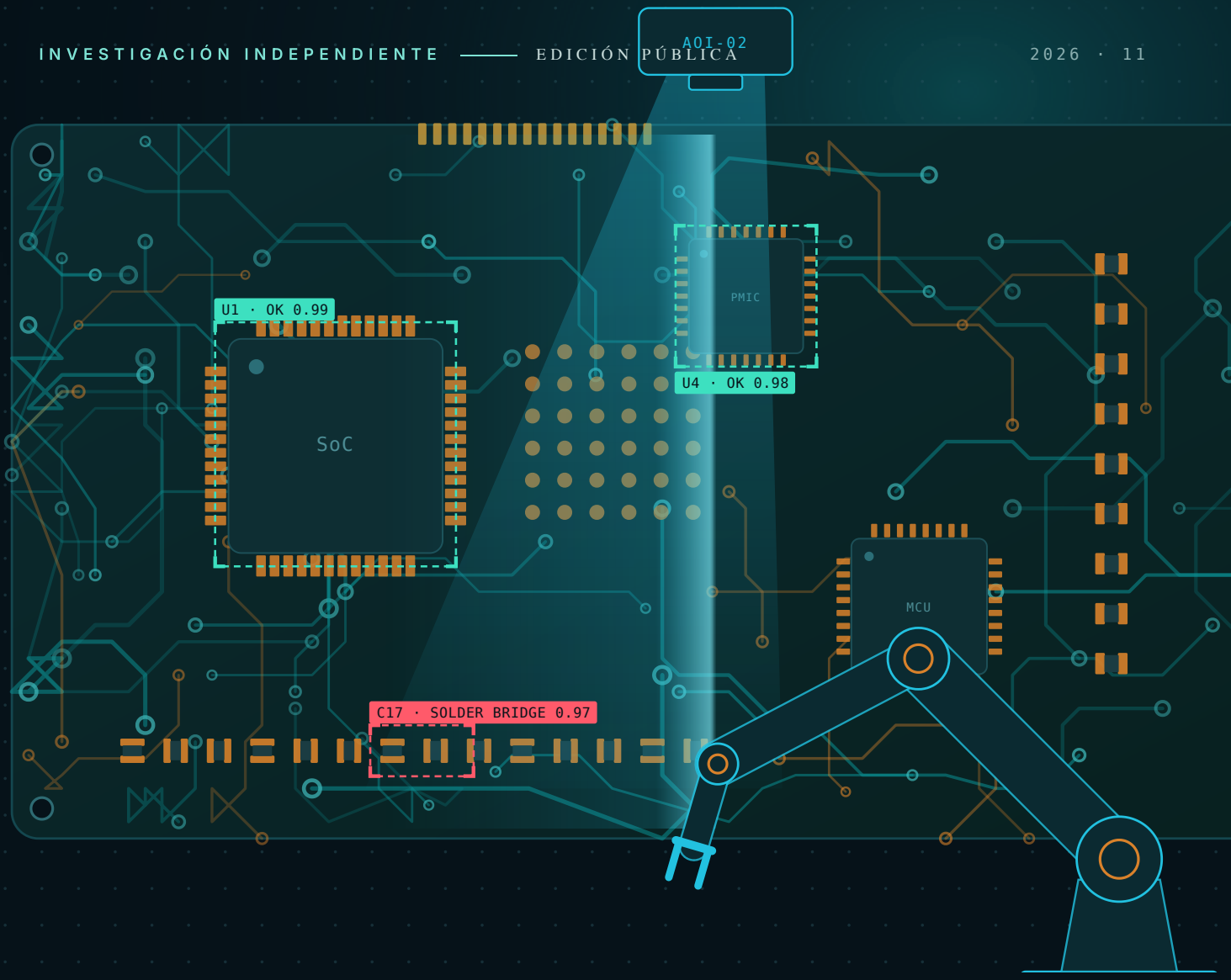


FABRICACIÓN ELECTRÓNICA · INSPECCIÓN CON IA × ROBÓTICA

Inspección visual con IA y brazos robóticos en fábricas de electrónica

Evaluación 2026

Configuraciones habituales · Viabilidad regional · Despliegue en local · Modelos profesor-alumno
Retorno de la inversión · Competencia · Lo que solo un proyecto piloto puede responder

HALLAZGO CLAVE

La revisión con IA merece la pena, de forma acotada. Los brazos robóticos corporeizados pueden esperar.

Tan Shuai
tanshuai.comi@tanshuai.com

Datos al 11 de noviembre de 2026
China continental · Taiwán · Sudeste Asiático · Estados Unidos

CONTENIDO

Contenido



◆ **RESUMEN EJECUTIVO** — Merece la pena, pero de forma acotada, ahora y con socios — 5

01 **La línea hoy** — Qué ejecutan hoy las fábricas de electrónica y qué es técnicamente viable — 9

02 **En local** — Ejecutar la IA dentro de la planta: salas limpias, aislamiento de red, hardware y cuantización — 17

03 **Profesor y alumno** — Profesor y alumno: los modelos de frontera pueden pre-etiquetar datos de inspección; para los brazos robóticos, los verdaderos profesores siguen siendo las personas — 26

04 **Costes y retorno** — Costes y retorno: la revisión con IA se amortiza en meses, la alimentación de máquinas en años, los brazos de retirada nunca — 35

05 **Mercado y competencia** — Mercado, competencia y cómo pagan los compradores — 43

06 **IA corporeizada** — IA corporeizada: primero cambia cómo se enseña a los robots, después cómo se mueven — 52

07 **Preguntas abiertas** — Lo que el registro público no puede responder, y cómo lo resuelve un piloto — 58

◆ **Sobre el autor** — Tan Shuai — 61

Etiquetas de confianza en el texto

- **verified** coinciden dos o más fuentes independientes, o dos documentos primarios independientes
- **single source** una sola fuente, incluidos los documentos primarios
- **vendor claim** publicado por el vendedor del producto o servicio
- **estimate** cálculo o inferencia del autor a partir de datos citados

Cómo leer este informe

Las cifras clave llevan una etiqueta de confianza. **verified** significa que coinciden dos o más fuentes independientes, o dos documentos primarios independientes; **single source** significa una sola fuente (incluidos los documentos primarios); **vendor claim** significa publicado por el vendedor del producto o servicio, aunque esté muy repetido; **estimate** significa mi propia aritmética o inferencia a partir de datos citados, con las entradas referenciadas. «Primaria» describe el tipo de fuente (memoria anual, folleto, regulación, norma, estadística oficial); no cambia la clase de confianza. Cuando las fuentes entran en conflicto, las cifras se muestran lado a lado y no se promedian. Las conversiones de divisa usan la paridad central del Banco Popular de China del 9 de noviembre de 2026, $1 = RMB 6.7188 (Xinhua), y para Taiwán la fijación de SMBC del mismo día, $1 = NT$32.12 (SMBC). Los datos de mano de obra y subvenciones del continente usan el delta del río Perla, donde más se concentra la fabricación de electrónica (datos salariales oficiales de Shenzhen), con el delta del río Yangtsé para comparar. Las afirmaciones de entrevistas del sector se tratan como afirmaciones a verificar, nunca como hechos. Las preguntas que el registro público no puede responder se recogen en el capítulo 7.

FIGURAS

Figuras



01	Dónde encaja la IA hoy en la planta	7
02	Una línea SMT en volumen: las personas se sientan detrás del AOI	10
03	Alrededor del 97–99 % de las alarmas AOI son falsas	11
04	Una cámara de uso general no es un AOI	13
05	Un despliegue aislado (air gap)	17
06	Las zonas limpias y ESD deciden dónde puede ubicarse el cómputo	18
07	Límites de velocidad por uso	20
08	Cómputo local: memoria, precio y qué se puede comprar legalmente en China continental	21
09	Una gran estación de trabajo no es un asistente compartido	22
10	Cuantización: 8-bit es seguro, 4-bit depende de la receta, 3-bit y menos se descarta	23
11	Candidatos a maestro: las puntuaciones genéricas se duplicaron en una generación	27
12	«¿Hay un defecto?» sigue siendo difícil para los modelos generales	28
13	Dos canalizaciones maestro–alumno	29
14	Qué tamaño debe tener el conjunto de referencia	30
15	Latencia de política edge	32
16	Alumnos robot: cerca del nivel de producción solo cuando el puesto está estructurado	33
17	Lo que cuesta al año un trabajador de inspección o manipulación	36
18	La escalera de precios	37
19	Amortización por región	39
20	Por qué extraer placas malas no se amortiza	40
21	Arancel de EE. UU. sobre robots, AOI y ordenadores edge	41
22	Los tamaños de mercado difieren en más de diez veces	44
23	Ascendente: 40.000 líneas → 16.000 servibles → ≈320 M RMB puntuales	45
24	La ventana de retrofit es de unos cinco años	46
25	La cuña de revisión está rodeada por tres lados	46
26	El margen está en el hardware industrializado	48
27	Cómo pagan los compradores	49
28	Días de ventas pendientes	50
29	La pila de IA corpórea	52
30	Qué divulgan realmente los despliegues robóticos en planta	54
31	Mano de obra neta sustituida por robot	54
32	Cronología de normas y regulación	56

RESUMEN EJECUTIVO

Merece la pena, pero de forma acotada, ahora y con socios

HALLAZGO CLAVE

El veredicto: para las fábricas de electrónica, lo que está maduro en 2026 y por lo que las plantas pagarán es la revisión con IA y la inspección de nicho sobre imágenes de AOI, más la alimentación programada de máquinas con robots, no «un modelo grande mirando placas más robots corporeizados recogiéndolas».

0.46–0.65 años

Retorno de la revisión con IA en EE. UU., Taiwán y Shenzhen, a RMB 50,000 por línea

≥95% · <20 PPM

Cobertura y fugas de la revisión con IA interna de una EMS líder: el listón para vender a EMS de primer nivel

100–700 RMB M

Mercado de reacondicionamiento de terceros aprovechable en China continental, pago único, en una ventana de unos cinco años

0 puntuaciones

Puntuaciones publicadas en pruebas de defectos industriales de los modelos insignia de noviembre de 2026: profesores, no inspectores

1.1–1.5 años

Retorno de una célula de alimentación de máquinas con cobot a dos turnos en EE. UU., Taiwán y Shenzhen

0.003–0.14 FTE

Mano de obra que ahorra por turno un brazo robótico de retirada en una línea SMT: nunca se amortiza

Inspección
ROI

El aprendizaje profundo supervisado aplicado a las imágenes que un AOI existente ya produce reduce las falsas alarmas o las horas de revisión en un 60–90% (afirmaciones de vendedores). DBG, una gran empresa de EMS, informa en su memoria anual de que su sistema interno cubre más del 95% de los componentes señalados, con fugas por debajo de 20 PPM y más de un 85% menos de revisión humana; ese es el listón que un producto de terceros debe superar para vender a EMS de primer nivel. A RMB 50,000 por línea, la revisión con IA se amortiza en menos de un año en China continental, Taiwán y EE. UU. (Shenzhen 0.65 años, Taiwán 0.53, EE. UU. 0.46), las células de alimentación de máquinas con cobot en 1.1–1.5 años y en Sudeste Asiático en 2–6 años, mientras que un brazo robótico añadido a una línea SMT solo para retirar placas defectuosas nunca se amortiza en ningún sitio, porque con un rendimiento de primera pasada superior al 90%, retirar ocupa solo 0.003–0.14 de una persona por turno estimate.

Modelos
ROBOTS

Los modelos grandes multimodales todavía no pueden ser el inspector principal. A 11 de noviembre de 2026, ninguno de los modelos insignia (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 y otros) tiene una puntuación publicada en ninguna prueba de defectos industriales o de PCB, y en el desafío PCBA de ACM MM 2026 los cinco mejores sistemas, entrenados con 6,600 imágenes reales de línea, puntuaron solo 77–82 en «¿hay un defecto?». Los modelos de frontera pueden preetiquetar y clasificar por prioridad. En cuanto a los brazos robóticos, los verdaderos profesores siguen siendo las demostraciones humanas, el aprendizaje por refuerzo sobre robot real y la simulación: los alumnos alcanzan un 90–99.4% en estaciones fijas y estructuradas, pero solo un 30–80%

en cintas transportadoras en movimiento, piezas no vistas y productos nuevos; un agente de frontera completó 104 de 108 tareas cuasi estáticas sobre robot real, a 280 segundos cada una.

En local
DESPLIEGUE

El despliegue aislado de la red en plantas limpias y controladas es totalmente práctico. Una sala SMT ordinaria suele ser un área protegida contra ESD más que una sala limpia clasificada: una caja Jetson sin ventilador o una caja NPU nacional va a la línea y los servidores GPU a la sala de servidores. Según las propias clasificaciones de exportación de NVIDIA, Jetson Thor y DGX Spark pueden venderse a una fábrica continental ordinaria sin licencia estadounidense; DGX Station no. La precisión por defecto es de 8 bits; el 4-bit cuantizado consciente del vendedor es utilizable con condiciones; 3 bits o menos queda descartado.

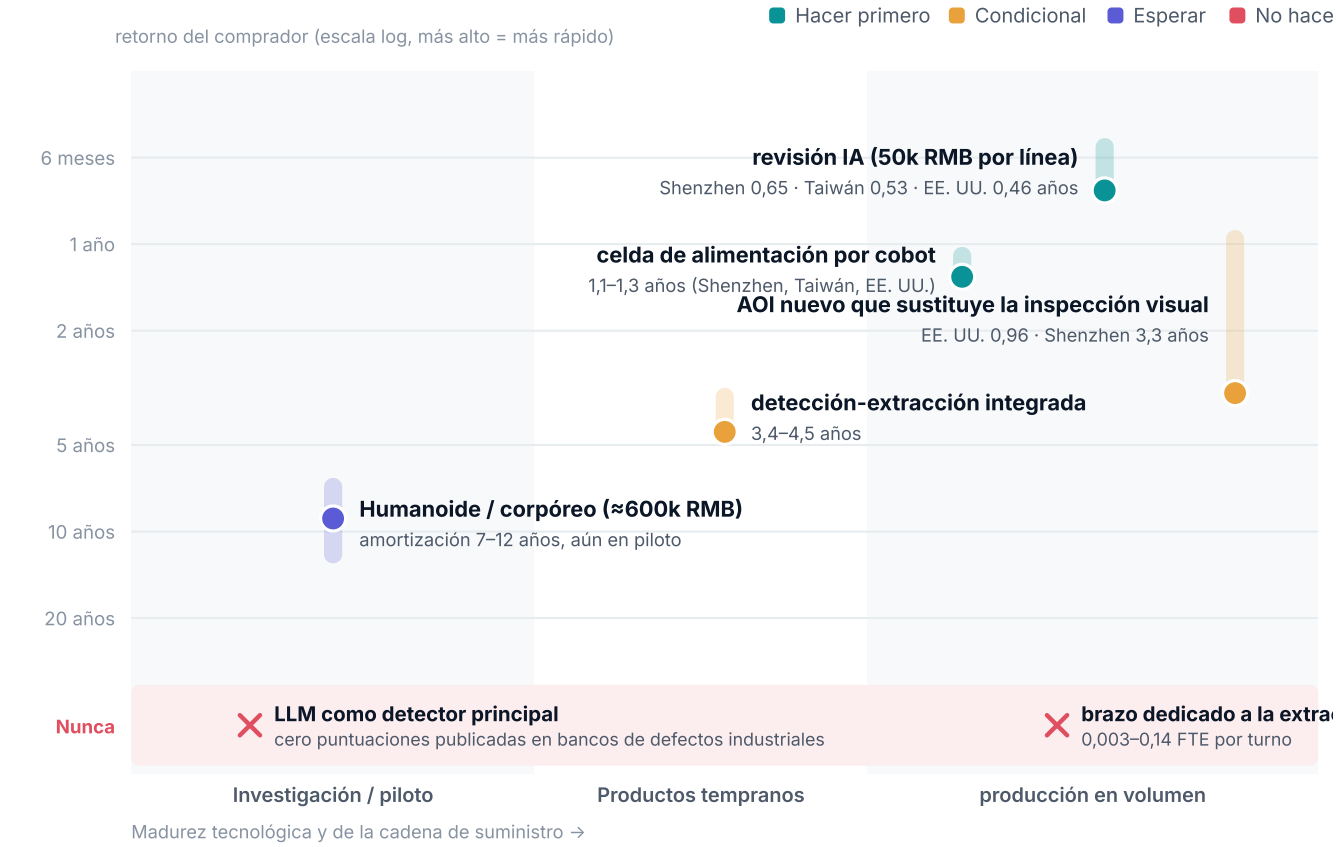
Mercado
COMPETENCIA

El mercado es real pero limitado. China continental tiene unas 35,000–46,000 líneas SMT y unos 26,000–66,000 AOI en línea sin IA; el mercado aprovechable para reacondicionamientos de terceros es de unos RMB 100–700 millones en pago único, en una ventana de unos cinco años, y los fabricantes de AOI han empezado a integrar la revisión en las máquinas nuevas estimate. La competencia ya es fuerte: la cuña de revisión está rodeada por los fabricantes de AOI, los equipos internos de las EMS y los proveedores de plataformas, y el margen reside en el hardware convertido en producto.

Pregunta	Juicio	Cifras clave (fuentes y confianza en los capítulos)
① **¿Qué ejecutan hoy las plantas de electrónica mayoritarias?**	Una línea SMT de volumen es «SPI 3D + AOI antes y después del reflujo + rayos X por muestreo + ICT/FCT + reparación e inspección visual», con revisión humana detrás de cada AOI. La IA entra en la revisión por tres vías: integrada por los fabricantes de AOI, construida internamente por las EMS líderes o añadida por terceros. El ensamblaje y la manipulación usan brazos, SCARA y AMR con percepción aprendida y movimiento programado; la IA corporeizada sigue en fase piloto	Takt ≈13–22 s por placa; un ODM presupuestó RMB 30.6M de AOI para 18 líneas, ≈3–5 AOI por línea; ≈97–99% de las alarmas de AOI son falsas
② **¿Hasta dónde llega la tecnología, puede el dinero comprar cero personas y qué puede usar cada región?**	La inspección es de grado industrial como «detector dedicado + revisión con IA + unas pocas personas»; ni siquiera dinero ilimitado llega a cero personas. El continente usa solo modelos nacionales o pesos abiertos locales; las plantas privadas de Taiwán pueden usar casi cualquier cosa, pero las fábricas de semiconductores y la infraestructura crítica excluyen en la práctica las marcas continentales; el Sudeste Asiático tiene las menores restricciones; EE. UU. tiene toda la pila más aranceles y certificación	La revisión con IA reduce las falsas alarmas o las horas de revisión un 60–90% vendor claims ; una cámara 4K sobre un panel completo ≈78 µm por píxel frente a 8–20 µm de un AOI; los brazos fabricados en el continente pagan ≈40% de arancel estadounidense combinado, los instrumentos tipo AOI ≈37.5%
③ **¿Puede ejecutarse en local en salas limpias y plantas aisladas de la red, con qué hardware y con la velocidad y precisión suficientes?**	Sí; las limitaciones son de procedimiento, no de cómputo. Una caja sin ventilador de ≤100 W en la línea, servidores GPU en la sala de servidores; Thor para estaciones, un Mac Studio o un equipo nacional para el Q&A de planta; sin DGX Station en el continente. La velocidad es suficiente cuando el tamaño del modelo se ajusta a la tarea; mantener 8 bits, usar 4-bit cuantizado consciente del vendedor con condiciones, nunca 3 bits	≈0.5 s por recorte para un modelo de visión 8B en Thor estimate ; kit de desarrollo Thor a RMB 59,104 en China; DGX Station $123–140k, 139 tokens/s en total para 32 usuarios en un modelo 743B

Pregunta	Juicio	Cifras clave (fuentes y confianza en los capítulos)
4 **¿Puede el modelo insignia más nuevo actuar como profesor durante docenas de rondas y destilarse en alumnos locales, incluso para fabricación flexible y brazos robóticos?**	Inspección: pre-etiquetado y clasificación por prioridad; aceptar automáticamente las etiquetas del profesor solo para clases con «≥299 defectos, cero fallos» en un conjunto dorado. Robots: los profesores siguen siendo las demostraciones humanas, el RL sobre robot real y la simulación, y los alumnos se acercan al grado de producción solo en estaciones estructuradas	Ningún modelo insignia de noviembre tiene una puntuación de defectos industriales; alumnos al 97.8% en conectores entrenados, 29–67% en no vistos, 37–59% en cintas en movimiento; modelos de frontera moviendo articulaciones directamente 0–5.5%
5 **¿Qué cuesta y cuál es el ROI por región?**	La revisión con IA se amortiza en menos de un año en China continental, Taiwán y EE. UU.; la alimentación con cobot en 1.1–1.5 años; el Sudeste Asiático en 2–6 años; un brazo de retirada en una línea SMT nunca se amortiza	Caja de revisión de RMB 50k por línea: Shenzhen 0.65 años, Taiwán 0.53, EE. UU. 0.46, Vietnam 2.2 estimate ; precio aceptable de complemento de terceros ≈RMB 10–30k por línea en pago único o RMB 5–10k al año
6 **¿Son ya los competidores suficientemente fuertes y es mejor empezar ahora o esperar?**	La demanda no está saturada, pero la oferta sí está abarrotada; la IA de los OEM limita la ventana de reacondicionamiento a unos cinco años. Entrar ahora con un producto acotado mediante socios de canal, sin pelear de frente con los proveedores de plataformas; esperar a los datos de producción de 2027–2028 antes de apostar por la IA corporeizada	≈26–66k AOI en línea sin IA en el continente, reduciéndose un 8–12% al año estimate ; el proveedor líder de visión con IA tuvo ingresos en 2025 de RMB 1.086B y una pérdida neta de RMB 991M; los ingresos 3C del comparable cotizado más cercano cayeron un 79.9% interanual entre enero y mayo de 2026

FIG 01 **Dónde encaja la IA hoy en la planta** ***La revisión IA de falsas alarmas AOI y la alimentación de máquinas por cobot son maduras y se amortizan rápido; un brazo de extracción dedicado y un LLM como inspector principal no merecen la pena.***



Fuente: Amortización del modelo del autor (capítulo 4); el eje de madurez es un juicio del autor.

