto a revisão com IA realmente se vende e o que os compradores pagarão; como os fabricantes de AOI agrupam a IA e atualizam máquinas antigas; quão antigos são os AOIs instalados e quantos já executam IA; quantas pessoas e horas a revisão realmente demora por linha; e como os pilotos se convertem e como os compradores pagam. Em cima disso vem o desempenho real dos modelos mais recentes, das compilações quantizadas e dos braços robóticos sob as condições de uma dada fábrica. Cada uma pode ser transformada num teste com um limiar claro. Os limiares abaixo são regras de decisão que derivei da evidência neste relatório estimate ; recalcule-os assim que os primeiros dados de fábrica chegarem.

Questão em aberto	Como resolvê-la	Como é um bom resultado
Preço e disponibilidade para pagar pela revisão com IA	Um pedido de cotação padrão (por exemplo, quatro linhas com AOIs 2D antigos e 3D mais recentes misturados, dois turnos, cerca de 2,000 placas por turno, 12 classes de defeito, apenas no local, integração MES ou IPC-CFX) enviado a vários fornecedores de terceiros; um inquérito de sensibilidade ao preço de quatro perguntas a 20–30 fábricas de EMS	A maioria das cotações dentro de RMB 40–120k de pagamento único ou 20–40k por ano para quatro linhas; pelo menos metade das fábricas aceita RMB 10–30k por linha com retorno em 12 meses
Agrupamento de IA dos fabricantes de AOI e atualizações para máquinas antigas	Entrevistas estruturadas curtas com cerca de uma dúzia de fabricantes e distribuidores de AOI	Se a revisão com IA é padrão, opcional ou ausente nos modelos 2025–26; se as máquinas anteriores a 2020 podem ser atualizadas e a que preço
Idade dos AOIs instalados e quota com IA	Levantamento no local de cerca de 30 fábricas de EMS, estratificado por dimensão e região, com listas de equipamento revistas sob NDA	Idade mediana dos AOIs ≥5 anos e ≥60% sem IA significa que o parque de retrofit é real
Mão de obra e horas de revisão	Entrevistas e 2–4 semanas de registos de alarme de AOI, revisão, reparação e turnos em duas ou mais fábricas, com amostragem de trabalho	≥1 revisor dedicado por linha por turno e ≥3 horas de revisão por linha por turno, com pelo menos 0.5 FTE por linha que a fábrica esteja disposta a redistribuir
Desempenho da própria revisão com IA	Conjunto dourado mais modo sombra numa linha real (capítulo 3): acordo de dados na semana 0, linha de base e etiquetagem nas semanas 1–2, execução em sombra nas semanas 2–4, execução assistida opcional nas semanas 4–8	Falsos alarmes ou horas de revisão cortados em ≥50–60%; limites de fuga por classe não piores do que a linha de base humana; zero omissões em classes críticas; latência p95 ≤½ takt; falha aberta automática para pessoas
Qualidade de etiqueta de professor dos modelos de fronteira	Executar dois ou três candidatos no mesmo conjunto dourado, três redações de prompt cada, uma boa imagem de referência, recortes de componente único ≥42–48 px	Aceitação automática apenas para classes com zero omissões em ≥299 defeitos e concordância de nível humano; caso contrário, apenas pré-etiquetagem e triagem

Questão em aberto	Como resolvê-la	Como é um bom resultado
Compilações quantizadas	Alterar apenas a precisão; fixar hashes de pesos, modelo, pré-processamento, descodificação, motor e tamanho de lote; desativar ou reiniciar a cache de prefixo; verificar que a compilação de referência reproduz os resultados do fornecedor; pontuar pares no conjunto dourado e na fatia de deriva	Sem novas omissões em classes críticas; ≤1 nova omissão em 1,000 defeitos; resolução abaixo ≤2 pontos; ≥99.5% de decisões idênticas em três execuções; caso contrário, recuar para FP8/INT8
Hardware de linha, limpeza e ESD	≥2 unidades de cada caixa candidata testadas sob stress ao dobro da taxa de alarme esperada; imersão de 72 horas com o fluxo de ar real do armário; verificação ESD pelo coordenador do cliente; teste ISO 14644-14 para clientes em ISO 7 e mais limpa	p95 ≤½ takt; sem bloqueios nem limitação ao longo de 72 horas; assinatura ESD; relatório de partículas na classe certa
Termos de pagamento, regras de piloto e conversão	Cláusulas de pagamento de contratos recentes de equipamento de inspeção nas fábricas inquiridas; entrevistas sobre regras de piloto, critérios de aceitação e tempo de qualificação	Cerca de 30/60/10 com dinheiro em 60 dias da aceitação; pilotos ≤8–12 semanas com um dono de assinatura nomeado; ≥1 em 3 ensaios a converter; um acordo-quadro cobrindo ≥5–10 linhas no primeiro ano
Dados, TI e responsabilidade	Entrevistas e revisão de contrato com aconselhamento jurídico	Uma caixa local pode ligar-se e aceder aos dados; responsabilidade limitada a defeitos de software
Conformidade regional antes de ir para o estrangeiro	Taiwan: estatuto escrito do cliente sob a Lei de Gestão da Cibersegurança, autoavaliação E187; EUA: decisão aduaneira sobre como uma célula é classificada, cotações NRTL, verificação ITAR; Sudeste Asiático: ligação local BOI e elegibilidade para o subsídio malásio, casa-mãe do cliente; continente: âmbito CCC para PCs e fontes de alimentação agrupados	O cliente não é um órgão controlado, e o retorno fica ≤2 anos com direitos e certificação incluídos
Aceitação de uma célula de robô ou incorporada	A tabela de aceitação de 12 itens do capítulo 6: definições assinadas, ≥3,000 ciclos de produção consecutivos em ≥3 turnos e ≥2 números de peça, injeção de falhas roteirizada, ≥30 ciclos humanos cronometrados, testemunhados	Limite inferior ≥99.5% inline (≥99.0% offline); ciclo ≤1.15 × humano; ≤0.5 intervenções por hora e zero teleoperação; ≥95% de tempo de atividade; MTTR ≤20 min; ≥90% de recuperação
Políticas ACT e VLA em chips de borda chineses	Quatro semanas: portar e micro-benchmark (≥1,000 inferências, distribuição de latência, potência, imersão de 60 minutos); simulação em malha fechada com latência medida injetada; A/B em robô real contra uma GPU; resistência de 8–24 horas (≥3,000 ciclos)	Blocos ACT p95 ≤100 ms, VLA ≤300 ms; sucesso em robô real dentro de 3 pontos da GPU e ciclo ≤1.2 × humano; caso contrário, manter a política em Jetson ou GPU
Preços de cobot, AMR e integração	Cotações modelo a modelo de vários distribuidores de cobots, e cotações de célula completa de dois ou três integradores para carga de fixação de teste (uma estação, dois turnos)	Célula completa ≤RMB 200k e retorno ≤1.5 anos em dois turnos em Shenzhen
Contagens de linhas SMT em Taiwan, Sudeste Asiático e EUA	Dados da IPC (antiga New Venture Research) ou base instalada de distribuidores de equipamento locais	Linhas aproveitáveis suficientes por região para justificar engenharia de aplicação local

De todas estas, quatro decidem os números mais sensíveis deste relatório: o preço que os compradores pagarão, a mão de obra que a revisão com IA realmente liberta, se a revisão com IA realmente não omite mais do que as pessoas, e a rapidez com que o dinheiro volta. Vale a pena resolvê-las antes de qualquer compromisso significativo, por uma fábrica que considere uma compra ou por um fornecedor que considere um produto.