Qué construir y qué no construir

	Recomendación
Construir	Revisión con IA y clasificación de defectos sobre imágenes de AOI, para EMS medianas con AOI antiguos o de marcas mixtas y sin actualización de IA del fabricante del AOI; inspección que el AOI de SMT cubre mal (agujeros pasantes y soldadura manual, recubrimiento conformal, cables, conectores y FPC, apariencia final y tornillos); células de carga y descarga con cobot donde se manipula cada pieza (utillajes de prueba, bandejas, embalaje), con la inspección como complemento; un conjunto dorado y el modo en sombra en una línea real como base de toda afirmación
No construir	Una plataforma general de «IA más brazo robótico»; un LLM multimodal como inspector principal; un brazo robótico en una línea SMT solo para retirar placas NG (un desviador o búfer de NG lo hace); humanoides y robots como servicio; modelos de frontera dentro de un bucle de control o una función de seguridad; montones de memoria GPU en las estaciones para ejecutar modelos «a tamaño completo»; venta directa a fábricas de semiconductores, infraestructura crítica taiwanesa o plantas de electrónica de defensa estadounidense
Con quién trabajar	Fabricantes y distribuidores de AOI (canal e interfaces de imagen; los distribuidores aportaron el 46.68% de los ingresos de 2025 de Sinic-Tek); fabricantes de cobots (brazos certificados y funciones de seguridad); empresas de EMS con 10–40 líneas (expansión progresiva y depósitos); integradores locales en Tailandia y otros lugares (para cumplir las reglas de vinculación local de BOI)

La línea hoy

01 Qué ejecutan hoy las fábricas de electrónica y qué es técnicamente viable

La línea de volumen estándar: varios AOI, rayos X por muestreo, revisión central, 13–22 segundos por placa

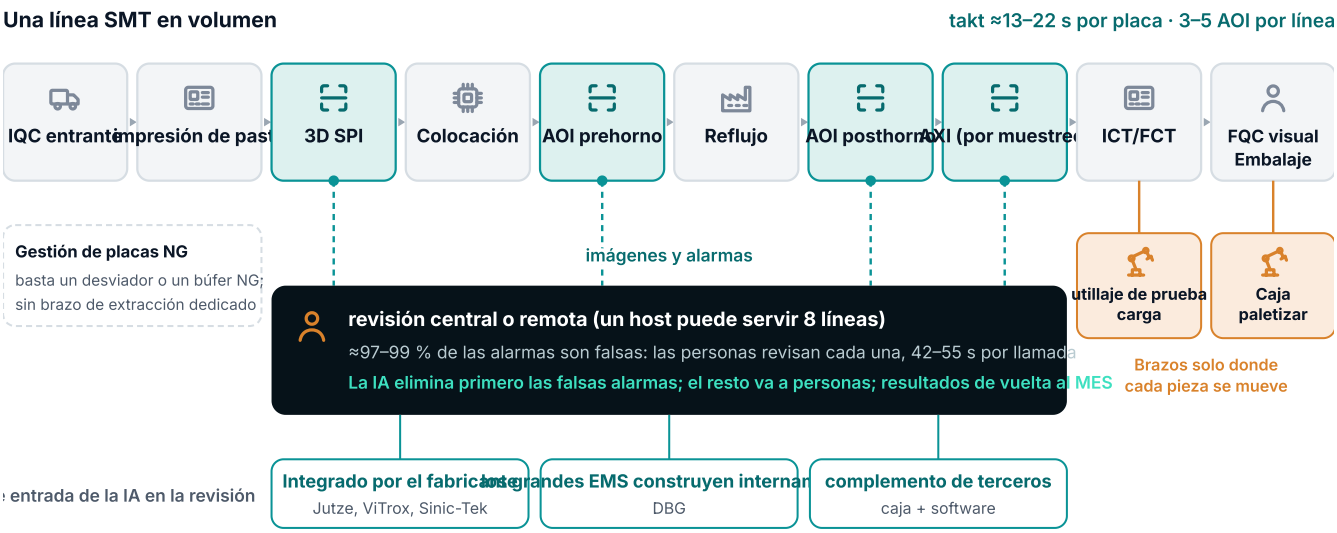
Las presentaciones primarias de ODM cotizados muestran cómo es realmente una línea de volumen de 2025–2026. El folleto de Hong Kong de Longcheer (datos al 30 de septiembre de 2025) enumera sus controles de calidad SMT como «inspección por rayos X, inspección de pasta de soldadura (SPI), inspección óptica automatizada (AOI), prueba funcional de PCBA e inspección visual»; su flujo de proceso ejecuta AOI antes del reflujo, AOI después del reflujo, rayos X por muestreo, placas NG a una estación de reparación SMT y, después, un control visual final. Sus plantas de Huizhou, Nanchang y Vietnam tienen 34, 24 y 8 líneas SMT y 67, 60 y 13 líneas de ensamblaje final (folleto de acciones H de Longcheer) single source, primary filing. La planta de portátiles de 2023 de Huaqin planificó 18 líneas SMT con un presupuesto de AOI de RMB 30.6 millones, unos RMB 1.7 millones por línea, lo que a precios nacionales de AOI 3D son de tres a cinco AOI por línea para la inspección antes y después del reflujo de placas de doble cara (folleto de acciones A de Huaqin) primary filing + estimate. Una línea de volumen, por tanto, suele tener de dos a cuatro puntos de revisión, no uno.

Las máquinas ya están muy automatizadas: Huaqin informa de un 96% de automatización en sus líneas SMT (folleto de acciones H de Huaqin) single source. Las personas restantes están en revisión, reparación, inspección visual, prueba y ensamblaje final, y ninguna empresa publica cuántos revisores tiene por línea (capítulo 7).

Tres estimaciones independientes coinciden en un takt de **13–22 segundos por placa** (una declaración ambiental de 2026 para cinco líneas en Hebei (EIA del proyecto SMT de Hebei), las líneas de portátiles de Huaqin y un caso de iFactory), o 7–11 segundos por cara para placas de doble cara estimate. La parte difícil de una línea de alta mezcla no es el volumen, sino la introducción de nuevos productos y los cambios de modelo, y cada cambio de modelo es otro arranque en frío para un modelo de IA.

En el lado del software, las estaciones de revisión de los propios fabricantes de AOI ya están centralizadas: el software de estación de reparación de Sinic-Tek permite que «un host controle el equipo de ocho líneas SMT» (folleto de Sinic-Tek) single source, primary, 2022, y después de que Jabil desplegara el control remoto de AOI ViTrox V-ONE en 2026, un ingeniero podía gestionar 5–8 AOI de forma remota frente a 1–2 en local (AtlasPCB) vendor claim. Un producto de revisión con IA de terceros se conecta exactamente a esta estación central, a través de la exportación de imágenes y resultados del AOI (IPC-CFX, IPC-HERMES-9852 o el formato propio de un vendedor). Los propios AOI están añadiendo IA rápidamente: Yamaha dice que su juicio con IA tarda aproximadamente un segundo por llamada, ayuda a los inspectores tras 30 imágenes NG y alcanza una calidad «muy alta» tras 1,000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Una línea SMT en volumen: las personas se sientan detrás del AOI** ***La mano de obra se concentra en los puestos de revisión tras la inspección óptica automática. Ahí entra la IA.***



Fuente: Documentos públicos y material de proveedores (Huaqin, DBG, Sinic-Tek y otros); compilado por el autor.

Etapa	Productos y tecnología mayoritarios	Proveedores representativos	Hasta dónde ha llegado la IA
Después de la impresión	SPI 3D (luz estructurada, moiré, algo de triangulación láser)	Koh Young, Sinic-Tek, TRI	Juicio de pasta con IA del OEM en producción; Jutze dice que algunos modelos son «totalmente IA»
Después de la colocación y el reflujo	AOI 2D/3D con LED multiángulo multicolor y luz estructurada, 8–20 µm por píxel, cámaras de 8–12 MP	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek y otros	Revisión con IA del OEM (Jutze en producción, ViTrox V510Ai, Sinic-Tek con Tencent); las EMS líderes construyen la suya (DBG); complementos de terceros
Uniones ocultas y de terminación inferior	Rayos X por muestreo o en línea (AXI)	—	Hikvision afirma que su modelo Guanlan juzga mal los huecos de rayos X menos del 1% de las veces vendor claim
Revisión y reparación	Estación de revisión central (un host, ocho líneas), revisión remota, estación de reparación	Software de los fabricantes de AOI, ViTrox V-ONE	La IA cierra falsas alarmas, las personas se encargan del resto
Inspección visual, IQC, FQC	Mayormente manual; algunas máquinas de apariencia con IA con carga robótica	SmartMore y otros	Máquinas de apariencia en producción; sin datos públicos de personal para IQC/FQC
Manipulación y alimentación de máquinas	Brazos de seis ejes y cobots, SCARA, AMR, guiado por visión 2D/3D	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Percepción aprendida más movimiento programado en producción; VLA y humanoides aún en fase piloto
Datos y sistemas	MES, IPC-CFX, IPC-HERMES-9852	—	Resultados escritos de vuelta al MES; el soporte en equipos antiguos es desigual

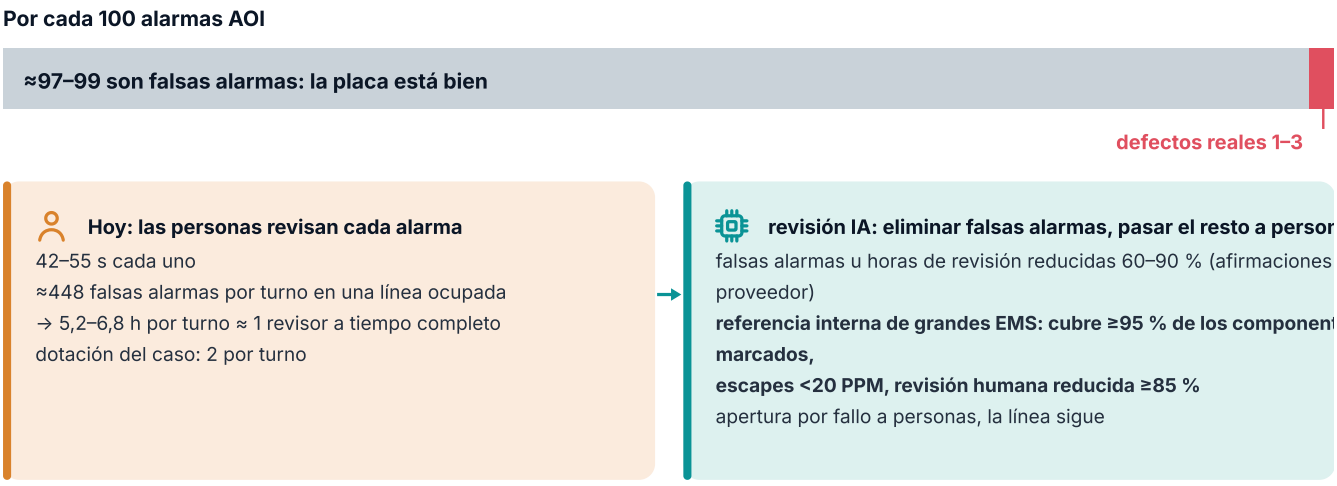
Fuentes de la tabla: las presentaciones anteriores y (Koh Young, memoria anual 2025 de Jutze, memoria anual 2025 de Sinic-Tek, ViTrox, Sohu sobre Hengtianyi, Xinhua sobre Hikvision).

Las personas se sientan detrás del AOI, porque la mayoría de las alarmas son falsas

La cadena de inspección en una línea SMT/PCBA mayoritaria es: SPI 3D después de la impresión para pasta faltante, insuficiente o excesiva, puentes y desplazamiento; AOI después de la colocación (opcional) y después del reflujo para piezas faltantes, desplazadas, giradas, invertidas o en lápida y volumen de soldadura; rayos X cuando hay BGA u otras piezas de terminación inferior; después prueba en circuito o con sonda volante, prueba funcional e inspección visual manual. Cada AOI tiene un puesto de revisión y una estación de reparación detrás (folleto de SinicTek, JLCPCB) verified. El AOI después del reflujo es el más común y el de mejor valor, pero solo ve forma y área; las uniones, los huecos y el mojado bajo un BGA requieren rayos X o prueba eléctrica (JLCPCB).

La mano de obra está en la revisión porque la mayoría de las alarmas son falsas. Una EMS británica registró **54,019 falsas alarmas frente a 544 defectos reales en 917 números de pieza, alrededor del 99% de las alarmas** (arXiv 2505.03576) single source, preprint; en la planta de Rastatt de Siemens, **el 97.5% de las placas que el AOI rechazó eran buenas** (caso de estudio de Siemens) vendor claim. Una línea ocupada ve unas 448 falsas alarmas por turno a 42–55 segundos cada una, 5.2–6.8 horas por turno, aproximadamente un revisor a tiempo completo (iFactory) vendor claim; las horas son una estimación.

FIG 03 **Alrededor del 97–99 % de las alarmas AOI son falsas** ***Hoy una persona confirma cada llamada; la revisión IA elimina primero las falsas y pasa el resto a las personas.***



Fuente: Caso de estudio Siemens, iFactory (afirmación del proveedor), informe anual DBG 2025; horas de revisión estimadas por el autor.

A partir de «horas de revisión por turno = placas × falsas alarmas por placa × segundos por llamada», un revisor puede cubrir unas 0.7–1 líneas de volumen con un AOI basado en reglas. Una vez que la IA elimina el 60–90% de las falsas alarmas, solo el tiempo de juicio permitiría 2–14 líneas por persona, pero los desplazamientos, la gestión del búfer de NG y los traspasos de reparación limitan la cobertura local a dos o cuatro líneas adyacentes; solo la revisión central o remota se acerca al techo teórico estimate; los segundos por llamada proceden de una sola fuente. Como los AOI mayoritarios escanean 60–80 cm² por segundo (guía de AOI de Yunheng 2026), una cara de 20×15 cm tarda 4–5 segundos; la IA solo tiene que procesar las pocas regiones señaladas dentro de un takt, no reescanear la placa.

Qué ya funciona en producción: clasificación supervisada sobre las propias imágenes del AOI

El patrón probado es la clasificación supervisada por aprendizaje profundo sobre las imágenes del AOI que una planta ya posee. El complemento de Fujitsu sobre el AOI de ADLINK en Taiwán redujo las falsas alarmas en **más de un 90%** (caso Fujitsu); OKI usa IA desde julio de 2026 para reducir la confirmación manual tras el AOI en aproximadamente un **80%** en su propio negocio de EMS (Circuits Assembly); Siemens informa de un 60% menos de análisis manual, €160,287 ahorrados el primer año y una amortización de ocho meses (caso de estudio de Siemens) todas

vendor claims. Dos cosas son comunes a estos casos: casi nadie publica una tasa de fuga, y todos mantienen personas en las alarmas restantes. El único caso que da una cifra de fuga (0.8% a 0.3%, contra un 0.5% contractual) es internamente inconsistente (iFactory). La IA reduce las horas de revisión, no el propio trabajo de revisión.

La revisión con IA ha pasado de complementos de terceros a funciones de OEM y sistemas internos, lo que cambia contra qué se mide a un recién llegado. La memoria anual de 2025 de Jutze enumera su «actualización de AOI 3D» como en producción en volumen y dice que «usa algoritmos de IA para eliminar la revisión manual en el extremo de inspección» (memoria anual 2025 de Jutze); Sinic-Tek ha desarrollado un «sistema de revisión manual asistida por IA multimodal» desde 2019 y está co-desarrollando un AOI 3D de arriba y abajo con Tencent, con el objetivo de una cobertura de defectos del 99.5%+ y una precisión de reconocimiento del 98% (memoria anual 2025 de Sinic-Tek) vendor claims en presentaciones primarias. Las grandes empresas de EMS construyen la suya. La memoria anual de 2025 de DBG enumera el «juicio con IA para inspección visual» como un proyecto completado que «cubre más del 95% de los componentes señalados por AOI», con una «tasa de fuga por debajo de 20 PPM» y «más de un 85% menos de revisión manual», totalmente trazable (memoria anual 2025 de DBG) single source, primary. Esa es la única cifra de planta de electrónica que encontré que da cobertura, fugas y reducción a la vez. **Un producto de terceros vendido a EMS de primer nivel tiene que igualar el 95% de cobertura, 20 PPM de fugas y 85% de reducción de DBG**; por debajo del primer nivel, el referente es la reducción del 50–85% que anuncian los vendedores.

La detección de anomalías entrenada solo con piezas buenas es rápida (EfficientAD-S tarda 2.2 ms por imagen (EfficientAD, WACV 2024)) pero débil en anomalías «lógicas» como piezas faltantes o incorrectas, y los mejores métodos promedian menos de 60% de AU-PRO en el más difícil MVTec AD 2 (MVTec AD 2) single source. Las piezas faltantes e incorrectas se comprueban mejor contra CAD y BOM.

El techo a cualquier coste: capas de máquinas, muy pocas personas, pero no cero

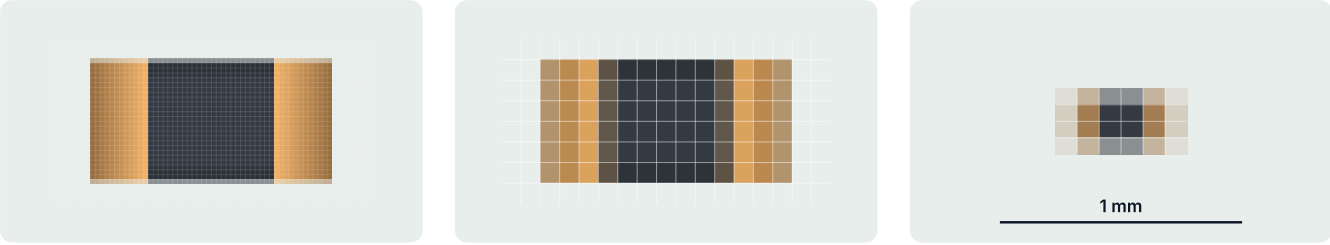
No encontré ninguna planta de electrónica que eliminara por completo la revisión manual y alcanzara cero fugas. La práctica para productos IPC-A-610 Clase 3 (alta fiabilidad) sigue haciendo que inspectores entrenados decidan los casos límite que los umbrales automáticos no pueden clasificar, y un ensamblador de PCB califica de «idea errónea común» que «la inspección automatizada puede sustituir la firma humana» (PCBCart) single source. La línea base humana también necesita corrección: citando un estudio de Sandia de 2012, un solo inspector visual detecta en el mejor de los casos cerca del **80%** de los defectos, dos en serie cerca del 96% (Instrumental) single source, citado por un vendedor. La afirmación defendible es «la IA más un pequeño bucle de revisión humana supera a la inspección totalmente manual», no «la IA sustituye a la inspección».

Sin límite de presupuesto, el techo es SPI 3D, AOI 3D, rayos X, revisión con IA y estaciones de apariencia con IA apilados, con personas solo para casos límite, productos nuevos y cambios de modelo, nuevos tipos de defectos y la firma del cliente. Las plantas insignia informan de un 85–92% menos de inspección manual (I-Connect007, 2021, WiAdvance) vendor claims; mi estimación del techo sin presupuesto es una mano de obra de revisión al 8–15% de la actual estimate. El ensamblaje final sigue siendo intensivo en mano de obra: la automatización de los botones del iPhone 16 se abandonó por altas tasas de defectos (9to5Mac) single source.

Una cámara general de alta resolución no es un AOI

Los AOI de producción capturan baldosa a baldosa a 8–20 µm por píxel, con cámaras de 8–12 MP, iluminación programable multiángulo multicolor y 3D de luz estructurada (Koh Young, Saki) verified. El sector señaló hace tiempo que una pieza 01005 cubre solo unos 8×16 píxeles a 25 µm por píxel, «posiblemente insuficiente» para un juicio fiable (Circuits Assembly, 2007). Con la misma aritmética estimate, una cámara 4K sobre un panel de 300 mm obtiene unos **78 µm por píxel**: una pieza imperial 0402 son unos 13×6 píxeles, suficiente para la presencia y el desplazamiento grueso pero no para la forma de la soldadura, y una 01005 son unos 5×3 píxeles, efectivamente invisible. Alcanzar 15 µm por píxel sobre un panel completo requeriría unos 330 megapíxeles, por lo que los AOI mueven una cámara sobre la placa. Cuando alguien dice «0402», comprueba cuál quiere decir: la 0402 métrica es la 01005 imperial, mucho más allá de cualquier cámara de gran angular.

FIG 04 **Una cámara de uso general no es un AOI** ***Una cámara 4K sobre un panel de 300 mm da ≈78 µm por píxel: un 0402 se reduce a ≈13×6 píxeles y un 01005 desaparece.***



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

Cámara 4K sobre un panel de 300 mm: ≈78 µm/px
El 0402 se reduce a ≈13×6 px: solo alcanza para presencia y desviación gruesa

Misma cámara sobre un 01005
≈5×3 px: efectivamente invisible

Fuente: Circuits Assembly; cálculo del autor para un panel de 300 mm y un sensor 4K. Las rejillas de píxeles son ilustrativas.

En las líneas que ya tienen AOI, el mejor lugar para la IA es encima de las imágenes del AOI. Solo las líneas sin AOI, o las estaciones de final de línea, justifican una estación de cámara cerrada con iluminación controlada; una cámara abierta de alta resolución sobre la línea es la más expuesta a la luz ambiental y las sombras (CSDN).

Ensamblaje y manipulación: percepción aprendida más movimiento programado ya está en producción

El informe de 2026 de IFR dice que «la visión artificial para inspección y guiado de robots está madura y es de grado industrial», mientras que «la IA física de propósito general y el aprendizaje a partir de simulación y experiencia aún no son de grado industrial» (resumen ejecutivo de IFR World Robotics 2026) single source, primary . La electrónica es el mayor comprador de robots (142,235 unidades instaladas en todo el mundo en 2025, el 59% de todas las instalaciones en China (IFR, comunicado de IFR China)), y sin embargo China tiene solo 166 robots por cada 10,000 trabajadores de fabricación frente a 307 en EE. UU. (IFR, abril de 2026). El patrón en producción es percepción aprendida más movimiento determinista; el movimiento aprendido de extremo a extremo está cerca del grado de producción solo en estaciones estructuradas (capítulos 3 y 6).

Qué puede usar cada región

China continental depende de modelos nacionales y pesos locales; en Taiwán decide la identidad del cliente; en EE. UU., los aranceles y la certificación.

Elemento	China continental	Taiwán (plantas privadas)	Sudeste Asiático	Estados Unidos
Claude (cualquier canal)	● No disponible	● Condicional: el proveedor no debe estar mayoritariamente en propiedad o bajo control de personas o entidades de China continental, Hong Kong o Macao	● Igual	● Igual
Nube de GPT y Gemini	● No disponible	● Disponible	● Disponible (listas no comprobadas país por país)	● Disponible
Modelos chinos en la nube (Qwen, DeepSeek, Doubao, GLM, Kimi)	● Disponible (modelos registrados; ofrecer un servicio requiere registro)	● Legal, pero el alojamiento continental rara vez supera la revisión de seguridad del cliente; prohibido en el sector público, empresas estatales e infraestructura crítica	● Legal (nodos internacionales)	● Legal, revisión estricta

Elemento	China continental	Taiwán (plantas privadas)	Sudeste Asiático	Estados Unidos
Pesos abiertos locales de origen chino (Qwen, GLM, DeepSeek)	🟢 Primera opción; el despliegue privado en planta no necesita registro de modelo ni de algoritmo (lectura de bufetes)	🟢 Disponible (sector público restringido)	🟢 Disponible	🟡 Condicional: revisión de origen
Pesos abiertos locales de origen no chino (p. ej. Gemma 4)	🟢 Disponible	🟢 Disponible	🟢 Disponible	🟢 Disponible
Modelos de visión clásicos no generativos	🟢 Disponible (fuera de las reglas de IA generativa)	🟢 Disponible	🟢 Disponible	🟢 Disponible
Jetson Thor/T5000/T4000, DGX Spark	🟢 Disponible: NVIDIA los clasifica 5A992.c, sin licencia estadounidense para una planta civil ordinaria; Spark tiene un SKU para China; precios de canal unos 40–60% por encima del listado estadounidense	🟢 Disponible	🟢 Disponible	🟢 Disponible
RTX 5090, RTX PRO 6000, DGX Station	🔴 No disponible: los dos primeros son 4A090.a, DGX Station 5A992.z; las alternativas legales son 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	🟢 Disponible	🟢 Disponible, pero los compradores con sede en China o propiedad última china cuentan como chinos; Malasia exige permisos de exportación y transbordo	🟢 Disponible
Mac Studio M5 Ultra 512 GB	🟢 Disponible (tienda de Apple China)	🟢 Disponible	🟡 Presumiblemente disponible	🟢 Disponible
Huawei Ascend y otros aceleradores chinos	🟢 Disponible	🔴 Alto riesgo	🔴 Alto riesgo	🔴 No utilizable en la práctica
Brazos fijos, AOI y cámaras industriales fabricados en el continente	🟢 Disponible	🟡 Importable; desde el 1 de diciembre de 2025 las autoridades pueden actuar directamente sobre las marcas continentales, y las agencias, empresas estatales, ciertas fundaciones y proveedores de infraestructura crítica pueden ser restringidos o excluidos de su uso	🟢 Disponible	🟡 Legal, pero el arancel combinado es de unos 40% en brazos, 37.5% en instrumentos tipo AOI y cámaras, 25% en ordenadores de borde
AMR y manipuladores móviles de fabricación extranjera	🟢 Disponible	🟢 Disponible	🟢 Disponible	🟡 Los modelos nuevos necesitan aprobación condicional del Departamento de Guerra de EE. UU.
Cámaras Hikvision y Dahua	🟢 Disponible	🟡 Uso privado permitido, sector público prohibido	🟢 Disponible	🔴 Nuevas autorizaciones FCC bloqueadas; prohibidas para contratistas federales

Elemento	China continental	Taiwán (plantas privadas)	Sudeste Asiático	Estados Unidos
Seguridad y certificación de robots	● GB 11291.1-2011 y GB 11291.2-2013 obligatorias, en revisión desde abril de 2026; certificación CR voluntaria; GB/T 39404-2026 cubre la ciberseguridad del controlador	● Vallado por defecto; la colaboración sin vallas debe cumplir CNS 14490 o ISO 10218 con una evaluación escrita	● Según cliente y destino de exportación, confirmado por escrito antes del envío	● El equipo eléctrico debe estar «aprobado» (en la práctica una marca NRTL o evaluación de campo); UL 1740 exige que las células robóticas cumplan ISO 10218-2 desde el 11 de agosto de 2025
Fábricas de semiconductores y plantas de productos controlados	—	● Los contratos de compra de TSMC exigen SEMI E187; se está retirando equipo continental de las fábricas de 2 nm	—	● Los datos técnicos ITAR incluyen «fotografías»; política de denegación para China; los proveedores controlados desde el continente no pueden en la práctica atender plantas de electrónica de defensa

Fuentes de la tabla: (regiones compatibles de Anthropic, países compatibles de OpenAI, disponibilidad de Gemini, medidas provisionales de la CAC sobre IA generativa, datos de clasificación de exportación de NVIDIA, eCFR 15 CFR 774, NVIDIA China DGX Spark, Apple China, Covington sobre nuevas reglas BIS, Greenberg Traurig, junio de 2026, Fulcrum, Tom's Hardware sobre guía Ascend, lista de importación BOFT de Taiwán, historial de la Ley de Gestión de Ciberseguridad de Taiwán, reglas de revisión de Taiwán para productos que ponen en peligro la ciberseguridad nacional, Taipei Times, capítulo 99 del HTS del USITC, Latham & Watkins, IPVM, plataforma de normas nacionales de China, GB 11291.2, actualización Intertek UL 1740, eCFR 22 CFR 126.1).

China continental: una pila completa, pero ni los modelos de frontera cerrados ni las máquinas individuales más fuertes pueden entrar. Los modelos nacionales registrados en la nube, los pesos abiertos locales (Qwen bajo Apache-2.0, GLM y DeepSeek bajo MIT), los modelos de visión clásicos y los AOI, cámaras y cobots nacionales son todos utilizables. Las propias clasificaciones de exportación de NVIDIA enumeran el kit Jetson AGX Thor, los módulos T5000/T4000 y DGX Spark como 5A992.c, de modo que una planta ordinaria de EMS o PCBA no necesita licencia estadounidense, mientras que la RTX 5090 y la RTX PRO 6000 (4A090.a) y DGX Station (5A992.z, TPP 58,484) no pueden comprarse en el continente (clasificaciones de exportación de NVIDIA, eCFR) primary; la lectura legal es mía y no es asesoramiento jurídico. RS China lista el kit Thor a RMB 59,104 impuestos incluidos, unas 1.4 veces el precio de lista estadounidense (RS China).

Cuatro bufetes coinciden en que un despliegue privado en planta para uso propio de la planta no necesita registro de modelo ni de algoritmo, mientras que un servicio alojado multiinquilino queda en una zona gris (Covington, FPF, JunHe) verified: lecturas coherentes, sin resolución oficial. GB 11291.1-2011 y GB 11291.2-2013 siguen siendo la línea base obligatoria de seguridad robótica y están en revisión desde el 28 de abril de 2026; GB/T 39404-2026 sobre ciberseguridad del controlador entró en vigor el 1 de noviembre de 2026 (plan de revisión 20262542, GB/T 39404-2026) single source, primary. Hasta que llegue el nuevo GB 11291 (2028–2029 estimate), diseña según ISO 10218:2025 y mantén los expedientes de cumplimiento GB 11291.

Taiwán: las plantas privadas pueden usar casi cualquier cosa; decide la identidad del cliente. Las API de frontera, los pesos abiertos y el equipo continental son todos legales (data.gov.tw). Pero desde que la Ley de Gestión de Ciberseguridad enmendada entró en vigor el **1 de diciembre de 2025**, las autoridades pueden actuar directamente sobre productos que sean «marcas continentales» y restringirlos en organismos gubernamentales, empresas estatales y proveedores de infraestructura crítica (Base de Datos de Leyes y Reglamentos, reglas de revisión) verified; la lista de proveedores de infraestructura crítica es confidencial; y TSMC ha incorporado SEMI E187 en sus contratos de compra (CNA) y se informa de que está retirando equipo continental de sus fábricas de 2 nm (TrendForce) single source. Las EMS privadas, el ensamblaje y las pequeñas plantas de PCBA están abiertas; las fábricas de semiconductores, las utilidades de parque y los clientes relacionados con el gobierno están cerradas en la práctica. Un proveedor controlado desde el continente debería diseñar para E187, enviar una lista de materiales de software, ofrecer actualizaciones sin conexión y nunca conectarse de forma remota desde el continente.

Sudeste Asiático: las menores restricciones; los problemas son la amortización lenta y las matrices chinas. Todo es utilizable, pero la guía estadounidense del 31 de mayo de 2026 exige una licencia para exportar artículos de computación avanzada a entidades con sede en China o propiedad última china incluso cuando el destinatario está en un tercer país (Greenberg Traurig) single source, y Malasia tiene sus propios permisos para chips avanzados (Fulcrum). La demanda se está moviendo aquí: TrendForce esperaba que Tailandia, Vietnam y Malasia se acercaran al 50% de la capacidad SMT de ODM de servidores en 2026 (TrendForce) single source, una previsión de 2023.

Estados Unidos: la pila completa, al precio de aranceles, certificación y datos controlados. Según el capítulo 99 del arancel estadounidense, un brazo robótico fabricado en el continente paga cerca de un 40% de arancel combinado, los instrumentos tipo AOI y las cámaras cerca del 37.5% y los ordenadores de borde el 25%; fabricado en Malasia o Indonesia un 10–12.5%, en Vietnam, Tailandia o Filipinas un 12.5–15%, en Taiwán un 10%, con los ordenadores al 0% fuera del continente (capítulo 99 del HTS del USITC, Great Lakes Customs Law) primary schedule + estimate. El aviso del USTR del 10 de noviembre extendió 178 antiguas exclusiones de la Sección 301 hasta el 9 de enero de 2027, ninguna de ellas cubriendo robots, instrumentos de inspección, cámaras u ordenadores (Federal Register 2026-22925) verified.

OSHA exige que el equipo eléctrico del lugar de trabajo esté «aprobado» (en la práctica una marca NRTL, o una evaluación de campo de unos $3,000 a $20,000+ por célula), y desde el 11 de agosto de 2025 UL 1740 somete las células robóticas completas a ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). La electrónica de defensa está cerrada a los proveedores controlados desde el continente en la práctica: ITAR trata las fotografías de producción como datos técnicos y aplica una política de denegación a China (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Una célula destinada a EE. UU. debería usar un brazo fijo, cámaras sin restricciones y modelos locales o en la nube estadounidense, con ensamblaje final en el Sudeste Asiático donde eso cambie genuinamente el origen.

A cualquier presupuesto, el techo difiere por región pero en ningún sitio alcanza cero personas estimate. El continente puede apilar SPI 3D, AOI 3D, rayos X, revisión con IA y un equipo interno de IA nacional para bajar la mano de obra de revisión al 8–15%, pero sin modelos de frontera cerrados ni DGX Station, el Q&A con LLM de toda la planta tiene que ejecutarse sobre pesos abiertos locales y cómputo nacional (un modelo a tamaño completo para muchos usuarios necesita al menos dos servidores Ascend de ocho tarjetas). Taiwán y el Sudeste Asiático pueden técnicamente usar lo mejor de todo, incluidas las API de frontera y DGX Station, limitados por la identidad del cliente, las marcas continentales y las licencias de matrices chinas. EE. UU. no tiene límites técnicos, al precio de aranceles, certificación NRTL y separación de personal para datos controlados. En todas partes quedan personas para los casos límite, los productos nuevos y los cambios de modelo, los nuevos tipos de defectos y la firma del cliente.

Un producto vendido entre regiones, por tanto, necesita dos pilas de modelos: una pila continental sobre modelos nacionales registrados o pesos abiertos locales, y una pila extranjera sobre API de frontera más un modelo local de origen no chino. La política de Anthropic vigente desde el 12 de noviembre de 2026 juzga por propiedad y control, de modo que un proveedor controlado desde el continente no puede construir sobre Claude ni siquiera a través de una filial extranjera, una plataforma en la nube o un revendedor (Anthropic) single source, primary.

En local

02 Ejecutar la IA dentro de la planta: salas limpias, aislamiento de red, hardware y cuantización

En una planta aislada de la red, la limitación es el proceso, no el cómputo

La inspección visual totalmente sin conexión es rutinaria; las limitaciones son de procedimiento. La guía MIIT de China de 2024 sobre seguridad de sistemas de control industrial pide zonificación, cortafuegos o diodos de datos, mantenimiento remoto auditado de tiempo limitado, soportes escaneados y parches probados (texto de la guía) single source ; los clientes de semiconductores añaden SEMI E187 y E188, que TSMC incorporó a sus contratos de compra en 2023 (SEMI, TXOne) verified . Los resultados vuelven al MES por IPC-CFX y la identidad de placa viaja por IPC-HERMES-9852 (norma Hermes). La trampa común es la licencia que asume una red: el modo sin conexión de Roboflow almacena los pesos en una concesión de 30 días (Roboflow). Un producto sin conexión necesita activación sin conexión, actualizaciones firmadas en soportes escaneados, una lista de materiales de software, control USB, registros de auditoría locales y un adaptador CFX.

FIG 05 **Un despliegue aislado (air gap)** ***Una cajita en línea decide; la sala de máquinas responde preguntas; los fallos recaen en las personas y los datos nunca salen de la planta.***



Fuente: Síntesis del autor (capítulo 2).

Los modelos de frontera cerrados, en la práctica, no entran en la sala de servidores de una planta. Gemini se ofrece aislado de la red solo a través de Google Distributed Cloud para gobiernos y grandes empresas (Google Cloud); Foundry Local de Microsoft abre los modelos grandes solo a «clientes cualificados» y no nombra ningún modelo GPT (Microsoft); Claude se ofrece solo en regiones gubernamentales de la nube, sin despliegue en local (Claude for government); ninguno de los tres se ofrece en China continental. El equipo de planta tiene que diseñarse en torno a pesos abiertos; los modelos de frontera como mucho pueden etiquetar o explicar imágenes en la sede, sin conexión, cuando el cliente permite que salgan imágenes anonimizadas.

El hardware de la sala de servidores debería tratarse como de grado oficina: el Mac Studio está clasificado para 10–35 °C y DGX Spark para 5–30 °C, con informes de apagados térmicos bajo inferencia sostenida (Apple, foro de

NVIDIA) verified. La línea recibe cajas industriales sin ventilador, como un PC industrial Jetson T4000 clasificado para −25 a 60 °C (AAEON) vendor claim. Como una hora de inactividad cuesta a una planta grande de $36,000 a $2.3 millones (Siemens True Cost of Downtime 2024) [encuesta de vendedor], cada paso de IA en la línea debe **fallar abierto**: ante cualquier fallo, vuelve al proceso existente de AOI más personas.

Zonas limpias y de ESD: una sala SMT ordinaria es un área protegida contra ESD, no una sala limpia clasificada

Lo que limita dónde puede situarse el cómputo es la estática y el flujo de aire, no la clase de limpieza. La mayoría de las áreas de ensamblaje SMT «no están formalmente controladas en partículas», y las plantas de clase mundial «pueden alcanzar o acercarse a ISO 8» (panel de expertos de Circuitnet) single source. La PCBA médica y los módulos de cámara suelen estar en ISO 7, el dado desnudo y la alineación óptica en ISO 5–6 local (PCBSync) single source; los constructores de salas limpias citan ISO 7–8 para SMT en general (Angstrom Technology) vendor claim.

El requisito universal es el control de ESD: ANSI/ESD S20.20-2021 exige conductores aislados por debajo de 35 V y mantiene los aislantes de campo fuerte al menos a 30 cm de los dispositivos sensibles (ANSI, ESDA, Desco) verified. Las fábricas de semiconductores sitúan el cómputo en el sub-fab o el centro de datos, en bastidores refrigerados por líquido que cumplen SEMI S2/S8/F47 (Prodrive) vendor claim.

Si un equipo puede entrar en una zona limpia se prueba por emisión de partículas bajo ISO 14644-14; los robots tienen dichos certificados (Fraunhofer IPA), pero no encontré ningún PC industrial, caja de IA de borde o servidor GPU con uno. CE, FCC e IEC 60068 muestran que una caja tolera su entorno, no que pueda entrar en un área protegida contra ESD; eso lo comprueba el coordinador de ESD del cliente. Los servidores GPU nunca van a la línea: una DGX Station consumió 1,292 W en la GPU en las pruebas (StorageReview).

FIG 06 **Las zonas limpias y ESD deciden dónde puede ubicarse el cómputo** ***Una nave SMT ordinaria es una zona protegida ESD, no una sala limpia. Los servidores GPU y NPU van a la sala de máquinas.***



Fuente: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, normas SEMI y material de proveedores; compilado por el autor.

Zona	Entorno típico	Cómputo en la línea	Servidores GPU/NPU
Sala SMT/PCBA ordinaria (consumo, industrial)	Área protegida contra ESD bajo S20.20 o IEC 61340-5-1, normalmente no una sala limpia clasificada (≈equivalente ISO 8); 40–60% de HR	Caja sin ventilador en una carcasa metálica conectada a tierra (Jetson T4000/Thor, caja NPU china) junto al AOI o al puesto de revisión; monitores y carcasas de plástico fuera del área o ≥30 cm de las placas	Sala de servidores o bastidor de TI climatizado

Zona	Entorno típico	Cómputo en la línea	Servidores GPU/NPU
Médica, módulos de cámara, electrónica de VE (sala limpia ISO 7–8)	Sala limpia más protección ESD	Sin ventilador, sellada IP65/66, lisa y fácil de limpiar; pedir datos ISO 14644-14 o un certificado IPA	Fuera de la sala limpia, conectada por red
Dado desnudo, unión por hilo, estaciones ópticas (ISO 5–6 local)	Miniambiente	Solo cámaras y luces dentro; la caja de cómputo fuera	Sala de servidores
Sala limpia de fábrica de semiconductores (ISO 5 y más limpia)	Planes de ESD y atracción electrostática específicos de fábrica, equipo según SEMI E78; ciberseguridad según SEMI E187/E188	Solo controladores integrados en herramienta y cualificados de fábrica; sin PC adicionales con ventiladores	Bastidores refrigerados por líquido del sub-fab o el centro de datos de la fábrica

La tabla es mi síntesis de las fuentes anteriores estimate. Para un producto, eso significa una caja sin ventilador de línea por debajo de unos 100 W, una carcasa metálica unida a tierra de protección, entrada de 12–24 V CC, operación en amplio rango de temperatura con el flujo de aire real del armario, y un informe de emisión de partículas ISO 14644-14 listo para clientes en ISO 7 y más limpio. Cualquier diseño donde «un servidor GPU sirve a cuatro líneas» pone el servidor en la sala de servidores y deja solo una caja pequeña o una cámara en la línea.

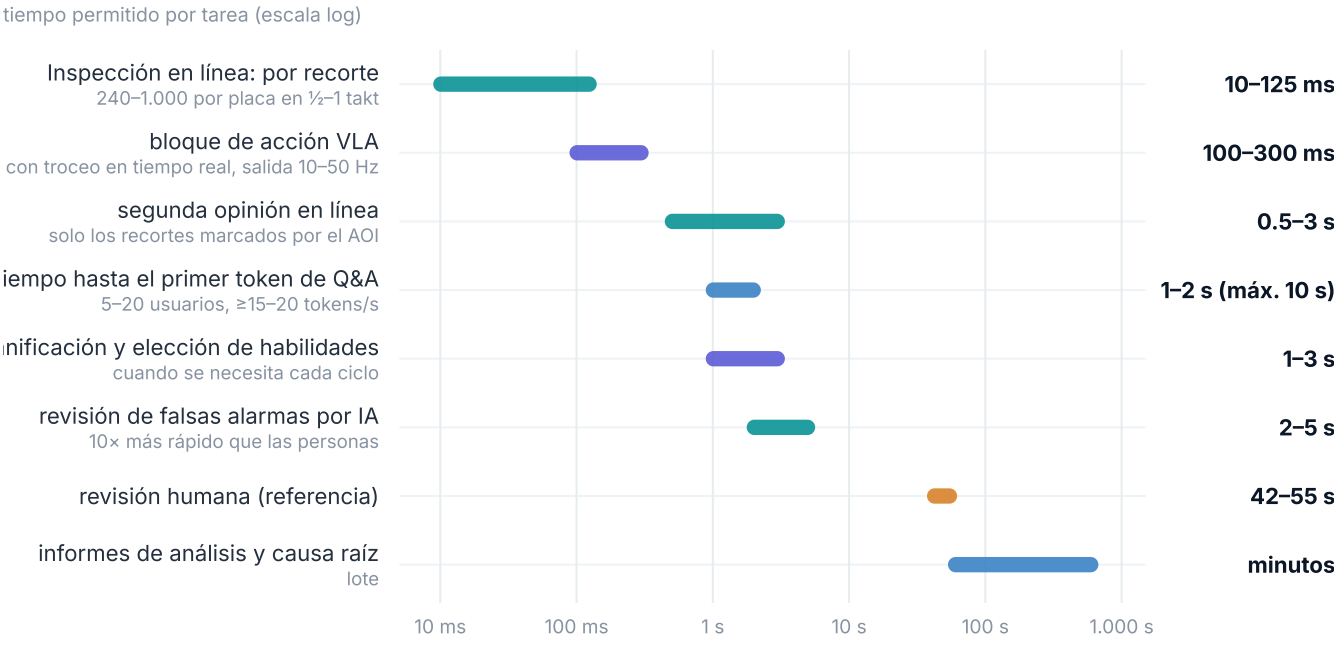
Límites de velocidad por uso: la precisión es más difícil que la velocidad

Las plantas y los vendedores casi nunca publican objetivos de latencia o disponibilidad de IA. Los umbrales de abajo se derivan del takt, el tiempo de revisión humana, las tasas de control de robots y la investigación sobre interacción estimate:

Uso	Límite de velocidad	Límite de precisión	Disponibilidad
Inspección en línea de cada pieza	Dentro de ½–1 takt: 240–1,000 recortes en un takt de 10–30 s, 8–100 recortes por segundo	Fugas no peores que el «AOI más personas» actual (ejemplo contractual <0.5%)	Fallar abierto, nunca parar la línea; objetivo ≥99.9% (sin SLA público)
Segunda opinión en línea solo sobre recortes señalados por AOI	Señales (normalmente ≤5–20 por placa) dentro de ½ takt: 0.5–3 s por recorte	Igual; entrenado con las propias imágenes de la línea	Vuelve a la revisión humana
Revisión con IA de falsas alarmas	2–5 s por llamada, más de diez veces más rápido que los 42–55 s de una persona	Fugas no peores que un revisor humano	La degradación añade personas, nunca para la línea
Control de robot (VLA)	100–300 ms por fragmento de acción con fragmentación en tiempo real, salida 10–50 Hz	≥99% de éxito de tarea para superar al movimiento programado	Sistema de seguridad independiente del modelo
Planificación de robot y elección de habilidad	1–3 s cuando se necesita cada ciclo; ≤1 min si solo en excepciones o cambios de modelo	La salida debe ser código verificable o un ID de habilidad	Vuelve al programa fijo
Q&A (SOP, consultas MES)	Primer token idealmente ≤1–2 s, como mucho 10 s; ≥15–20 tokens/s por usuario; 5–20 usuarios concurrentes	Citas precisas; fiel en contexto largo	Grado TI de oficina (≈99–99.5%)
Informes de análisis y causa raíz	Los minutos están bien (por lotes)	Máximo; fiel en contexto largo	Grado TI de oficina

Physical Intelligence midió 97 ms de inferencia para un VLA de clase 3B (unos 108–139 ms con red y preprocesado), encontró que la fragmentación en tiempo real mantiene el rendimiento sin cambios hasta 200 ms de retardo extra, y completó tareas finas como encender una cerilla o enchufar un cable de red a más de 300 ms (Physical Intelligence) [investigación de vendedor]; la investigación de interacción sitúa el límite para mantener el hilo de pensamiento en un segundo y para la atención en diez (NN/g). La precisión es la parte difícil: el Qwen3-VL-8B abierto promedia solo 56.8 en zero-shot en cinco conjuntos de anomalías industriales (IndusAgent) single source. Ningún hardware rescata un modelo de visión-lenguaje que no se ha afinado con las propias imágenes de la línea.

FIG 07 **Límites de velocidad por uso** ***La inspección en línea se mide en milisegundos, el Q&A en segundos, el análisis en minutos. Las personas tardan 42–55 s por llamada.***



Fuente: Síntesis del autor (capítulos 1–2).

Hardware: Thor para la estación, un Mac Studio o equipo nacional para el Q&A de planta, y sin DGX Station en el continente

Dispositivo	Memoria y ancho de banda	Precio (noviembre de 2026)	Para qué es bueno	China continental
Kit de desarrollo Jetson AGX Thor / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 en el lanzamiento → $5,499 (julio de 2026); T5000 $4,999, T4000 $2,999; kit de desarrollo de RS China RMB 59,104 imp. incl.	Modelos de visión 2–8B recomprobando recortes; VLA de clase 3B en tiempo real	Sí (5A992.c, sin licencia para plantas civiles)
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; el modelo China del mismo chip ASUS GX10 a RMB 32,999 en JD (oct. 2025)	MoE 8–35B multiusuario	Sí (5A992.c, modelo China, canales autorizados)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (jun. 2026) → ≈$16,005 listado medio de eBay en noviembre	27–35B FP8 multiusuario; MoE de clase 120B a 4 bits	No (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB unificada, 1.2 TB/s; 480 W máx. sostenido	Desde $22,299 en EE. UU.; ≈RMB 173–186k estimado en China	DeepSeek 671B a 4 bits para 1–3 usuarios	Sí

Dispositivo	Memoria y ancho de banda	Precio (noviembre de 2026)	Para qué es bueno	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 en la GPU	$123–140k precio de calle en noviembre (≈RMB 830–940k antes de aranceles); $95–117k en agosto	Alta concurrencia en MoE 120–235B que cabe en HBM; solo unos pocos usuarios en modelos 433–743B	● No (5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	RMB 13,300 de un comerciante de JD	Revisión por lotes de bajo volumen; pocos modelos compatibles	● Sí (nacional)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	No obtenido; DeepSeek completo necesita dos en W8A8, cuatro en BF16	Q&A y análisis multiusuario de toda la planta	● Sí (nacional)
RTX 5090 D V2 (modelo China)	24 GB	RMB 16,499 en el lanzamiento (ago. 2025), al alza desde entonces con los precios de memoria	Modelos de visión 8B, VLA pequeños	● Sí

Fuentes: (NVIDIA China Thor, lanzamiento de NVIDIA Thor, CNX sobre subidas de precio, datos de clasificación de exportación de NVIDIA, RS China, Implicator, IThome sobre ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, especificaciones de Apple, TechRadar, IThome sobre precios del M5 Ultra en China, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, matriz de compatibilidad vLLM-Ascend, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). Los precios continentales del Mac Studio de 512 GB se extrapolan de las primas de configuración estadounidenses y son de baja confianza estimate.

FIG 08 **Cómputo local: memoria, precio y qué se puede comprar legalmente en China continental** ***Las clasificaciones de exportación deciden la lista corta antes que los benchmarks.***



Fuente: Clasificaciones de exportación de NVIDIA, Apple, RS China, JD.com y precios de revisión (noviembre de 2026).