Uma sequência de piloto que funciona para ambos os lados

Para uma fábrica, e para um fornecedor que queira uma referência credível, a ordem é a mesma:

1. **Semanas 0–2: acordar dados e definições.** Assinar o acordo de dados, definir as classes de defeito e o que conta como fuga, e registar uma linha de base: falsos alarmes por placa, segundos de revisão por chamada, fugas encontradas a jusante.
2. **Semanas 1–4: construir o conjunto dourado** a partir do arquivo de AOI e das placas NG retidas, etiquetado às cegas e arbitrado (capítulo 3).
3. **Semanas 2–4 (ou mais): modo sombra.** O modelo regista as suas decisões enquanto as pessoas decidem como antes; comparar por classe.
4. **Semanas 4–8: execução assistida em turnos aleatórios,** com o modelo a resolver apenas as classes que passaram o portão e tudo o resto a ir para as pessoas.
5. **Aceitação, por escrito, antes de escalar:** os resultados por classe, a latência, o comportamento de falha aberta e quem detém modelos e etiquetas. Escalar linha a linha e revalidar em cada novo produto, versão de modelo, quantização ou mudança de motor.

Limitações deste relatório

Isto é investigação documental. Não conduzi entrevistas primárias com fábricas para esta edição; muitos dados de linha vêm de fornecedores e não foram auditados de forma independente. Vários sites governamentais, bases de dados de concursos e plataformas de recrutamento não puderam ser acedidos, pelo que as taxas de subsídio de atualização de Shenzhen em 2026, a orientação salarial do Delta do Rio Yangtzé, os preços de transação da revisão com IA, os revisores por linha e as taxas de conversão de pilotos são estimativas ou analogias; cada um aparece acima como uma questão em aberto. As taxas de câmbio são a paridade central do Banco Popular da China de 9 de novembro de 2026 ($1 = RMB 6.7188) e, para Taiwan, a fixação do SMBC no mesmo dia ($1 = NT$32.12); o dongue vietnamita usa dados de maio de 2026. Todos os limiares são regras de decisão derivadas da evidência citada, não normas.

Sobre o autor

Tan Shuai

Tan Shuai é um tecnólogo e fundador com mais de vinte anos de trabalho em chips, hardware e software de sistema, e 20 patentes concedidas. O seu trabalho levou-o por Silicon Valley, Nova Iorque, Chicago, Taiwan, a Grande Área da Baía (Shenzhen e Hong Kong), Vietname e Malásia, com um longo foco na indústria avançada e nas cadeias de abastecimento. Hoje trabalha como investigador e programador de IA, concentrado em como a IA é realmente usada no chão de fábrica.

Acolhe conversas com pessoas que trabalham em IA na indústria, quer estejam a construir uma equipa, a procurar um parceiro ou a comparar notas: i@tanshuai.com · tanshuai.com

Notas e aviso legal

Este relatório é investigação documental independente baseada em documentos públicos disponíveis até 11 de novembro de 2026. É fornecido apenas para informação e não é aconselhamento de investimento, jurídico, fiscal ou de engenharia. Os nomes de produtos, os preços, os lançamentos de modelos, as classificações de exportação, os direitos aduaneiros e os regulamentos mudam rapidamente; verifique qualquer coisa em que se apoie contra a fonte primária. Os valores marcados como vendor claim ou estimate devem ser lidos como tal. Os nomes de empresas e produtos pertencem aos seus donos e são usados apenas para identificação; nenhuma empresa reviu ou patrocinou este relatório.

© 2026 Tan Shuai. Pode partilhar este relatório inalterado com atribuição.

FABRICO DE ELETRÓNICA · INSPEÇÃO COM IA × ROBÓTICA

Inspeção visual com IA e braços robóticos em fábricas de eletrónica: avaliação 2026

Cada número-chave traz a sua fonte e uma etiqueta de confiança: verified, single source, vendor claim ou estimate. O que o registo público não consegue responder está reunido no capítulo 7, com a forma como um piloto resolve cada questão.

Tan Shuai
tanshuai.comi@tanshuai.com

EDIÇÃO PÚBLICA · DADOS ATÉ 11 DE NOVEMBRO DE 2026