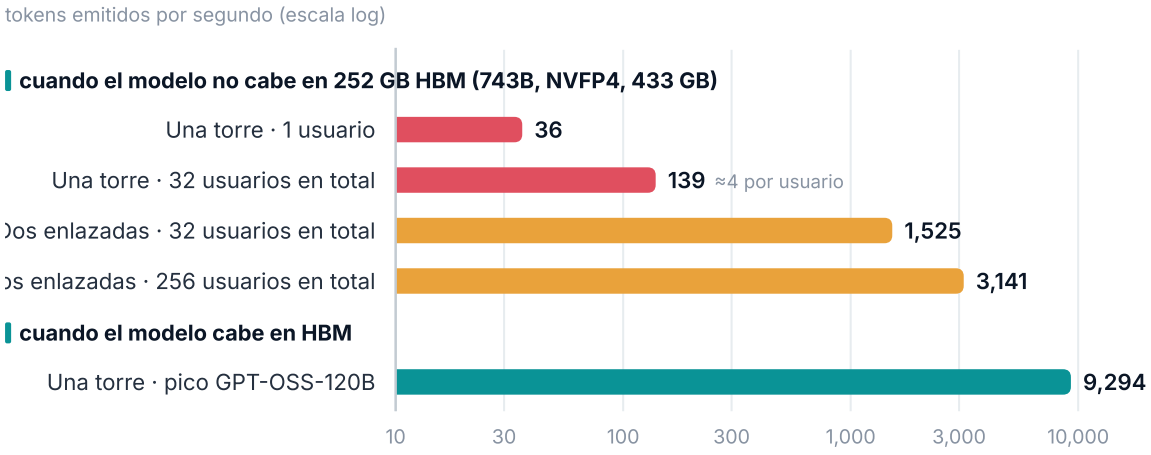
Jetson Thor hace funcionar el despliegue a nivel de estación. Las pruebas TensorRT Edge-LLM de NVIDIA muestran Qwen3-VL-8B (NVFP4) prellenando 2,048 tokens en 104.5 ms y decodificando 44.4 tokens/s en Thor (TensorRT Edge-LLM) vendor claim; para un recorte de AOI y una respuesta de 20 tokens, eso son unos 0.5 s con el modelo 8B y 0.2 s con el 2B, o 8.5–21 recortes por segundo en lote estimate. La longitud de salida, no el número de parámetros, fija la latencia, de modo que la decodificación restringida a un token de clase o JSON corto forma parte del

diseño. GR00T N1.7 (3B) corre a 25.1 Hz en Thor (Jetson AI Lab) vendor claim. Una Orin NX tarda unos 1.8 s por recorte y una Atlas 300I Duo tiene un p95 por encima de 23 s, apta solo para trabajo por lotes (CSDN) single source. En el continente, presupuesta los productos NVIDIA un 40–60% por encima del precio de lista estadounidense estimate.

Un Mac Studio de 512 GB hace posible un modelo a tamaño completo en una máquina, para unos pocos usuarios. El M5 Ultra (agosto de 2026) ofrece hasta 512 GB a 1.2 TB/s desde unos $22,299 (Apple, TechRadar); DeepSeek 671B a 4 bits (≈380–404 GB) decodifica 17–20 tokens/s en un M3 Ultra (Pinggy) verified y un estimado 25–29 en el M5 Ultra. Los prompts largos son el punto débil: un prompt de recuperación de 10,000 tokens tarda unos 17–70 s estimate, más allá del límite de interacción de 10 segundos.

DGX Station es la máquina individual más fuerte, pero no está disponible en el continente y cuesta de cinco a seis veces un Mac. El GB300 que se envía tiene **748 GB**, de los cuales solo 252 GB son HBM rápida (Wccftech, Tom's Hardware, ago. 2026) verified. Las pruebas independientes muestran unos 9,300 tokens/s de salida en GPT-OSS-120B, que cabe en HBM, pero un modelo 743B que se desborda a LPDDR5X dio a 32 usuarios solo 139 tokens/s en total, unos 4 cada uno; dos unidades enlazadas alcanzaron 1,525 (StorageReview, ago. 2026, StorageReview, oct. 2026) single source. Los precios de calle subieron a $123–140k en noviembre (CyberPowerPC, monitors.com). En el continente, las opciones realistas para el mismo trabajo son el Mac Studio, servidores CPU+GPU estilo KTransformers (KTransformers) [prueba de vendedor] o servidores Ascend nacionales.

FIG 09 **Una gran estación de trabajo no es un asistente compartido** ***Cuando el modelo no cabe en memoria de gran ancho de banda, 32 usuarios simultáneos obtienen unos 4 tokens por segundo cada uno.***



Fuente: StorageReview (ago. y oct. 2026); ServeTheHome. Pruebas independientes de una sola fuente.

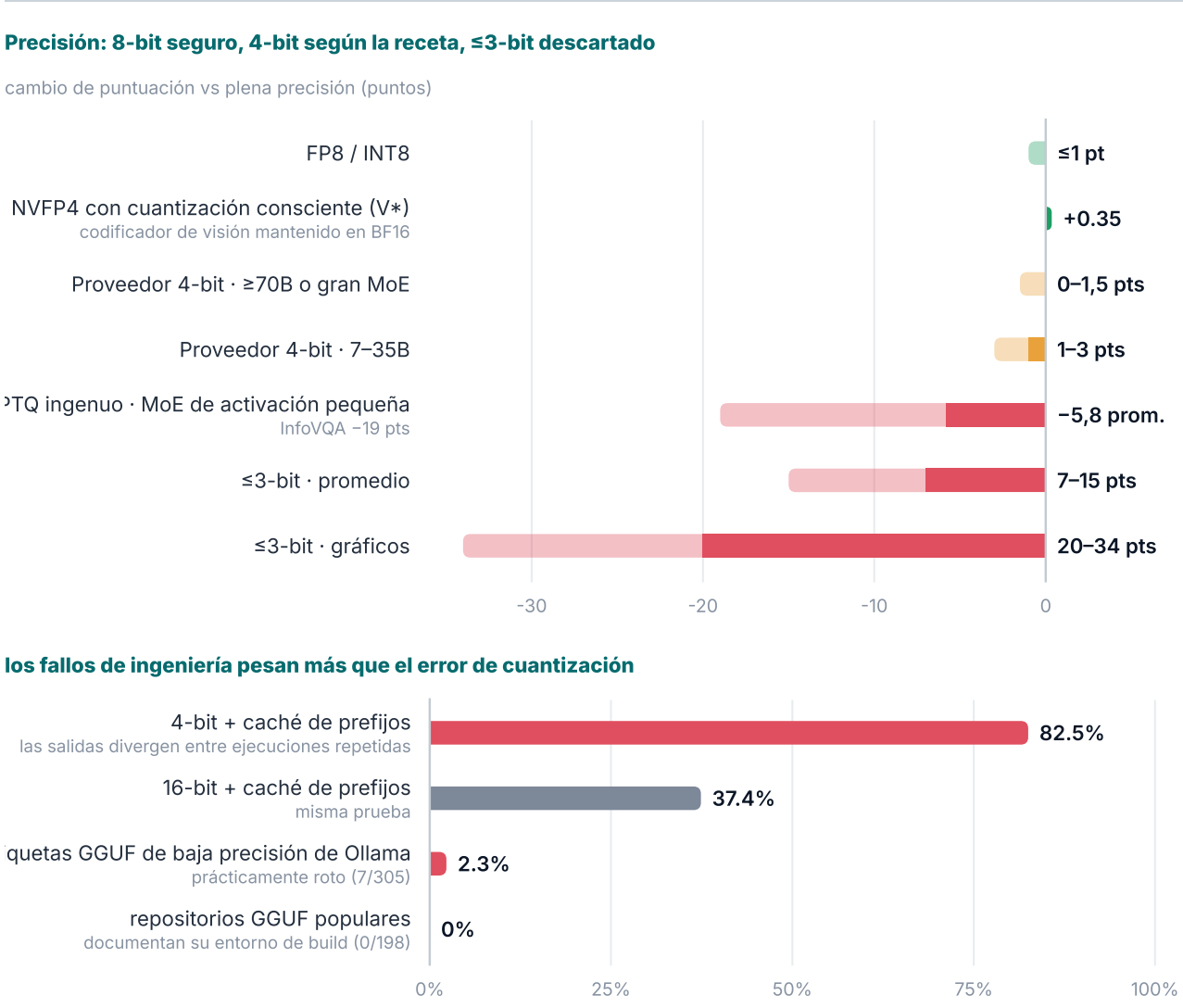
La escasez de memoria movió las mismas piezas un 20–100% en un año (DGX Spark pasó de $3,999 a $6,950 (TechRadar)), de modo que las cotizaciones de equipo deberían indexarse en lugar de fijarse.

Los aceleradores nacionales ejecutan modelos a tamaño completo; el mantenimiento de software es el cuello de botella. Ascend 910B tiene 64 GB de HBM, pero vLLM-Ascend marca DeepSeek V3/R1 como «ya no mantenido» y solo admite unos pocos modelos en tarjetas más pequeñas (vLLM-Ascend) single source, primary, y cada modelo nuevo tarda uno o dos meses en adaptarse (Caijing/Anue); envía una versión de modelo congelada y validada. Servir DeepSeek 671B completo a 20–100 usuarios necesita al menos dos servidores Atlas 800I A2 de ocho tarjetas (China Telecom Cloud) verified, y sin embargo las licitaciones públicas de equipos LLM en China totalizaron solo RMB 27.5 millones entre enero y septiembre de 2025 (C114) single source: las plantas pagan por integración y servicio, no por una caja de cómputo.

Cuantización: 8 bits es seguro, 4 bits depende de la receta, 3 bits o menos queda descartado

8 bits (FP8 o INT8) es esencialmente sin pérdidas para los modelos de visión-lenguaje (Red Hat, Hugging Face) verified. A 4 bits, la pérdida depende del tamaño y la receta: la compilación AWQ de 72B de Qwen pierde menos de un punto en MMMU mientras que la 7B pierde 2.8 (Qwen AWQ) verified, y las tareas con muchos detalles caen más (el GPTQ 4-bit ingenuo cuesta a Qwen3-VL-30B-A3B 19 puntos en InfoVQA (VEQ)). A 3 bits o menos, la pérdida media es de 7–15 puntos y los gráficos pierden 20–34 (VLMQ) verified; más allá de 64k tokens, 4 bits pierde hasta un 59%, peor en entrada no inglesa (arXiv 2505.20276). Mantén el codificador de visión en FP16/BF16.

FIG 10 **Cuantización: 8-bit es seguro, 4-bit depende de la receta, 3-bit y menos se descarta** ***Los errores de ingeniería en la pila de servicio cuestan más precisión que la propia cuantización.***



Fuente: Ficha del modelo NVIDIA Nemotron, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 y otros.

Dos advertencias importan más. Primero, **a 11 de noviembre de 2026, ningún estudio ni tarjeta de modelo ha medido la cuantización en imágenes de defectos de PCB/SMT**, así que revalida en el conjunto dorado. Segundo, los errores de la pila de servicio suelen ser mayores que el error de cuantización: el mismo modelo puntuó unos 30 puntos menos bajo Ollama que bajo la implementación de referencia (SiliconBench); 7 de las 305 etiquetas oficiales de código de Ollama estaban esencialmente rotas aunque los archivos estuvieran intactos (arXiv 2609.05881); ninguno de los 198 repositorios GGUF populares registró su compilación (arXiv 2610.19498); y con reutilización de caché de prefijo, el 82.5% de los episodios divergieron a 4 bits frente al 37.4% a 16 bits (arXiv 2609.04748) single sources. El 4-bit cuantizado consciente del vendedor está cerca de ser sin pérdidas en tareas de detalle (NVIDIA Ne-

motron 3.5) vendor claim, pero los VLA a 4 bits pierden 1.5–4.3 puntos bajo perturbación y corren más lento en Thor (LoopQ) single source. Gestiona un modelo cuantizado como un calibre nuevo: fija los hashes de pesos, registra la compilación, restablece la caché de prefijo y revalida tras cualquier cambio de precisión, motor o plantilla.

Precisión	≥70B denso o ≥100B MoE	7–35B	≤4B o tareas con muchos detalles	Veredicto
FP8 / INT8	Seguro (≤1 punto)	Seguro	Seguro; el valor por defecto para inspección	Usar
4-bit AWQ/GPTQ (receta del vendedor, codificador de visión FP16)	0–1.5 puntos	1–3 puntos; GPTQ ingenuo en MoE de activación pequeña −5.8 de media, −19 en InfoVQA	2–3 puntos; validar primero las tareas de detalle	Condicional
NVFP4 (receta del vendedor)	≈0.3–1 punto por debajo de FP8 (datos del vendedor)	Compilaciones conscientes de cuantización casi sin pérdidas en V*, OCRBench	Sin datos públicos	Condicional (preferir compilaciones conscientes de cuantización)
GGUF Q4_K_M	Sin probar	Alta varianza en tareas de dominio	No para texto pequeño ni defectos	No para juicio de detalle
GGUF Q5_K_M y superiores	Presumiblemente seguro	Casi sin pérdidas en dos pruebas	El suelo para detalle en llama.cpp	Usar
≤3 bits (incluidos modelos 1T comprimidos a 1.8–2 bits en una máquina)	No para decisiones	No	No	No

Qué configuración para qué trabajo

Configuración	Inspección en línea	Revisión con IA de falsas alarmas	Planificación y control de robot	Q&A y análisis
Detectores dedicados o modelos de anomalías (YOLO, RF-DETR, clase PatchCore), INT8/FP16, cualquier caja de borde NPU o GPU	● Sí: el juez principal, 2–12 ms por recorte	● Sí (clasificador afinado en la línea)	● Solo percepción	—
Modelos de visión 2–4B, NVFP4/AWQ/FP8 con codificador de visión FP16, en Thor/T4000 o DGX Spark	● Condicional: recomprobar recortes señalados por AOI a 0.2–0.3 s cada uno	● Condicional, por debajo de 1 s tras afinar	● Columna vertebral de VLA (VLA 2–3B a ≈40–130 ms por fragmento)	● Débil
Modelos de visión 7–9B (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 o 4-bit del vendedor, en Thor, Spark o una tarjeta de 24–32 GB	● Condicional: ≈0.5 s por recorte señalado	● Sí: 0.5–1 s por llamada, el mejor valor	● Planificación y elección de habilidad (1–3 s), no en el bucle de control	● Q&A básico de SOP para uno o unos pocos usuarios
MoE 26–35B (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● No (demasiado pesado)	● Sí, como explicador o para casos difíciles	● Planificación	● Sí (8 usuarios en un Spark, ≈0.1 s primer token)

Configuración	Inspección en línea	Revisión con IA de falsas alarmas	Planificación y control de robot	Q&A y análisis
MoE 106–241B a 4 bits en un Mac Studio, RTX PRO 6000 o Spark	No	No	Planificación sin conexión	Sí, para uno o unos pocos usuarios
DeepSeek 671B o Kimi K2.x 1T a 4 bits en un Mac de 512 GB, servidor KTransformers o DGX Station	No	No	Solo generación de programas sin conexión	Condicional: 1–3 usuarios y prompts cortos; los prompts largos tardan 16–170 s
Cualquier cosa a ≤3 bits; GGUF Q4_K_M para trabajo de detalle de modelo pequeño; Ollama para muchos usuarios	No	No	No	No

Fuentes: los datos de TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research y KTransformers anteriores, y (pruebas de DGX Spark, pruebas de RTX PRO 6000, Hardware Corner sobre Atlas, Ultralytics YOLO26, RF-DETR) **pruebas de comunidad y vendedores, mayormente single source** .

Así que, a las dos preocupaciones habituales: **no tiene por qué ser lento** si el modelo se ajusta al trabajo (los modelos de visión pequeños y medianos en Thor o Spark cumplen todos los presupuestos de línea y revisión; los modelos 671B–1T se quedan fuera de la línea), y **la cuantización no tiene por qué hacerlo incorrecto** con FP8/INT8 por defecto, o 4-bit de receta de vendedor por encima de 7B validado en el conjunto dorado. Los riesgos mayores son la precisión industrial zero-shot de los modelos generales de unos 57 puntos y los errores de la pila de servicio. Da a las peticiones de inspección tamaños de lote fijos y su propia cola (temperatura 0 no es reproducible a nivel de bits bajo lotes (documentación de vLLM)), mantén un repuesto en caliente N+1 y haz que cada paso de la línea falle abierto.

Profesor y alumno

03 Profesor y alumno: los modelos de frontera pueden pre-etiquetar datos de inspección; para los brazos robóticos, los verdaderos profesores siguen siendo las personas

Candidatos a profesor de noviembre de 2026: la visión genérica se duplicó, faltan las puntuaciones de defectos industriales

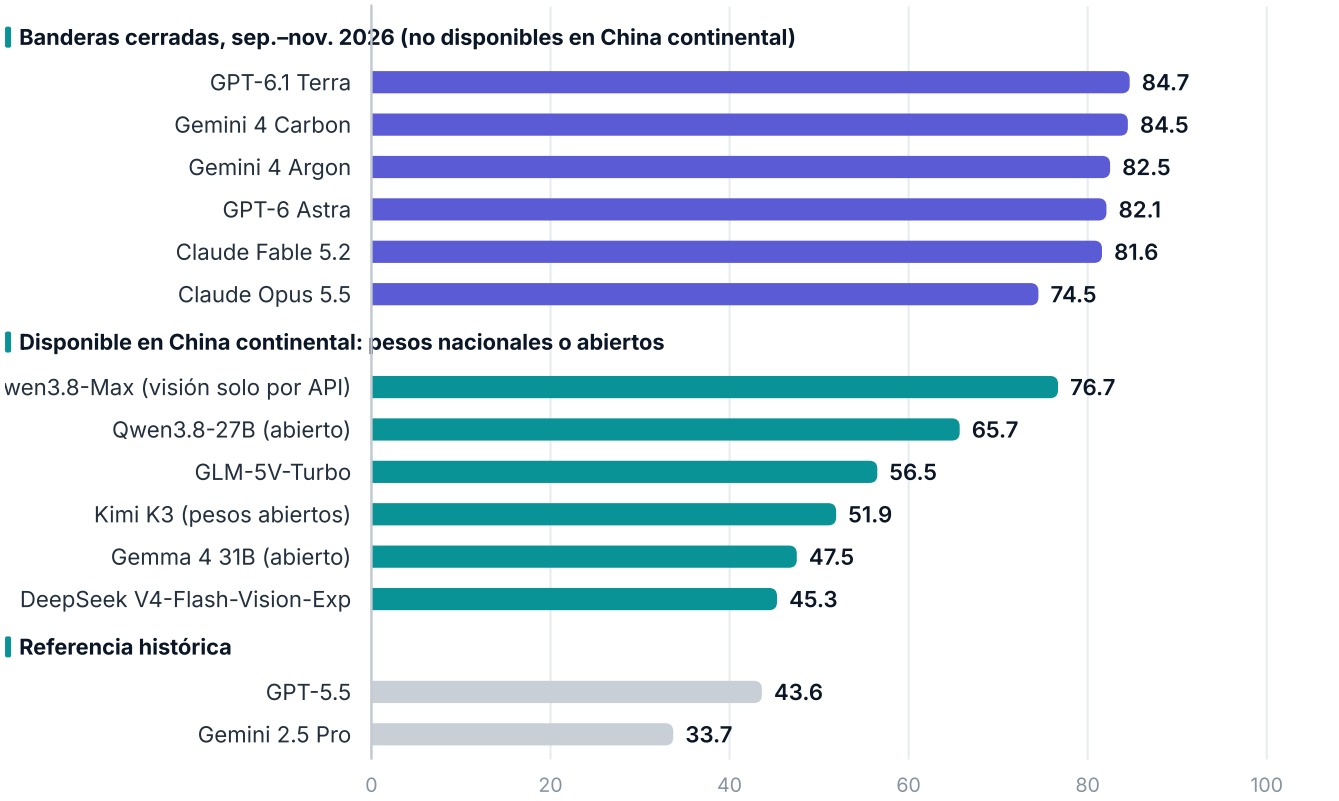
Todos los modelos insignia cerrados que podrían actuar como profesor en noviembre de 2026 aceptan imágenes: Claude Fable 5.2 (disponible en general el 9 de noviembre, $8/$40 por millón de tokens) y Opus 5.5 ($4/$20) (Anthropic, resumen de modelos de Claude) verified; GPT-6 Astra ($10/$50) y GPT-6.1 Terra ($1/$5) (precios de OpenAI); y Gemini 4 Argon y Carbon, abiertos por etapas (Google Carbon). Utilizables en China continental están Qwen3.8-Max (visión vía API) y modelos abiertos como Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp y Gemma 4 (blog de Roboflow, Hugging Face).

La laguna clave debe enunciarse con claridad: ninguno de estos modelos tiene una puntuación publicada en MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB o cualquier conjunto de datos de PCB/SMT. MMAD todavía no tiene tabla de clasificación (GitHub de MMAD), y los dos artículos de anomalías industriales más nuevos (agosto–septiembre de 2026) siguen usando Gemini 2.5 Pro y Claude Sonnet 4 como sus líneas base generales más nuevas (InspectorGPT, Anomaly-LR). Una búsqueda de texto completo en arXiv el 11 de noviembre encontró cero coincidencias para «Fable 5.2» o «Gemini 4 Carbon» en este contexto (API de arXiv) single source, ausencia de evidencia. Los sustitutos más cercanos:

- El modelo de frontera más nuevo con una puntuación MMAD es GPT-5.5: 69.95% en «defecto o no» y 76.20% promediado en siete tareas sobre un subconjunto de 500 imágenes, frente a 95.24% y 86.65% de expertos humanos (AnomalyClaw) single source; GPT-4o puntuó 68.63% y 74.92% en 2024 (MMAD, ICLR 2025), unos 1.3 puntos de progreso en dos años.
- En LDU-Bench, imágenes reales de revisión SEM de litografía IC y el único benchmark de electrónica con modelos de 2026, la clasificación bueno-o-malo es decente (macro F1 0.932 para GPT-5.4, 0.884 para GLM-5V-Turbo, 0.876 para Claude Opus 4.6, 0.504 para Gemma 4 31B), pero clasificar 11 morfologías de defectos alcanza solo 0.20–0.41 (LDU-Bench) single source.
- En las evaluaciones de visión independientes de Roboflow (actualizadas el 10 de noviembre de 2026), el mAP@50 de detección de objetos es 84.5 para Gemini 4 Carbon, 84.7 para GPT-6.1 Terra y 81.6 para Fable 5.2, frente a 33.7, 43.6 y 38.6 para Gemini 2.5 Pro, GPT-5.5 y Claude Opus 4.8 (Roboflow Vision Evals) single source, evaluador independiente, conjuntos de datos privados.

FIG 11 **Candidatos a maestro: las puntuaciones genéricas se duplicaron en una generación*Pero ninguna de las banderas 2026 tiene una puntuación publicada en un banco de defectos industriales.***

detección genérica Roboflow mAP@50



Bancos de defectos industriales (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): cero puntuaciones publicadas para estas band
las puntuaciones de detección genéricas se duplicaron en una generación, pero eso no es juicio de defecto; pruébelo en su propio conjunto de referencia

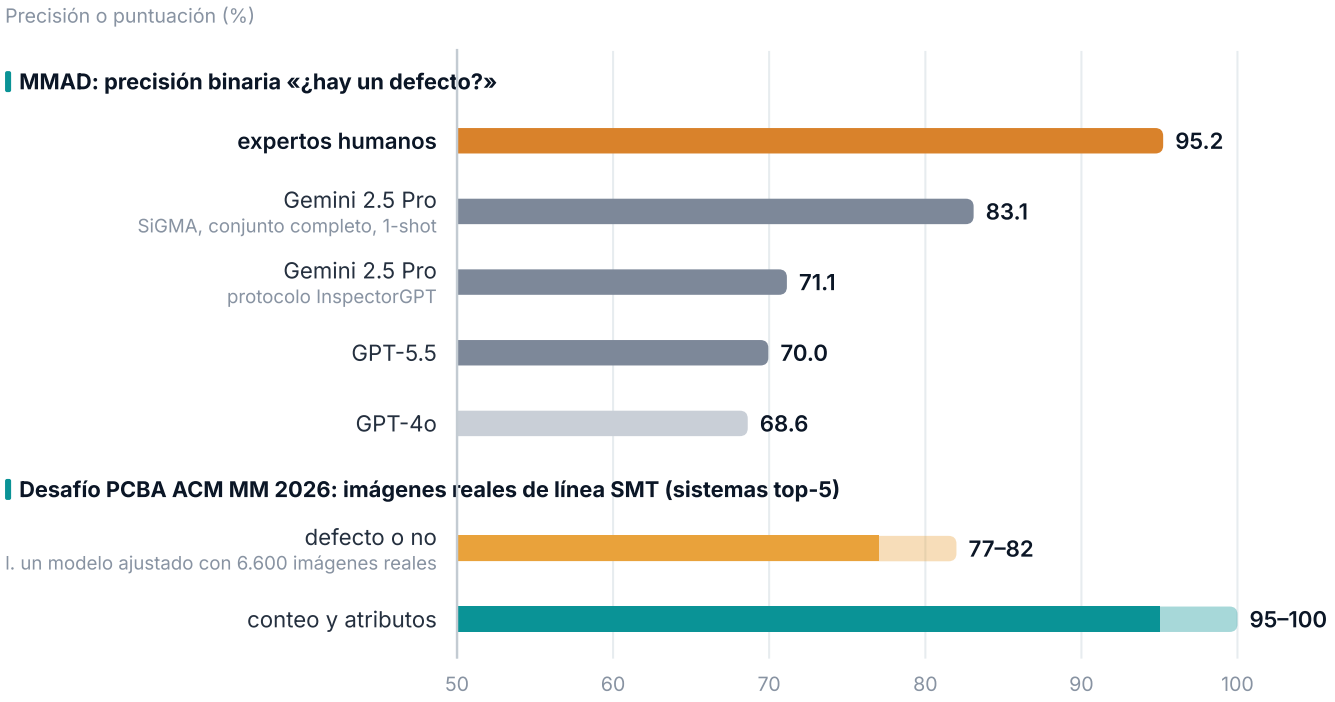
Fuente: Roboflow Vision Evals; notas de versión de modelos; páginas de MMAD, MVTec AD 2 y otros bancos (a 11 nov. 2026).

Candidato a profesor (nov. 2026)	Precio por M de tokens (entrada/salida)	mAP@50 / conteo de detección de Roboflow	Benchmarks de defectos industriales	Utilizable en China continental
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Ninguno	No
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Ninguno	No
GPT-6 Astra	$10 / $50	82.1 / 80.2	Ninguno	No
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Ninguno	No
Gemini 4 Carbon	$2 / $10 (promocional)	84.5 / 90.1	Ninguno	No
Qwen3.8-Max (solo visión vía API)	$2 / $6	76.7 / 81.1	Ninguno	Sí
Qwen3.8-27B (abierto)	Local	65.7 / 64.9	Ninguno	Sí
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	Triage LDU-Bench 0.884	Sí
Gemma 4 31B (abierto)	Local	47.5 / 51.4	Triage LDU-Bench 0.504	Sí (pesos locales)
Referencia: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	Binario MMAD 69.95% / 68.63%	No

Candidato a profesor (nov. 2026)	Precio por M de tokens (entrada/salida)	mAP@50 / conteo de detección de Roboflow	Benchmarks de defectos industriales	Utilizable en China continental
Expertos humanos	—	—	Binario MMAD 95.24%	—

Los artículos más nuevos todavía prueban la generación anterior, pero apuntan en la misma dirección. En el conjunto completo MMAD, los expertos humanos puntúan 95.2% en defecto-o-no y Gemini 2.5 Pro 83.1% (SiGMA), y sin embargo el mismo modelo puntuó solo 71.14% en otro artículo (InspectorGPT): el protocolo de prueba mueve la puntuación tanto como una generación de modelo. En el desafío VAND 4.0 de CVPR 2026, los modelos de visión-lenguaje listos para usar alcanzaron un pico de 61.88 AP frente a 90.84 de los métodos supervisados, y los organizadores concluyeron que «no pueden sustituir actualmente a los clasificadores afinados» (VAND 4.0) single source. En el desafío PCBA de ACM MM 2026, los cinco primeros puntuaron 76.92–82.47 en defecto-o-no mientras que el tipo de componente y el conteo de pines alcanzaron 95–99.5 (arXiv 2609.21276) single source: reconocer es mucho más fácil que juzgar.

FIG 12 **«¿Hay un defecto?» sigue siendo difícil para los modelos generales** ***El mejor modelo general queda a más de 12 puntos de los expertos humanos en MMAD; los sistemas ajustados lo hacen mejor en imágenes reales de línea.***



Fuente: MMAD; SiGMA; InspectorGPT; desafío PCBA ACM MM 2026.

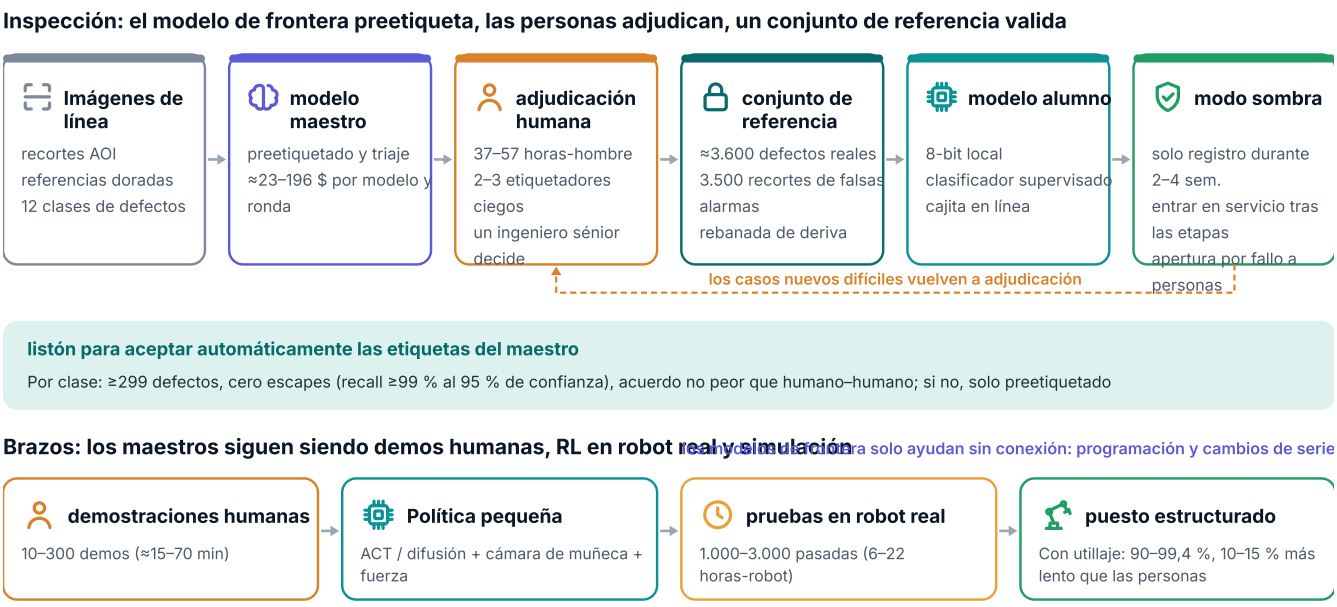
Se siguen tres conclusiones. Ver y contar se duplicaron en una generación, pero juzgar la anormalidad avanza despacio, depende mucho del protocolo y viene con sesgos sistemáticos; cuánto ganan los nuevos modelos solo puede medirse con las propias imágenes de una planta. Dentro de su distribución de entrenamiento, los especialistas 7B afinados siguen superando a los modelos generales (Anomaly-LR, UniPCB). Y los datos reales de línea colapsan las puntuaciones: en paneles electrónicos ensamblados, el AUROC a nivel de imagen cayó de 97–98 en MVTec AD a 31.5–78.6 (arXiv 2606.01992) single source. Para las empresas del continente, el profesor disponible más fuerte, Qwen3.8-Max, se sitúa unos 8 puntos por debajo de la frontera global.

Ejecutar docenas o cientos de rondas: el dinero no es el problema, la calidad de etiquetado sí

El dinero no es el problema. A precios de lista de noviembre de 2026, 1,000 imágenes de un megapíxel cuestan unos $3 (GPT-6.1 Terra) a $32 (GPT-6 Astra) por ronda, y cien rondas sobre 100,000 imágenes unos $32,000–135,000 antes de descuentos por volumen estimate (documentación de visión de Claude, guía de visión de OpenAI, Roboflow Vision Evals). Para etiquetas simples, una persona cuesta aproximadamente lo mismo que Terra; la ventaja del profesor es la velocidad y la escala, no el precio. Los NDA de los clientes suelen mantener las imágenes de producto dentro de la planta, de modo que el conjunto dorado se etiqueta en el sitio.

El obstáculo real es la calidad, y más rondas no corrigen los errores sistemáticos. Roboflow ejecuta cada tarea tres veces y las puntuaciones se mueven solo ±0.1–1.4 puntos (Roboflow), de modo que un modelo repite en su mayoría sus errores y el voto mayoritario no puede salvarlo. Las rondas ayudan solo cuando cada una añade información: una o dos buenas imágenes de referencia añaden unos 3–4 puntos, más no añaden nada (AD-Copilot); hacer que un modelo discuta contra una buena referencia durante varios turnos añade 6.23 puntos de AUROC entre dominios para GPT-5.5 y 2.95 en MMAD, con pico tras unas tres llamadas (AnomalyClaw) single sources. «Un alumno entrenado con etiquetas del profesor se acerca a las etiquetas humanas» solo se ha demostrado en imágenes naturales: una ResNet-101 entrenada con etiquetas de Qwen3-VL de COCO y Objects365 alcanzó el 95.7–99.4% del resultado de las etiquetas humanas, y los autores dicen que los datos de dominio no están probados (arXiv 2602.20972) single source.

FIG 13 **Dos canalizaciones maestro–alumno** ***Para la inspección, un modelo de frontera preetiqueta y las personas adjudican; para los brazos robóticos, los verdaderos maestros siguen siendo las demostraciones humanas.***



Fuente: Síntesis del autor (capítulo 3); costes y tamaños de muestra son estimaciones.

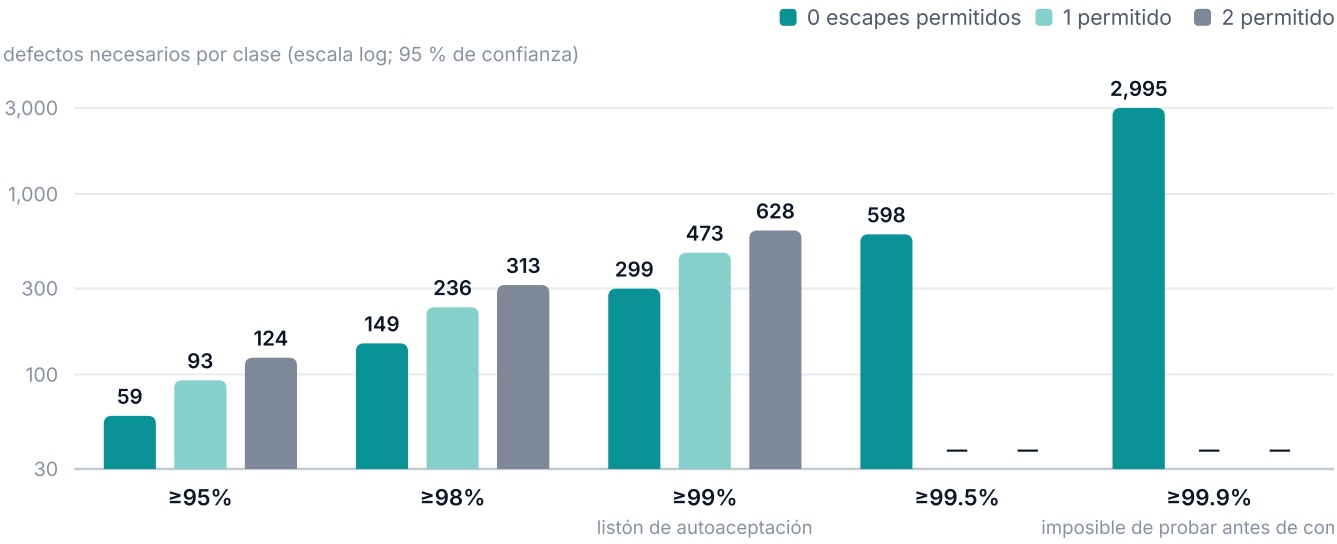
Un protocolo viable estimate:

1. **Reúne el conjunto dorado del propio archivo del AOI de la planta y de las placas NG conservadas**, estratificado en 12 clases verdaderas de defectos (faltante, desplazada, lápida, polaridad, pieza incorrecta, puente/cortocircuito, soldadura insuficiente, soldadura excesiva, bola de soldadura, terminal levantado/abierto, pieza dañada, material extraño) y cuatro causas de falsa alarma (sombra, flux, serigrafía, tolerancia), más una «rodaja de deriva» de al menos 300 defectos de una segunda línea o con diferente iluminación, color de placa o números de pieza.

2. **Etiqueta a ciegas con dos o tres inspectores formados en IPC-A-610**, con un ingeniero sénior decidiendo y rayos X, prueba eléctrica o secciones transversales resolviendo las muestras ambiguas. Acepta una definición de clase solo si el acuerdo entre inspectores alcanza κ ≥0.75 (≥0.90 para clases críticas); el acuerdo humano en 11 morfologías de defectos de electrónica en LDU-Bench fue solo 0.727, de modo que algunas clases necesitarán fusionarse o redefinirse. El acuerdo de atributos estilo AIAG (efectividad ≥90%, fallo ≤2%, falsa alarma ≤5%, κ >0.75; >0.90 para liberación) es un mínimo sensato (Fabrico, iFactory).
3. **Dimensiona con límites binomiales exactos (Clopper–Pearson) unilaterales al 95%** (intervalos de confianza binomiales, regla del tres:

Sensibilidad a demostrar	0 fallos permitidos	1 permitido	2 permitidos
≥95%	59 defectos	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **Qué tamaño debe tener el conjunto de referencia*Probar un 99 % de recall con cero escapes exige 299 defectos por clase; el 99,9 % no puede probarse antes de comprar.***



Fuente: Intervalos de confianza binomiales exactos (95 %).

Sesenta defectos por clase bastan para un cribado de pasa/no pasa; se necesitan 300 para decidir la aceptación automática. El conjunto completo, unos 3,600 defectos más 3,500 recortes de falsa alarma, tarda unas 37–57 horas de trabajo en etiquetar y arbitrar, y ejecutar un candidato a profesor sobre él cuesta $23–196; el coste real son de dos a seis semanas de recolección, fijadas por la clase más rara estimate. Envía recortes de AOI de un solo componente, no placas enteras, con el objetivo de al menos 42–48 píxeles de ancho (arXiv 2610.09313).

1. **Acepta las etiquetas del profesor automáticamente solo para una clase con cero fallos en ≥299 defectos dorados**, acuerdo modelo-verdad no peor que el acuerdo humano-humano, y al menos el 50% de las falsas alarmas resueltas. Las clases críticas como cortocircuitos, piezas faltantes, polaridad y piezas incorrectas necesitan 99.5–99.9% (598–2,995 defectos); de lo contrario, todo lo que la IA no pueda resolver con alta confianza va a una persona. Una tasa de fuga por debajo del 0.1% de los defectos no puede demostrarse antes de la compra (2,995 defectos con cero fallos; 0.01% requeriría 29,956), de modo que la línea mantiene muestreo

humano de las llamadas resueltas por IA, ICT/FCT aguas abajo como respaldo y SPC sobre fugas. Las etiquetas verificadas se destilan después en un detector o clasificador convencional y se revalidan en cada producto nuevo.

Los términos y las reglas de datos también deciden el profesor. Anthropic prohíbe usar sus servicios para entrenar modelos competidores (términos comerciales de Anthropic); OpenAI permite que las salidas entrenen clasificadores internos que no se vendan a terceros (términos comerciales de OpenAI); DeepSeek permite explícitamente la destilación (términos de DeepSeek) [términos primarios]. Ninguno define «modelo competidor», de modo que el uso comercial necesita revisión legal, y las reglas de datos de los clientes suelen mantener las imágenes en el sitio. El valor por defecto más seguro es un profesor de pesos abiertos ejecutado localmente, con modelos de frontera en la nube solo cuando el cliente lo permite por escrito.

Alumnos de inspección: pequeños y rápidos; los cuellos de botella son la óptica, los datos y la validación

El alumno no es el cuello de botella: YOLO26n tarda 3.80 ms por imagen en INT8 en una Orin Nano Super de $249–399 (Ultralytics) [prueba de vendedor], NVIDIA informa de un 98.51% en tres clases de defectos de PCB tras adaptar un modelo fundacional sobre unas 600 imágenes (NVIDIA) vendor claim, y los defectos sintéticos añaden solo 0.9–2.5 puntos de mAP (arXiv 2605.04635). Las limitaciones son la óptica, los datos de línea, la validación y las licencias: Ultralytics YOLO es AGPL-3.0 o con licencia empresarial (Ultralytics), de modo que los productos cerrados necesitan una licencia o un detector Apache-2.0 como RF-DETR.

Alumnos de robot: las estaciones estructuradas están cerca del grado de producción; el trabajo flexible y dinámico no

En noviembre de 2026 hay tres tipos de profesor de robot, y solo uno puede convertirse en alumno local: VLA abiertos y afinables (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Los modelos de robot cerrados (π0.6/π0.7 mediante una API beta, Gemini Robotics 2 para probadores de confianza) no pueden desplegarse en el sitio (The AI Insider, DeepMind). Los modelos generales de frontera son lentos y malos en manipulación fina: en la prueba de Artificial Analysis puntuaron 46–69 con un 28–45% de éxito de agarre y unos 30 minutos por tarea, «demasiado lentos y caros para cualquier uso de producción» (AA-Robotics); GPT-6 Astra logró una ranura de precisión solo 2 veces de 20 (RoboCurve); y mover articulaciones directamente da 0–5.5% en LIBERO a 0.2–0.4 Hz frente a los ≈83 Hz necesarios (Anthropic) verified en tres fuentes. El 64.9% de Fable 5.2 en el benchmark simulado RoboDojo es una ejecución interna (Anthropic) vendor claim.

Los estudios de terceros desde octubre coinciden: alrededor de 33–44% en LIBERO-10 para los tres agentes líderes (arXiv 2610.30051); agentes que afirman éxito falsamente en el 24–39% de las subtareas (arXiv 2611.08209); y 104 de 108 tareas cuasi estáticas sobre robot real para Opus 5.5, pero a 280 segundos cada una y sin inserción ni ensamblaje (AMP1) single sources. Los modelos de frontera pertenecen al entorno sin conexión: escribir programas, orquestar tareas y redactar criterios de éxito.

La calidad del alumno tiene una ventaja clara: buena dentro de la distribución de entrenamiento, colapsando fuera de ella. π0.7 «a menudo supera el 90%» en tareas vistas y alcanza el 60–80% en no vistas (artículo de π0.7) [investigación de vendedor]; ContactDP alcanza 97.8% en conectores entrenados pero 28.6–66.7% en no vistos (ContactDP); pequeños cambios de cámara o pose hacen caer el 95% por debajo del 30% en LIBERO-Plus (LIBERO-Plus, CVPR 2026); el agarre aprendido de cintas en movimiento alcanza un pico de 58.89% (ForeTime-VLA) single sources, un trabajo que el seguimiento por visión clásico lleva años haciendo en producción.

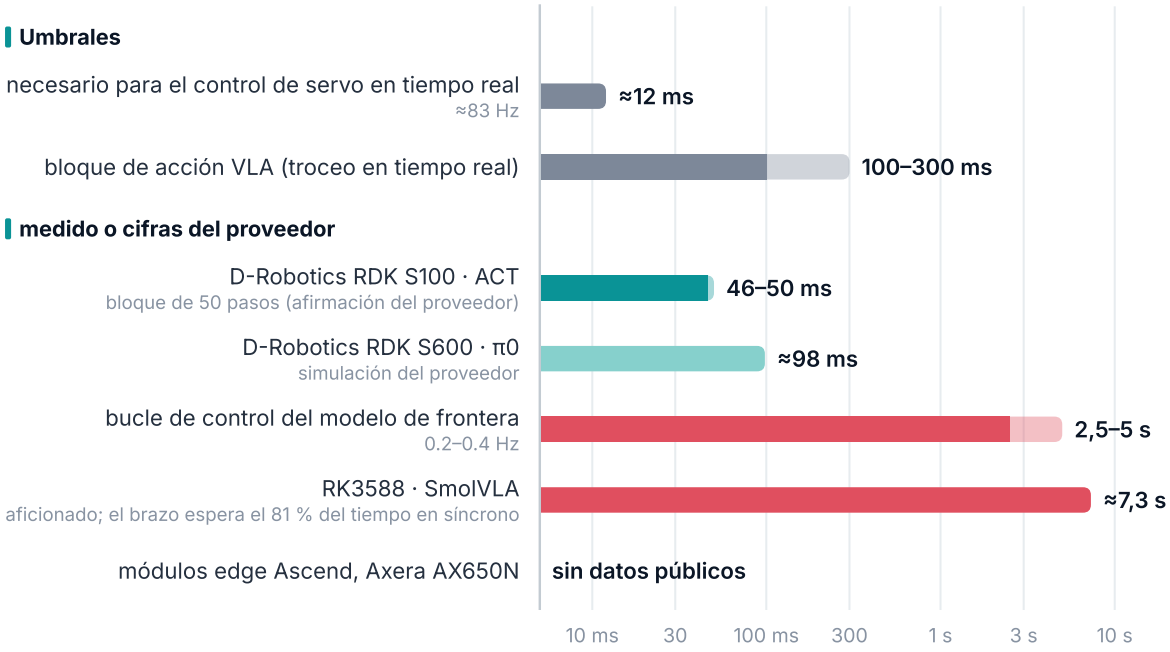
La evidencia cercana a producción proviene toda de estaciones estructuradas, demostraciones humanas y corrección sobre robot real. Neuromeka mantuvo una máquina de estados y movimiento pre-enseñado y aprendió solo agarre, inserción de cable y soldadura, cada uno con menos de 20 minutos de datos: 322 de 324 en producción, un 12.8% más lento que una persona (Neuromeka) single source, vendor. Fraunhofer alcanzó 92.8% en cuatro conectores con 68 minutos de demostraciones (Fraunhofer); HIL-SERL de Berkeley tuvo éxito 100 de 100 veces en inser-

ción de RAM, SSD y USB tras 1–2.5 horas de RL sobre robot real (HIL-SERL); AgiBot ejecutó 18–20 segundos por paso dentro de un takt de 20 segundos en una línea de prueba de tabletas (IThome) vendor claim .

La destilación compra velocidad, no generalización: π0.5 corre a 0.70 Hz en Jetson Orin en bruto y 10.30 Hz tras planificación asíncrona (EagleVLA); la destilación de capas lo lleva de unos 3 Hz a unos 10 Hz (Shallow-π) verified . Un alumno hereda solo lo que el profesor cubrió; cada producto nuevo sigue necesitando datos de tarea.

FIG 15 **Latencia de política edge** ***Una política ACT pequeña corre en ≈50 ms en una placa edge; un VLA portado a un chip aficionado tarda 7,3 s.***

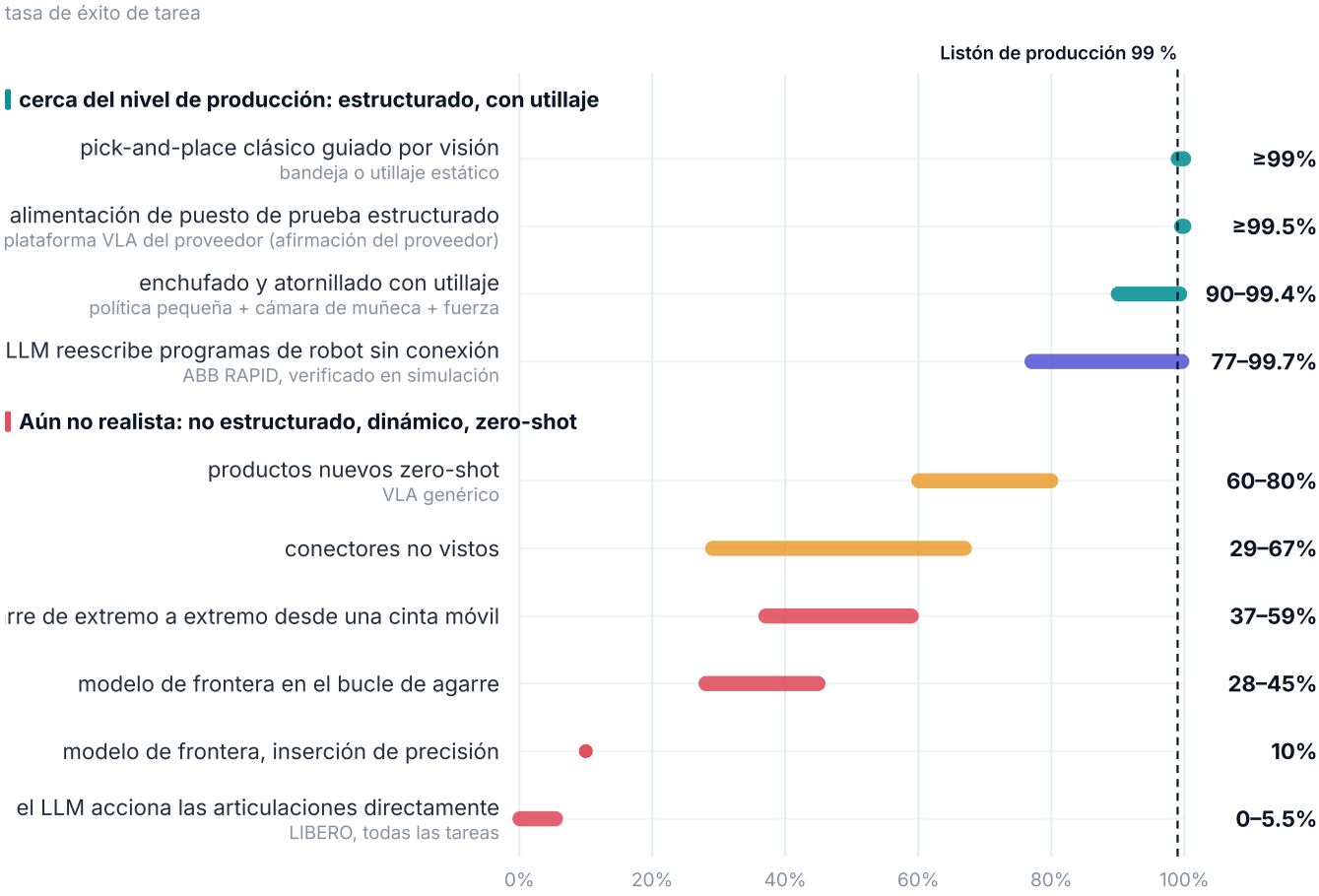
latencia de inferencia por bloque de acción o ciclo de control (escala log)



Fuente: Documentación y foro de D-Robotics (cifras del proveedor); ports comunitarios de GitHub; investigación de Anthropic.

En chips de borde chinos, la única política aprendida medida es una política ACT en el RDK S100 de D-Robotics a 46–50 ms por fragmento de 50 pasos, sin tasa de éxito publicada (documentación de D-Robotics RDK S) vendor claim ; un portado aficionado de SmolVLA en un RK3588 tarda unos 7.3 s por fragmento (GitHub). La cuantización no compra velocidad gratis: ninguna compilación de π0.5 a 4 bits fue más rápida que 16 bits a tamaño de lote 1 (LoopQ). Las demostraciones son caras (Reuters informó de unos 300 intentos por movimiento utilizable para un teleoperador novato (Notebookcheck)), y la validación cuesta más: demostrar una tasa de fallo ≤0.1% al 95% de confianza requiere unos 3,000 éxitos consecutivos, 17–22 horas-robot estimate . Ninguna política de manipulación aprendida tiene certificación de seguridad.

FIG 16 **Alumnos robot: cerca del nivel de producción solo cuando el puesto está estructurado*Los puestos con utillaje alcanzan 90–99 %+; las tareas no estructuradas, dinámicas y zero-shot aún no son realistas.***



Fuente: Artículos y material de proveedores citados en el capítulo 3; algunas cifras son afirmaciones del proveedor.

Tarea de electrónica	Alumno viable	Éxito esperado	Velocidad frente a una persona	Coste por cambio de modelo	Veredicto 2026–2028
Pick-and-place de alta mezcla, kitting, alimentación de máquinas (bandeja o utillaje estático)	Percepción aprendida más movimiento programado o planificado	≥99% con guiado por visión clásica	Igual que las estaciones convencionales	Horas	Viable ahora
Inserción de conector, FFC y cable con utillaje, atornillado, soldadura	Máquina de estados + pequeña política BC/difusión/ACT + cámara de muñeca + fuerza/par (idealmente tacto) + corrección humana	90–99.4%	10–15% más lento	10–300 demos (≈15–70 min) + 1,000–3,000 ejecuciones de validación (6–22 horas-robot)	Viable, intensivo en I+D; meses para la primera estación
Atención de estación de prueba estructurada y clasificación OK/NG	Plataforma VLA de vendedor + gemelo digital + RL sobre robot real	≥99.5% vendor claim, en conflicto	18–20 s por paso, dentro de un takt de 20 s	≤4 h de reentrenamiento + 15 min de calibración	Viable en una plataforma de vendedor
Programación de cambios de modelo	El LLM escribe o edita programas de robot sin conexión, verificados en simulación	77–99.7% en reescrituras ABB RAPID	Sin efecto en el takt	Minutos a horas por programa	Viable como herramienta de ingeniería

Tarea de electrónica	Alumno viable	Éxito esperado	Velocidad frente a una persona	Coste por cambio de modelo	Veredicto 2026–2028
Recogida aprendida de extremo a extremo desde cintas en movimiento	Alumno VLA	37–59%; ≤37% en cintas rápidas	—	≥458 demos por configuración	**No antes de 2028; usar seguimiento clásico más percepción aprendida**
Productos nuevos sin datos nuevos (zero-shot)	VLA genérico	60–80%; conectores no vistos 29–67%	—	—	**No realista**
LLM de frontera en el bucle moviendo el brazo	Agentes GPT-6, Claude, Gemini	Agarre 28–45%; inserción de precisión 10%	≈30 min o 280 s por tarea frente a ≈20 s	—	**No realista**
Arneses de cableado de rutas libres y ensamblaje flexible de varios pasos	—	Sin evidencia de producción en 2026	—	—	**No antes de 2028**

Fuentes: los estudios anteriores y (Universidad de Passau sobre reescritura RAPID, GraspVLA, Komax) **mayormente single source y datos de vendedor** . En 2026–2028, el valor de un modelo grande para los robots es como prior pre-entrenado y programador sin conexión, no como algo que «enseña» a un alumno. Todos los casos cercanos a producción que encontré (Neuromeka, AgiBot en Longcheer, Fraunhofer, ContactDP) dependen de demostraciones humanas, sensado de fuerza o tacto, simulación o gemelos digitales y RL supervisado. Ninguno depende de etiquetas de un modelo de frontera.

Costes y retorno

04 Costes y retorno: la revisión con IA se amortiza en meses, la alimentación de máquinas en años, los brazos de retirada nunca

Coste laboral: EE. UU. es unas 5.5 veces Shenzhen, Taiwán unas 1.5 veces, la mayor parte del Sudeste Asiático menos de la mitad

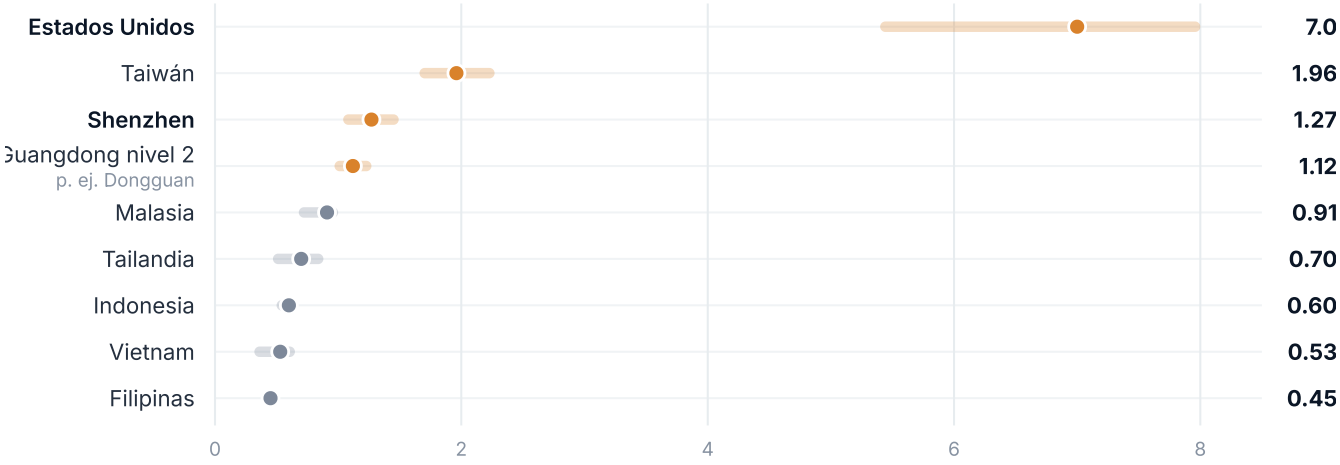
El retorno de sustituir mano de obra empieza por lo que cuesta la mano de obra. La guía salarial oficial de Shenzhen sitúa la mediana de 2024 de un inspector de calidad de fabricación en **RMB 74,074** al año; los salarios son el 86.92% del coste laboral total en la fabricación de electrónica, de modo que el coste totalmente cargado es de unos RMB 85,000. La encuesta de 2025 de JETRO sitúa el coste anual de un trabajador de fábrica chino en $12,667, unos RMB 85,100, dentro del 0.2% (guía salarial de Shenzhen, JETRO 2025) verified .

Región	Coste anual totalmente cargado de un inspector o manipulador (bajo / base / alto)	Base
China continental, Shenzhen	RMB 70k / **85k** / 100k ($10.4k / **12.7k** / 14.9k)	Mediana oficial × 1.15; coincide con JETRO verified
Áreas salariales de nivel 2 de Guangdong (Dongguan etc.)	RMB 65k / **75k** / 85k	Mediana de Guangdong 2023 × 1.15 más dos años de subidas estimate
Delta del río Yangtsé (Suzhou, Kunshan, Wuxi)	Como Shenzhen	Sin base para mano de obra de inspección más barata; el salario mínimo más alto de Jiangsu está un 5.6% por encima del de Shenzhen estimate
Taiwán	$16.6k / **19.6k** / 22.7k (NT$534k / **630k** / 729k)	Encuesta salarial ocupacional del Ministerio de Trabajo más ≈1.18 de costes legales; JETRO como extremo alto estimate
Malasia	$6.8k / **9.1k** / 10.0k	JETRO y salarios mínimos estimate
Tailandia	$4.7k / **7.0k** / 8.8k	Igual
Indonesia	$5.0k / **6.0k** / 6.3k	Igual
Vietnam	$3.2k / **5.3k** / 6.5k	Igual
Filipinas	$4.0k / **4.5k** / 4.9k	Igual
Estados Unidos	$54k / **70k** / 80k	Mediana de inspector BLS mayo 2025 $48,570 ÷ salarios al 65.5% de la compensación estimate

Fuentes: (JETRO 2025, guía salarial de Guangdong, oficina de RRHH de Wuxi, encuesta salarial ocupacional de Taiwán, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **Lo que cuesta al año un trabajador de inspección o manipulación** ***A plena carga, EE. UU. es ≈5,5× Shenzhen y Taiwán ≈1,5×.***

coste anual cargado por trabajador de CQ o manipulación (10k $; punto = base, barra = bajo–alto)



Fuente: Guía salarial oficial de Shenzhen, Ministerio de Trabajo de Taiwán, JETRO, BLS; estimaciones del autor.

Tres correcciones fijan la base. Primera, Shenzhen elevó el suelo de la cotización de pensiones de RMB 4,775 a 5,521 al mes en julio de 2026 (Shenzhen Bendibao) single source, de modo que un inspector de Shenzhen en 2026 más probablemente cuesta RMB 80–90k. Segunda, un puesto no es una persona: dos turnos de 12 horas necesitan unas 2.0 personas (seis días a la semana) a 2.4 (todo el año) en el delta del río Perla, unas 2.7–3.1 en el Sudeste Asiático bajo límites de horas extra, y 4.2–4.5 para 24×7 en EE. UU. y Taiwán estimate. La revisión con IA suele permitir que un revisor cubra de dos a cuatro líneas, de modo que el ahorro realista es **0.5–1.5 personas a tiempo completo por línea, no una persona por línea por turno**. Tercera, los compradores de Taiwán, Vietnam y Malasia no persiguen solo el coste: en la encuesta de JETRO, el 61.1% de las empresas de Taiwán, el 48.2% de Vietnam y el 44.9% de Malasia dijeron que contratar se había vuelto más difícil en los últimos dos años, frente al 18.6% en China continental (JETRO 2025) single source.

Precios de equipos: cuatro bases que no deberían promediarse

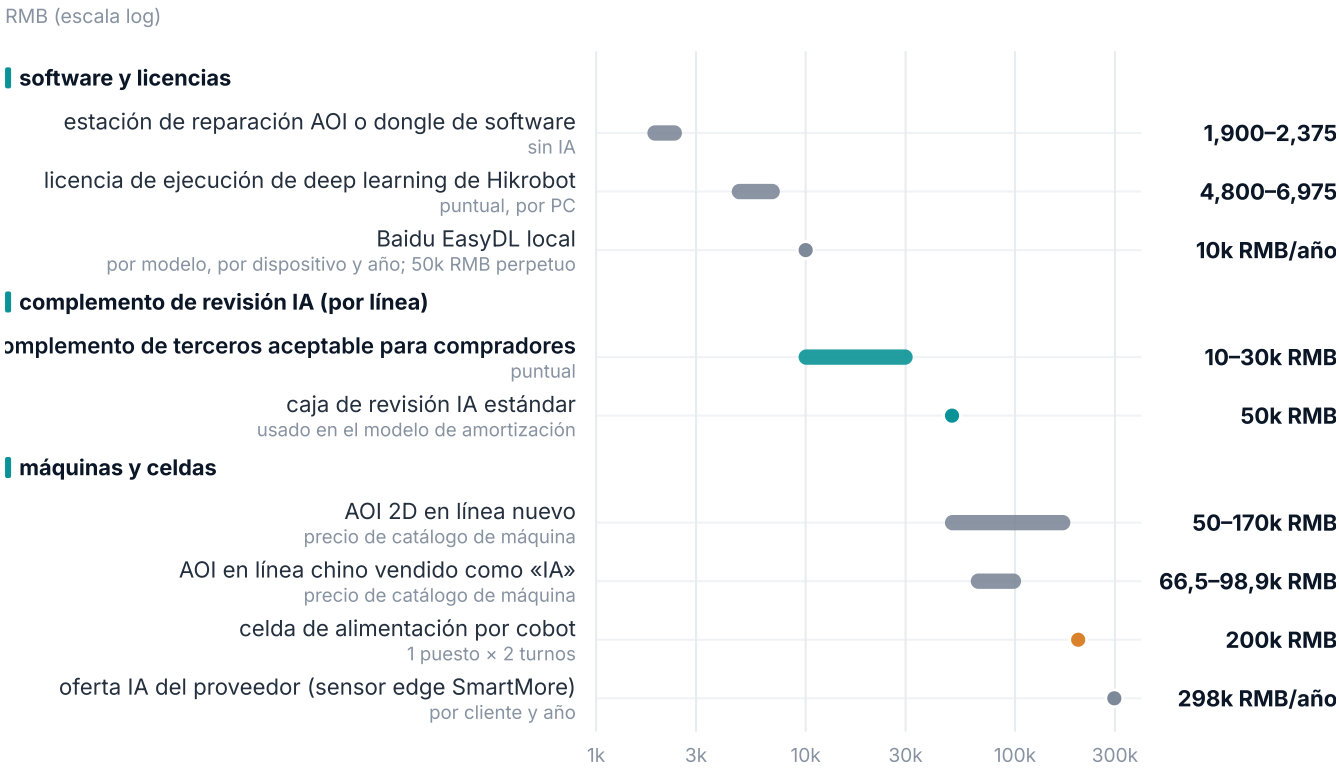
Los precios de AOI discrepan sobre todo porque miden cosas distintas. Los listados en línea (noviembre de 2026, nuevos, usados y de alquiler mezclados) van de unos RMB 17–80k para 2D sin conexión o de escritorio, RMB 50–170k para 2D en línea y RMB 80–475k para 3D en línea, hasta RMB 700k para 3D coreano (AOI en línea de 1688, AOI sin conexión de 1688, AOI 3D de 1688) single source, precios de lista. Una licitación de noviembre de 2025 limitó un AOI 3D más un SPI 3D a RMB 800k (aviso de licitación). El folleto de Sinic-Tek implica RMB 428–512k por AOI 3D en 2019–2022, unos RMB 340–470k según su informe de 2025 (folleto de Sinic-Tek, memoria anual 2025 de Sinic-Tek) primary + estimate. Un vendedor extranjero estima $3,000–20,000 para 2D de escritorio, $30–70k 2D en línea, $50–90k 3D de entrada y $100–200k+ 3D de gama alta, con la misma máquina en línea a $30–50k en China (Averroes) [estimación de vendedor de un vendedor de software de IA competidor].

La propia revisión con IA no tiene precio público en RMB; los vendedores cotizan caso por caso tras una evaluación gratuita en el sitio (Sohu sobre Hengtianyi), y las búsquedas de listados de comercio electrónico, licitaciones, artículos de vendedores y medios taiwaneses no arrojaron ninguna cotización pública (1688, 51SMT). Los anclajes: unos €27k por línea implicados por el caso de Siemens estimate; $21,924 de lista para el módulo de red de revisión remota de un distribuidor estadounidense (Manncorp) [precio de lista de vendedor]; unos RMB 298k por cliente al año por los sensores de IA de borde de SmartMore en 2025 (36Kr) single source. Un AOI 2D en línea nacional vendido como «IA» se lista a RMB 66.5k (AOI en línea de 1688), y un búfer de NG para reinspección manual a unos RMB 50k (revisión AOI de 1688). Así que el techo para un complemento de revisión de terceros es de unos RMB 10–30k

por línea, o una cuota anual similar; por encima de eso, el comprador preferiría sustituir la máquina por un AOI nacional con IA integrada estimate.

Ancla	Precio (RMB)	Qué compra	Fuente
Estación de reparación de AOI o dongle de software (TRI antiguo, Zhenhuaxing)	≈1,900–2,375	Clave de licencia, sin IA	1688 [precio de lista]
Tiempo de ejecución de visión de aprendizaje profundo general (dongle VisionMaster de Hikrobot)	≈4,800–6,975 pago único por PC	La plataforma; el cliente construye y mantiene los modelos	1688 [precio de lista]
Despliegue en local de Baidu EasyDL	10,000 por modelo por dispositivo por año, o 50,000 perpetuo	Servicio de inferencia empaquetado	documentación de Baidu AI [página de precios de 2022]
AOI en línea nacional vendido como «IA»	66,500–98,900 por máquina	Sustitución de máquina	1688 [precio de lista]
Sensores de IA de borde de SmartMore	≈298,000 por cliente al año	Solución de vendedor en varias máquinas	36Kr single source

FIG 18 **La escalera de precios** ***De unos pocos miles de RMB de licencia de software a 200–300k RMB por una celda robótica o una oferta de IA del proveedor.***



Fuente: Anuncios de 1688, documentación de Baidu AI (tarifa 2022), 36Kr, una entrevista industrial (sin verificar).

Una cuota anual estilo Baidu, con un servidor de inferencia para una a cuatro líneas y uno a cuatro modelos cada uno, implica una suscripción de RMB 2.5–40k por línea al año, realísticamente unos RMB 5–10k; al 20% de una licencia perpetua, la suscripción y la compra empatan en el quinto año, una estructura que vale la pena copiar estimate. Los compradores están acostumbrados a cuotas de software casi nulas: todos los ingresos «otros» de Sinic-Tek (repuestos, servicio, alquiler, accesorios) fueron el 7.65% de las ventas de 2025, unos RMB 2,300–3,100 por máquina instalada al año (memoria anual 2025 de Sinic-Tek) primary + estimate. Una suscripción de IA de RMB 5–10k

por línea al año es de dos a cinco veces lo que los compradores gastan hoy en software de AOI, de modo que tiene que justificarse por las horas de revisión que libera (unos RMB 75–90k por persona a tiempo completo al año), no por comparación con las cuotas de software existentes.

Para los brazos robóticos, el precio medio de cobot de China cayó a unos RMB 58,600 en 2025 (GGII) estimate, y el promedio de seis ejes de Dobot cayó un 19.1% ese año a medida que los modelos más baratos ganaban cuota (Huxiu) verified, aunque los precios de lista de los mismos modelos subieron un 5–15% en 2026 (Sina Finance). En EE. UU., un FANUC CRX-10iA se listaba a $51,434 (Vention) y las recopilaciones sitúan un UR7e en $30–48k (cobot-cost) estimate. El brazo es solo el 40–50% de una célula estadounidense (Standard Bots); en China, cámaras, pinzas, seguridad e ingeniería fácilmente cuestan de dos a cuatro veces un brazo de RMB 40k, de modo que presupuesta el coste de ingeniería absoluto en lugar de un múltiplo estimate.

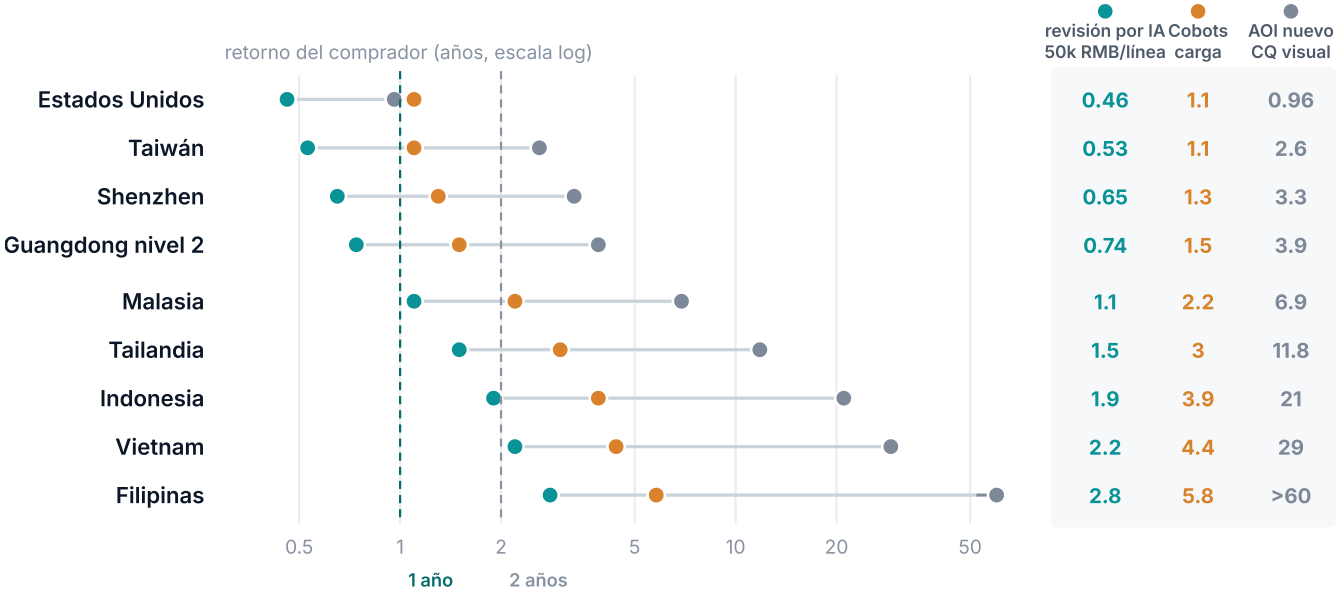
Retorno: revisión en meses, alimentación en años, retirada nunca

Retorno simple = inversión ÷ (personas ahorradas × coste laboral cargado − coste de funcionamiento anual), para una línea de dos turnos sin subvenciones ni aranceles. El equipo de Taiwán se valora a 1.3 veces China, Vietnam, Malasia y Tailandia a 1.15 veces, Indonesia y Filipinas a 1.2 veces, y EE. UU. a precios extranjeros estimate.

Escenario	Inversión (China / EE. UU.)	Personas ahorradas	Coste de funcionamiento anual
(a) Caja de revisión con IA o caja de cámara con IA	RMB 50k / $30k por línea	1.0 por línea	15%
(b) AOI 2D en línea nuevo sustituyendo controles visuales manuales (queda 0.5 revisor por turno)	RMB 200k / $60k	1.0	12%
(c) Célula de carga con cobot (una estación × dos turnos)	RMB 200k / $140k	2.0	8%
(d) Célula integrada de detección y recogida (inspección 1.0 + recogida 0.07)	RMB 300k / $200k	1.07	8%
(e) Un brazo robótico añadido solo para retirar placas NG	RMB 150k / $100k	0.07	8%

Región	(a) Revisión con IA	(b) AOI nuevo	(c) Célula de cobot	(d) Detección y recogida	(e) Brazo de retirada
Shenzhen	0.65 años (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	Nunca
Guangdong nivel 2	0.74	3.9	1.5	5.3	Nunca
Taiwán	0.53	2.6	1.1	3.6	Nunca
Malasia	1.1	6.9	2.2	9.2	Nunca
Tailandia	1.5	11.8	3.0	15	Nunca
Indonesia	1.9	21	3.9	25	Nunca
Vietnam	2.2 (1.6–4.5)	29	4.4	34	Nunca
Filipinas	2.8	>60 (coste de funcionamiento ≈ ahorro)	5.8	≈100	Nunca
Estados Unidos	0.46	**0.96**	**1.1**	3.4	Nunca

FIG 19 **Amortización por región** ***La revisión IA se amortiza en meses en EE. UU., Taiwán y Shenzhen y en 1–3 años en el sudeste asiático; la alimentación por cobot, en años.***



Un brazo robótico solo para extraer placas NG nunca se amortiza, en ninguna parte
Con un rendimiento de primer paso superior al 90 %, la extracción ocupa solo 0,003–0,14 FTE por turno

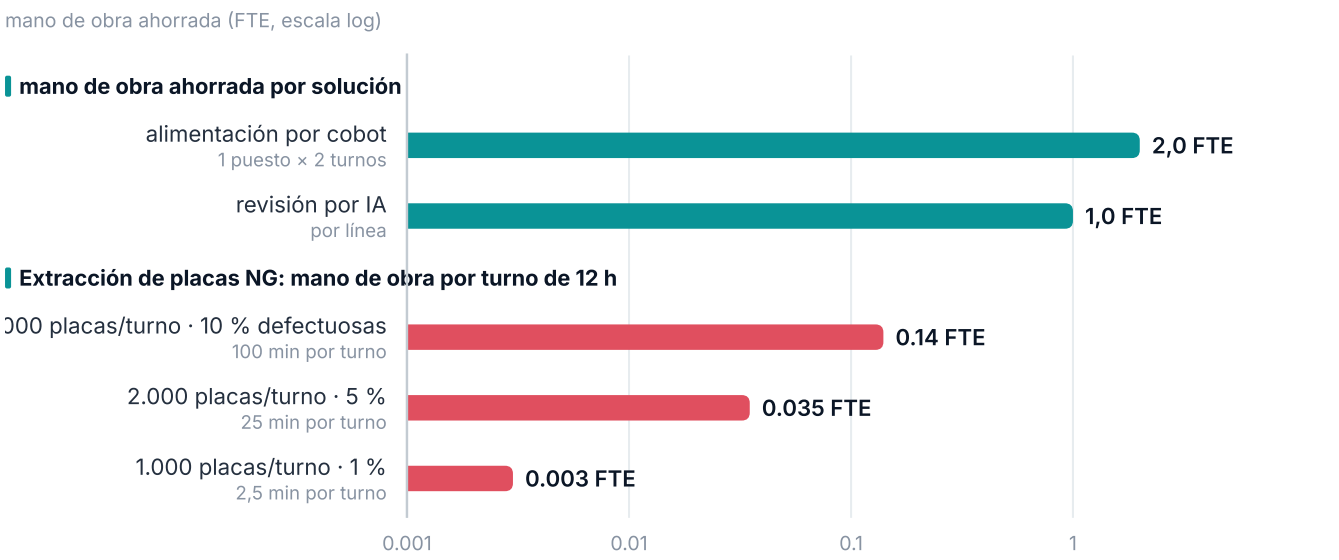
Fuente: Modelo del autor: caja de revisión IA de 50k RMB por línea; celda de alimentación por cobot a 200k RMB o 140k $, dos turnos.

Todos los valores son estimaciones a partir de las entradas anteriores. Sensibilidades: reduce a la mitad las personas ahorradas y todo retorno se duplica aproximadamente; ejecuta un turno y (b)–(d) se duplican; la depreciación del 100% del primer año en EE. UU. acorta el retorno en un 20%, mientras que los aranceles sobre equipo chino, japonés o coreano lo alargan un 10–40%; una subvención local del 15% en China lo acorta un 15%. No encontré ninguna encuesta independiente de umbrales de retorno de los compradores; los vendedores dicen que el AOI se amortiza «típicamente en 9–18 meses» (Averroes) y el caso de Siemens afirma ocho meses vendor claims. Uso uno y dos años como líneas de cribado. El modelo deja fuera las ganancias de calidad (garantía, devoluciones) y los costes de rotación y formación, de modo que es conservador.

La parte contraintuitiva de «detectar y recoger» es que la mano de obra está en la revisión, no en sacar las pocas placas NG reales de la línea. A 15 segundos de manipulación por placa NG estimate:

Placas por turno	Tasa de defectos	Tiempo de recogida por turno de 12 horas	Personas
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 qué extraer placas malas no se amortiza** ***Por encima del 90 % de rendimiento de primer paso, la extracción ocupa solo 0,003–0,14 de persona por turno.***



Fuente: Estimaciones del autor.

El caso base (2,000 placas por turno, 5% defectuosas, dos turnos) es 0.07 de una persona, un valor de unos RMB 6,000 al año en Shenzhen, menos que el coste de funcionamiento anual del brazo de unos RMB 12,000, de modo que nunca se amortiza en ningún sitio; incluso el peor caso (4,000 placas por turno, 10% defectuosas) tarda unos 13 años en Shenzhen y 9 en EE. UU. estimate . Si la preocupación es parar la línea para recoger, un desviador de NG o un transportador de búfer estándar lo resuelve sin un brazo. La gente del sector dice «por encima del 90% de rendimiento de primera pasada, para la línea y saca la placa a mano», y **eso es correcto para la recogida, pero no para la inspección**. El diseño de producto en una línea SMT es, por tanto, «inspeccionar más marcar o desviar», con brazos robóticos reservados para lugares donde se manipula cada pieza (carga de utillajes de prueba, manipulación de bandejas y cargadores, embalaje), donde la inspección se convierte en un complemento casi gratuito y se aplica el escenario (c).

Subvenciones, impuestos y aranceles mueven el retorno un 15–25%, pero no el orden

China continental: el apoyo en su mayoría reduce el coste de financiación; las máquinas individuales rara vez califican directamente. Una subvención nacional de préstamos paga 1.5 puntos de interés hasta dos años en 14 cadenas prioritarias (Yicai) verified , unos RMB 9,000 en una máquina de RMB 300k; la recompensa de actualización de Guangdong de hasta el 20% necesita un proyecto registrado con al menos RMB 7 millones de equipo nuevo y excluye Shenzhen (reglas de Guangdong), y Dongguan realmente pagó el 4.42% en julio de 2026 (asignación de Dongguan) single source, primary . Un dispositivo de inspección con IA se aprueba solo como equipo «con software de apoyo» dentro del proyecto registrado del comprador, de modo que el caso base asume sin subvención.

Taiwán: los incentivos van al comprador. Según el artículo 10-1 del Estatuto para la Innovación Industrial, enmendado en mayo de 2025 para incluir la IA, un comprador puede acreditar el 5% del gasto de 2025–2029 en nueva maquinaria inteligente o IA (incluida la inspección basada en CNN/DNN) contra el impuesto sobre la renta de ese año, o el 3% al año durante tres años (Estatuto, reglas de crédito) verified ; las subvenciones del lado del proveedor están cerradas al capital continental (MOEA) single source .

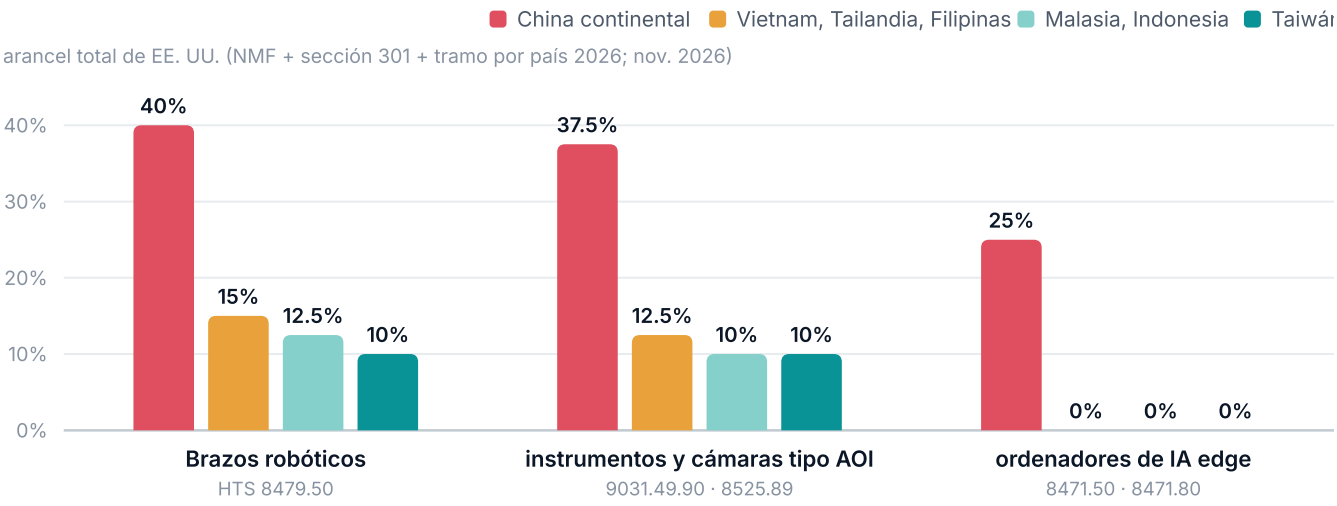
Sudeste Asiático: solo Malasia y Tailandia premian comprar automatización. La desgravación de capital de automatización del 200% de Malasia sobre los primeros RM10 millones vale alrededor de un 24% del capex para un comprador rentable (MIDA, PwC Malasia) verified ; el BOI de Tailandia da una exención fiscal de tres años de hasta el 50–100% de una inversión de actualización, con el límite más alto cuando participan proveedores tailandeses de automatización (BOI) single source, primary , de modo que entrar en Tailandia funciona mejor con un integrador local.

Vietnam, Indonesia y Filipinas premian construir plantas en lugar de comprar automatización (Vietnam Law Magazine, PwC Filipinas).

Estados Unidos: la depreciación acorta el retorno; los aranceles lo alargan. Los activos comprados después del 19 de enero de 2025 obtienen una depreciación permanente del 100% del primer año, alrededor de un escudo fiscal federal del 21% para un comprador rentable (IRS Notice 2026-11) single source, primary ; la investigación de la Sección 232 de robots y maquinaria industrial terminó el 28 de octubre de 2026 con negociaciones en lugar de nuevos aranceles, para ser reevaluada en 180 días (Casa Blanca, Mayer Brown) verified . El ensamblaje final en el Sudeste Asiático reduce el recargo sobre brazos e instrumentos tipo AOI de aproximadamente el 37.5–40% al 10–15% (capítulo 1).

Región	Instrumentos principales	Efecto sobre el retorno del comprador estimate
China continental	Subvención de préstamo de 1.5 puntos hasta dos años; recompensas provinciales y municipales de actualización (Guangdong necesita proyectos registrados ≥RMB 7M; Dongguan realmente pagó el 4.42%); medidas de Shenzhen 2026–2029 (tasas no públicas)	Un 0–5% para una máquina individual; 4–15% dentro de un proyecto grande
Taiwán	Crédito del Estatuto §10-1, 5% o 3% × 3 años	Un 5% más corto, si el comprador paga suficiente impuesto
Malasia	Desgravación de capital de automatización del 200% (primeros RM10M)	Un 15–25% más corto para compradores rentables
Tailandia	Exención fiscal BOI de tres años hasta el 50–100% de la inversión	Un 15–25% más corto para compradores rentables fuera de una exención fiscal
Vietnam, Indonesia, Filipinas	Incentivos de impuesto sobre la renta por construir plantas; depreciación extra del 20% de equipo en Filipinas	Casi nada para compras individuales de automatización
Estados Unidos	Depreciación del 100% del primer año; ≈37.5–40% de arancel combinado sobre equipo fabricado en el continente	Depreciación ≈20% más corto; aranceles ≈10–40% más largo

FIG 21 **Arancel de EE. UU. sobre robots, AOI y ordenadores edge*Los brazos fabricados en China continental pagan ≈40 %; enviados desde el sudeste asiático o Taiwán, 10–15 %.***



Fuente: Capítulo 99 del Arancel Armonizado USITC; Federal Register; USTR (noviembre de 2026).

Por región: en China continental, una caja de revisión con IA se amortiza en meses y una célula de cobot en 1.3–1.5 años, mientras que un AOI nuevo comprado solo para recortar mano de obra tarda 3–4 años y la detección-y-recogida integrada 4.5–5.3, ambas más allá de un cribado de dos años. Taiwán lo hace ligeramente mejor que Shenzhen (0.53 años para revisión, 1.1 para una célula de cobot), y contratar allí es más difícil. En Malasia y Tailandia, una caja de IA tarda 1–1.5 años y una célula de cobot 2–3, un 15–25% más rápido con los incentivos fiscales locales para compradores rentables. En Indonesia, Vietnam y Filipinas solo una caja de IA muy barata se amortiza en 2–3 años; la demanda allí proviene más de los requisitos de calidad de los clientes y la escasez de mano de obra. En EE. UU. todas las opciones excepto un brazo de retirada puro se amortizan en 0.2–3.4 años, con los aranceles, la reevaluación de la Sección 232 y la certificación NRTL como variables principales.

Mercado y competencia

05 Mercado, competencia y cómo pagan los compradores

Tamaño del mercado: las cifras publicadas difieren por múltiplos; solo un recuento ascendente es útil

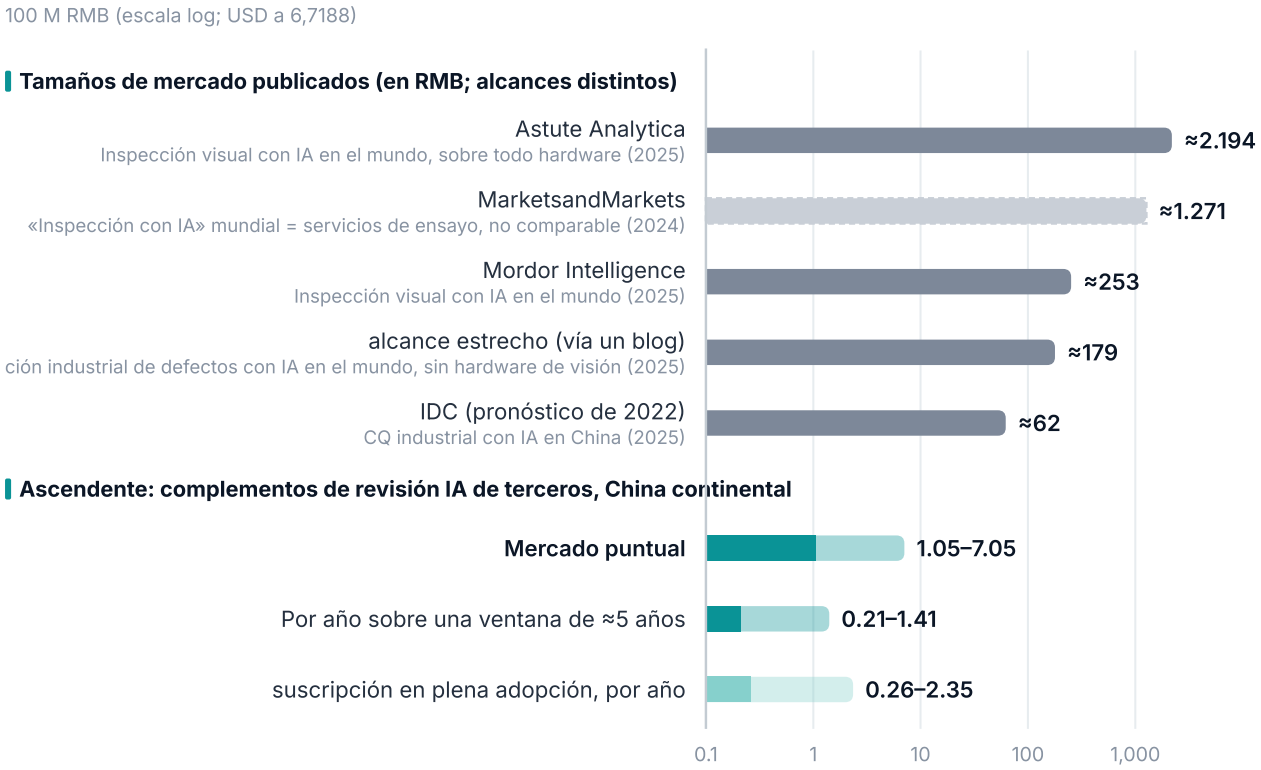
Cada cifra de mercado tiene que leerse con su alcance. El mercado de visión artificial de China fue de RMB 18.15 mil millones en 2024 (Zhiyan) verified, con la electrónica 3C como mayor comprador; las estimaciones globales de visión artificial difieren unas tres veces y las de AOI unas 6.5 veces (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). De forma ascendente a partir de los ingresos de Koh Young, TRI, Jutze, Sinic-Tek y ViTrox, el equipo global de inspección SMT es del orden de $1–2 mil millones estimate.

No existe un mercado independiente para el «software de inspección visual con IA». Las cifras publicadas difieren más de diez veces y todas incluyen hardware o servicios:

Alcance	Editor	Cifra
«Detección industrial de defectos con IA» global (estrecha, excluyendo hardware de visión)	Firma sin nombre, vía un blog de iFactory	$2.66 mil millones en 2025, 8.6% CAGR hasta 2035, electrónica 34%
«Inspección visual con IA» global	Mordor Intelligence	$3.76 mil millones en 2025, $12.8 mil millones en 2031, 22.67% CAGR
«Inspección visual con IA» global (mayormente hardware)	Astute Analytica	≈$32.66 mil millones en 2025
«Inspección con IA» global (servicios de pruebas y certificación; no comparable)	MarketsandMarkets	$18.92 mil millones en 2024
«Inspección de calidad industrial con IA» de China	IDC, previsión hecha en 2022	$958 millones en 2025 (≈RMB 6.2 mil millones), 91.5% en comunicaciones, electrónica y otras dos industrias

Fuentes: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) cada uno single source, mayormente resúmenes de informes de pago.

FIG 22 **Los tamaños de mercado difieren en más de diez veces** ***Solo el alcance estrecho y un recuento ascendente son útiles para decidir.***



Fuente: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; estimación ascendente del autor.

De forma ascendente, solo hay un recuento público de líneas SMT: New Venture Research estimó unas 73,231 líneas en todo el mundo en 2020, el 47% de ellas en China (19% en 2000), un 70% de gama alta y un 30% de gama baja (EMSNow, resumen de NVR SMT2020) single source, 2020 ; NVR ahora pertenece a IPC y los datos más nuevos son de pago (NVR). Creciendo un 3–5% al año con una cuota de China del 40–47% da unas 87,000–98,000 líneas en todo el mundo en 2026 y 35,000–46,000 en China continental (base 40,000) estimate .

Entrada	Bajo	Base	Alto	Base
Líneas SMT de China continental (2026)	35,000	40,000	46,000	Base NVR 2020 × crecimiento × cuota
Cuota sin revisión con IA	75%	80%	85%	Ventana de reacondicionamiento, abajo
Cuota que un tercero puede alcanzar (no interna, no ligada a IA de OEM, puede pagar)	40%	50%	60%	Supuesto: ≈30% de líneas de gama baja más algunas líneas de gama alta de marcas mixtas, excluyendo construcciones internas estilo DBG
Líneas aprovechables	**10,500**	**16,000**	**23,500**	Producto
Precio único por línea	RMB 10k	RMB 20k	RMB 30k	Escalera de precios y techo (capítulo 4)
Mercado único aprovechable (base instalada)	**RMB 105M**	**RMB 320M**	**RMB 705M**	Producto
Al año en una ventana de ≈5 años	RMB 21M	RMB 64M	RMB 141M	La ventana se reduce un 8–12% al año
Suscripción (RMB 2.5k / 5k / 10k por línea al año, adopción total)	RMB 26M al año	RMB 80M al año	RMB 235M al año	Ancla de cuota anual estilo Baidu

FIG 23 **Ascendente: 40.000 líneas → 16.000 servibles → ≈320 M RMB puntuales** ***El nicho de revisión IA de terceros en China continental es real, pero pequeño.***



Un nuevo entrante ganando 2–5 % de las líneas servibles en 3 años: ≈320–800 líneas, ≈2–5 M RMB al año
Como suscripción (5k RMB por línea y año): 1,6–4 M RMB recurrentes, solo el 0,3–2,3 % del pronóstico de IDC del CQ industrial con IA en China

Fuente: NVR (base 2020); estimaciones del autor.

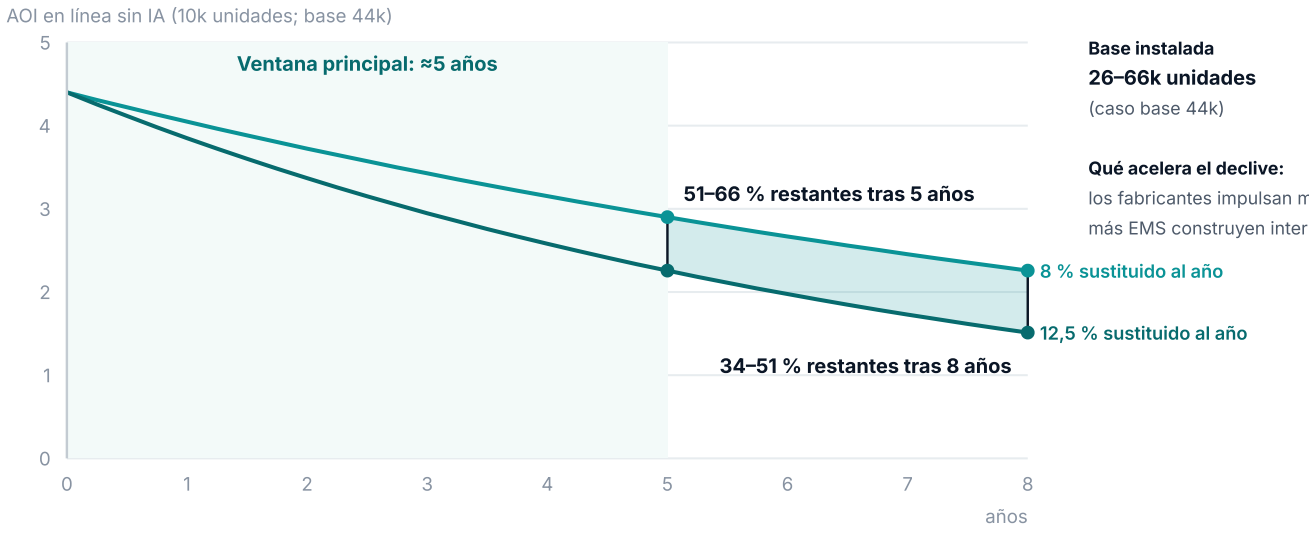
Todo esto son estimaciones; ni la cuota ni el precio se han medido públicamente. El nicho vale solo el 0.3–2.3% al año de la previsión de IDC para toda la inspección de calidad industrial con IA de China en 2025 (unos RMB 6.2 mil millones). Un recién llegado que ganara el 2–5% de las líneas aprovechables de base en tres años tendría unas 320–800 líneas: RMB 6.4–16 millones en total a RMB 20,000 por línea, unos RMB 2–5 millones al año, o RMB 1.6–4 millones al año en ingresos recurrentes a RMB 5,000 por línea estimate. Eso basta para una unidad de negocio dentro de una empresa, no para una empresa independiente de escala de riesgo. Cada 1,000 líneas instaladas en Taiwán, Sudeste Asiático o EE. UU. vale unos RMB 3–15 millones en pago único a precios continentales; los precios son más altos en Taiwán y EE. UU., pero allí los compradores usan más a menudo AOI 3D de OEM con opciones de IA estimate.

La ventana de reacondicionamiento es de unos cinco años

Ningún fabricante de AOI publica la vida útil de la máquina; las empresas de EMS deprecian el equipo en 3–10 años (memoria anual 2025 de DBG), y los AOI antiguos pierden actualizaciones de software y repuestos tras el fin de vida (I-Connect007) single source. Uso 8–12 años (base 10).

A 1.0–1.65 AOI en línea por línea, el continente tiene unos 35,000–78,000 AOI en línea en servicio (base unos 55,000), coherente con los envíos conocidos (Circuits Assembly, memoria anual 2025 de Sinic-Tek). Ninguna fuente mide cuántos ejecutan revisión con IA; suponiendo que las máquinas nacionales la ofrecen desde unos 2023 y que el 40–60% de las entregas recientes la usan, **unos 26,000–66,000 AOI en línea en China continental (base unos 44,000) no ejecutan IA** estimate. Si cada máquina nueva ahora se envía con IA, ese fondo se reduce un 8–12.5% al año: queda un 51–66% tras cinco años y un 34–51% tras ocho, más rápido si los fabricantes de AOI impulsan actualizaciones de IA a las máquinas más nuevas. **La mayor parte de la oportunidad cae en los próximos cinco años o así**, con las EMS medianas que ejecutan AOI 2D nacionales antiguos, marcas mixtas y sin vía de actualización de IA en su núcleo.

FIG 24 **La ventana de retrofit es de unos cinco años** ***Los AOI en línea sin IA se reducen un 8–12,5 % al año a medida que se sustituyen o mejoran.***



Fuente: Estimaciones del autor; base instalada y cuotas no verificadas de forma independiente.

Los competidores ya son fuertes: la cuña de revisión está rodeada

La cuña de revisión con IA se estrecha desde tres direcciones. Los líderes de AOI y visión están integrando IA en las nuevas plataformas: Jutze ha «adoptado totalmente los algoritmos de IA» (relaciones con inversores de Jutze, junio de 2026), los ingresos de IA industrial de OPT crecieron un 96.77% en 2025 (memoria anual 2025 de OPT) single source, primary , y ViTrox afirma hasta un 95% de precisión para la revisión autónoma (ViTrox) vendor claim . Los grandes fabricantes construyen la suya (iThome). Y las empresas de IA respaldadas por capital riesgo venden a escala: los clientes de sensores de borde de SmartMore crecieron de 65 a 251 en dos años (36Kr) single source , Hikvision vende su modelo Guanlan para inspección (Xinhua), y la integración «ojo, cerebro y mano» ya está en oferta de SmartMore, Hikrobot y Mech-Mind (36Kr, Securities Star).

FIG 25 **La cuña de revisión está rodeada por tres lados** ***El espacio para terceros está en EMS medianos y en nichos de inspección mal cubiertos por el AOI.***



Fuente: Informes anuales de Jutze, Sinic-Tek y DBG; material público de Hikvision y SmartMore; compilado por el autor.

El espacio que queda para un recién llegado es estrecho: EMS medianas con AOI 2D/3D antiguos y sin actualización de IA del fabricante; líneas que mezclan varias marcas de AOI (un vendedor de reacondicionamiento ya soporta Omron, VI, Koh Young y Mirtec (Sohu sobre Hengtianyi)); nichos que los AOI de SMT cubren mal, como uniones de agujeros pasantes y soldadura manual, recubrimiento conformal, cables, conectores y FPC, y apariencia final y tornillos; y plantas pequeñas que no pueden permitirse un AOI de RMB 230–280k.

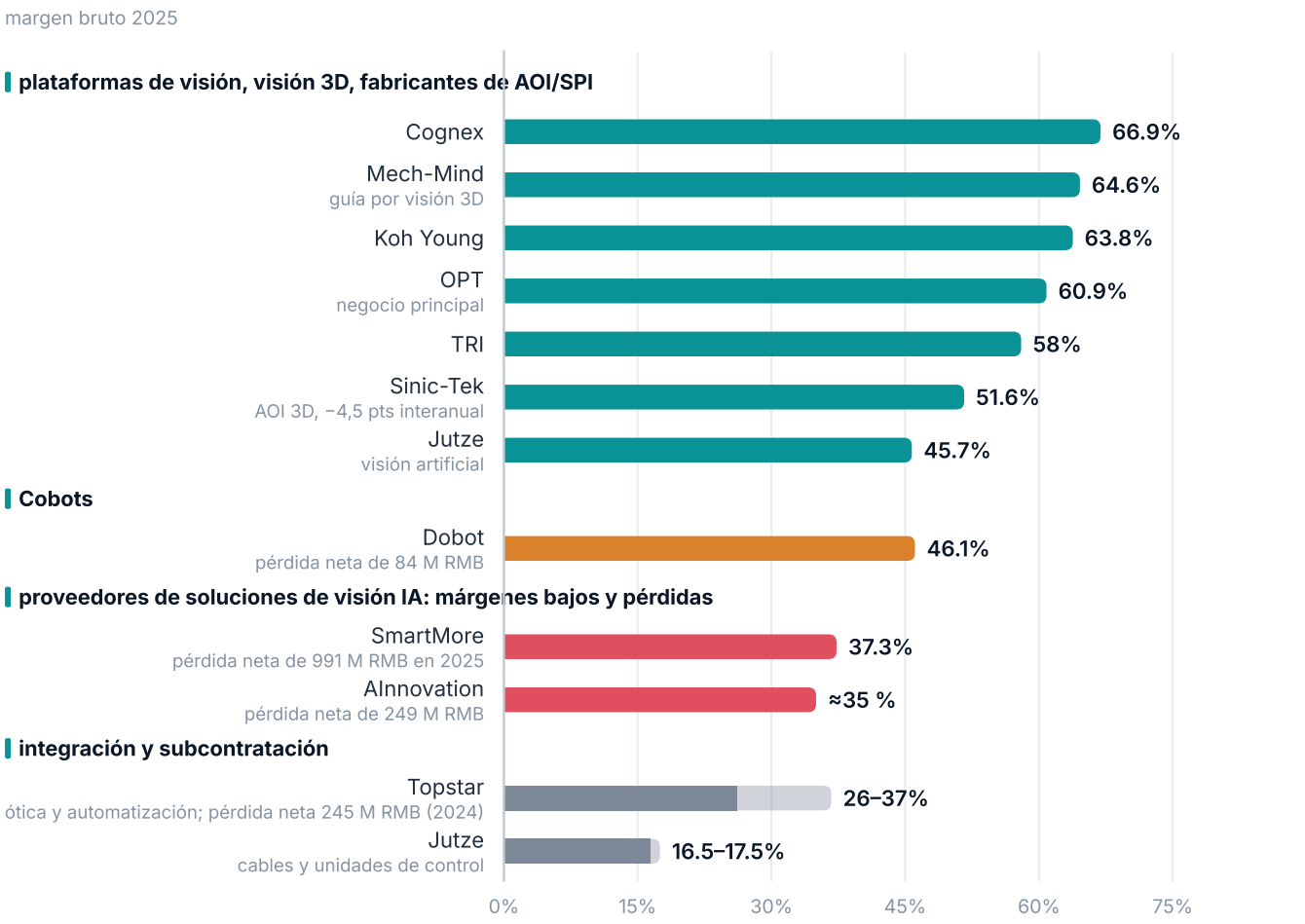
El comparable cotizado más cercano da una advertencia. Weiyi Zhizao, que vende equipos y robots de inspección con IA, vio caer los ingresos 3C un 79.9% interanual entre enero y mayo de 2026, reconoció una «competencia intensificada y presión de precios», pospuso su salida a bolsa en Hong Kong el 25 de octubre de 2026 y reveló 15 proyectos terminados antes de completarse, 12 de ellos proyectos de robots corporeizados (folleto de Weiyi Zhizao, aviso de aplazamiento) single source, primary. AInnovation terminó 2025 con el 29.8% de sus cuentas por cobrar provisionadas (memoria anual 2025 de AInnovation). La cuña de inspección 3C está abarrotada y el poder de fijación de precios por línea es débil.

El margen reside en el hardware convertido en producto

Posición en la cadena	Empresas representativas y margen bruto 2025	Beneficio y cobros
Plataformas de visión	Cognex 66.9%	Rentable; desinvirtiendo ≈$22M de ingresos de bajo margen
Equipos de AOI/SPI	Koh Young 63.8%; TRI 58%; negocio principal de OPT 60.85%; AOI 3D de Sinic-Tek 51.62% (−4.53 puntos interanuales); visión artificial de Jutze 45.74%	Rentable; días de cobro 139 (Jutze) a 247–274 (OPT)
Guiado por visión 3D	Mech-Mind 64.6%	—
Vendedores de soluciones de visión con IA	SmartMore 37.3%; AInnovation ≈35%	Con pérdidas: SmartMore pérdida neta RMB 991M en 2025, AInnovation RMB 249M
Cobots	Dobot 46.1%	Dobot pérdida neta RMB 84M; ingresos de UR de $326M (2022) a ≈$232M (2025, previsión)
Integración de sistemas	Robótica y automatización de Topstar 26.2–36.7%	Pérdida neta RMB 245M (2024); días de cobro 180–240
Hardware por contrato	Cables y unidades de control de Jutze 16.5–17.5%	—

Fuentes: (Cognex, StockAnalysis sobre Koh Young, Fugle sobre TRI, memoria anual 2025 de OPT, memoria anual 2025 de Sinic-Tek, memoria anual 2025 de 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 **El margen está en el hardware industrializado** ***Las plataformas de visión y fabricantes de equipos logran 46–67 % de margen bruto; los proveedores de soluciones IA, 35–37 % y pierden dinero.***



Fuente: Informes anuales o folletos de 2025.

Un recién llegado mantiene su margen solo convirtiendo la inspección en un producto, equipo o software; el ajuste proyecto a proyecto empuja los márgenes a alrededor del 35% con pérdidas persistentes. La mecánica también juega en contra de las empresas pequeñas: los ingresos solo se reconocen tras la aceptación en el sitio, los anticipos son pequeños (Jutze, OPT), los pilotos suelen ser gratuitos (A3/Mindtrace) y los propietarios de marca a menudo especifican el equipo de inspección en las líneas de sus fabricantes por contrato.

Cómo pagan los compradores: 30% a la firma, 60% tras la aceptación, 10% retenido un año

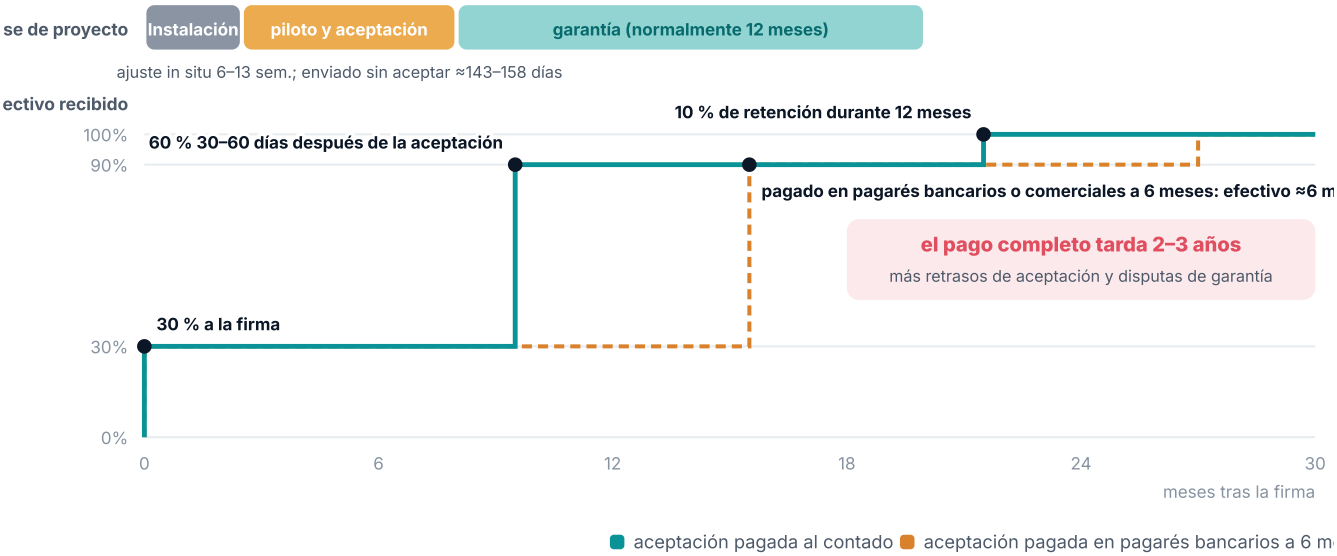
Los términos de pago que las plantas de electrónica continentales usan para el equipo de inspección se detallan en 13 contratos de compradores nombrados revelados en la respuesta a la consulta de cotización de Sinic-Tek (respuesta a la consulta de Sinic-Tek 8-1-1) single source: una presentación primaria, varios compradores independientes, 2020–2022 :

Comprador (año)	Firma	Entrega	Aceptación	Retención o saldo
Una unidad de Hikvision (2021)	30% en 15 días	—	65% en 30 días desde la aceptación final	5% de retención de garantía pagado 12 meses tras la aceptación; parte pagado en efectos de aceptación bancaria a 6+ meses
Nanchang Huaqin (2020–2021)	30% en 10 días laborables	—	60% a 60 días tras la aceptación	10% tras una garantía de un año

Comprador (año)	Firma	Entrega	Aceptación	Retención o saldo
Guangdong Yinuo (2021, AOI 3D)	30% en 10 días laborables	—	60% en 30 días	10% tras tres meses de funcionamiento sin fallos
Xiamen Hongxin, Jingmen Hongyi (2022)	30% en 10 días laborables	—	60% en 30 días laborables desde la aceptación final, efectos bancarios a 6 meses	10% tras una garantía de un año
AKM Panyu, Suzhou (2021–2022)	30% en 7–10 días laborables	—	60% un mes tras la aceptación y facturación, efectos bancarios a 180 días	10% tras 360 días
BYD Precision, Shenzhen, Huizhou (2020)	—	30% a la llegada	70% tras la aceptación y facturación	Todo en efectos comerciales a 6 meses
Shenzhen H&T (2020)	—	—	90% tras la aceptación y facturación	10% tras seis meses de uso sin fallos

Otras presentaciones primarias muestran la misma estructura: los grandes proyectos de Mech-Mind pagan 30% a la firma, 30% antes de la entrega, 30% tras la aceptación y 10% de retención de garantía a 12–15 meses (folleto de Mech-Mind); GKG dice que la retención en el sector de equipos suele ser del 5–10% del contrato, pagada un año tras la aceptación (respuesta a la consulta de GKG) single sources, primary. La norma, entonces, es alrededor de un 30% a la firma (20–30%, cero en algunas plantas como BYD), 60–65% tras la aceptación (pagado 30–60 días después, a menudo en efectos bancarios o comerciales a seis meses, lo que retrasa el efectivo utilizable otro medio año), y un 5–10% de retención durante 3–12 meses (12 lo más común) verified en varios contratos.

FIG 27 **Cómo pagan los compradores *30 % a la firma, 60 % tras la aceptación, 10 % retenido un año: el pago completo tarda 2–3 años.***



Fuente: Respuesta de consulta de Sinic-Tek, Mech-Mind, GKG, JT y otros documentos; esquema en condiciones habituales, calendario estimado.

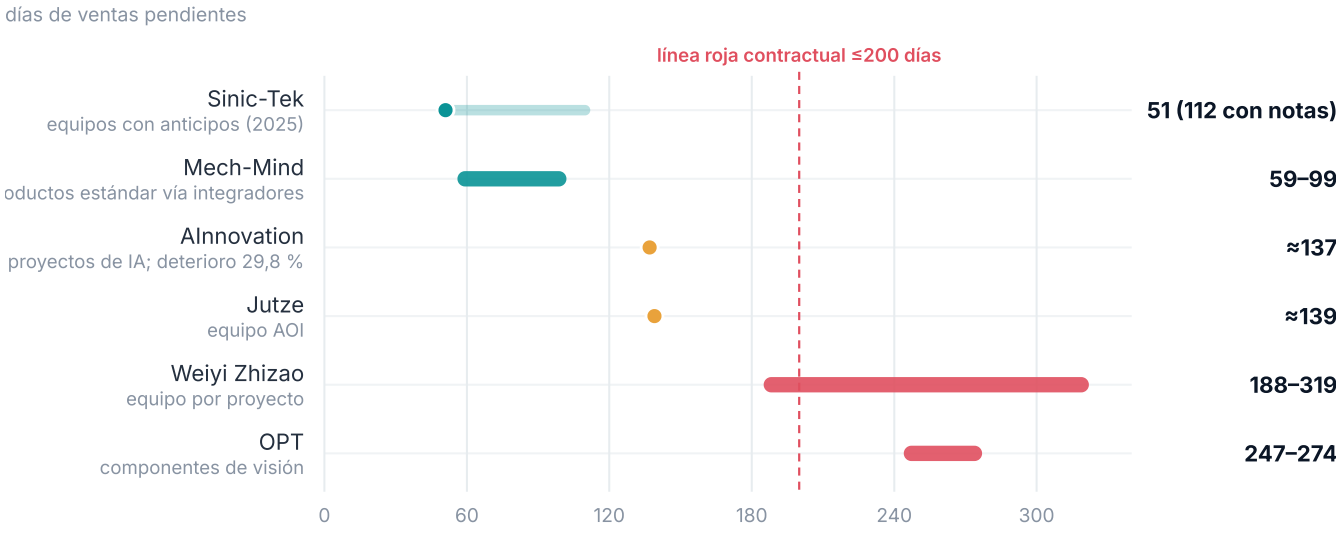
La aceptación y las pruebas son lentas y las paga el vendedor. Las mercancías enviadas pero no aceptadas de Sinic-Tek tardaron unos 143–158 días en rotar, porque los grandes clientes de EMS y marca necesitan pruebas de integración de línea completa y una ejecución de prueba extendida (respuesta a la consulta de Sinic-Tek 8-1-2); los proyectos de IA de Weiyi tardan 4–8 meses de la evaluación al despliegue (folleto de Weiyi Zhizao) single sources, primary. No hay datos de conversión para pilotos de revisión con IA; Sinic-Tek convirtió solo un 18–19% de sus má-

quinas de demostración directamente en ventas. Planifico uno de cada tres a uno de cada cinco para pilotos gratuitos y alrededor del 80% para pilotos pagados tras la preselección estimate.

Empresa	Modelo de negocio	Días de ventas pendientes de cobro
Sinic-Tek (2025)	Equipo con depósitos	≈51 (≈112 incluyendo efectos)
Mech-Mind	Productos estándar a través de integradores	59–99
Jutze	Equipo de AOI	≈139
AInnovation (2025)	Proyectos de IA	≈137 al cierre del año, con 29.8% de deterioro
Weiyi Zhizao	Equipo por proyecto	188–319
OPT	Componentes de visión	247–274

Fuentes: las memorias anuales y folletos anteriores; cuentas por cobrar (incluidos activos de contrato) ÷ ingresos × 365 primary + estimate.

FIG 28 **Días de ventas pendientes** ***Productos estándar 51–99 días; equipos por proyecto 188–319 días.***



Fuente: Documentos de empresa; cuentas por cobrar (incl. activos contractuales) ÷ ingresos × 365.

Las licencias de revisión con IA vendidas a través de fabricantes de AOI o integradores deberían cobrar como los 60–100 días de Mech-Mind; las ventas directas por proyecto como los 190–320 días de Weiyi. Junto como línea de tiempo de efectivo: piloto gratuito de uno a tres meses, cualificación en un ODM grande de hasta unos doce meses, construcción, entrega y aceptación de tres a siete meses; el 30% llega con el pedido, el 60% unos cuatro a nueve meses después y el último 10% unos 16–21 meses después. **Desde el primer piloto en una EMS grande hasta el pago completo pueden pasar de dos a tres años**; las EMS medianas sin licitaciones formales pueden acortar la cualificación a semanas estimate. Los criterios de aceptación en los contratos revelados cubren solo la integración de línea completa, el rendimiento (un proyecto taiwanés exigía 100 unidades por hora) y meses de funcionamiento sin fallos; ninguno especifica tasas de fuga o de falsa alarma (respuesta a la consulta de Anda) single source, primary. Los criterios de aceptación por escrito para la revisión con IA tienen que acordarse antes del piloto.

Cuatro patrones de fracaso

Los fracasos documentados caen en cuatro patrones: enviar antes de que el producto fuera fiable (Rethink (The Robot Report)); una capa de software sobre el hardware de un líder que el ecosistema del líder reemplazó después (READY Robotics (The Robot Report)); quedarse sin dinero en el fondo de un ciclo (Dextrous Robotics (Robots & Startups)); y empresas de modelos que terminan en adquisiciones de talento (Covariant (GeekWire)). Las lecciones:

productizar de forma acotada en lugar de ajustar proyecto a proyecto; mantener capital circulante para ingresos basados en aceptación y cobros largos; ofrecer servicio local para plantas que funcionan 24 horas; y no escalar antes de que el producto esté probado.

Momento: la inspección no gana nada esperando; el hardware de brazos se abarata

Los argumentos para esperar son débiles. El hardware no se abarató en 2026: los fabricantes de robots subieron precios un 5–15%, algunos productos un 25–50% (Sina Finance) single source, y los kits de NVIDIA subieron por tercera vez (36Kr). Los VLA de grado de producción no llegarán dentro de la ventana (IFR, Instrumental). Mientras tanto, cada sustitución de AOI elimina un objetivo de reacondicionamiento y cada despliegue de IA de OEM bloquea los datos y el proceso de una línea; la capacidad de los modelos se está mercantilizando, y lo escaso es el acceso a las plantas y los datos de fábrica etiquetados. El soporte local también tiene ventanas: las reglas de «ciudad pionera de IA» de Shenzhen de 2026, con hasta RMB 2 millones para proyectos de demostración de aplicaciones de IA y vales de modelo que cubren el 30% de las compras, duran hasta el **30 de junio de 2027** (Yicai, mayo de 2026) términos single source.

Lo que todavía se abarata rápido son los humanoides y las manos diestras (las manos cayeron de unos RMB 50,000 a menos de 10,000 en dos años (Sina Finance)), y los cobots de entrada a medida que la mezcla de productos se mueve hacia abajo; eso aboga por diferir la parte corporeizada, no la parte de inspección. **En conjunto: inspección ahora, como producto acotado; células de robot ahora, a través de socios e integración; IA corporeizada más tarde, con solo una estación de I+D hoy.**

IA corporeizada

06 IA corporeizada: primero cambia cómo se enseña a los robots, después cómo se mueven

Lo que está cambiando es la percepción y la programación; el movimiento sigue siendo determinista

Todos los líderes de brazos robóticos comercializan ahora la «IA física», pero lo que se ha enviado cae en tres capas: percepción aprendida sin enseñanza (AI Masterless Vision de FANUC, que se envía en enero de 2027 (FANUC); la selección neural de agarre de KUKA (KUKA)); modelos grandes que escriben programas y planifican tareas (iiQWorks.Copilot de KUKA, que advierte de que el código generado por IA debería verificarse en un gemelo digital (KUKA); Yaskawa con Gemini Robotics-ER, donde «la IA generativa decide qué hacer» y el robot decide «cómo moverse» (Yaskawa)); y plataformas de cómputo y datos (AI Accelerator de UR con una Jetson Orin (Universal Robots)) vendor claims. Ningún líder publica tasas de éxito de producción de sus funciones de IA, y los líderes japoneses se asocian con Google, SoftBank, Microsoft y otros en lugar de construir VLA de frontera (índice de noticias de FANUC): la capa de política aprendida vendrá de las empresas de modelos, los brazos y controladores certificados de los fabricantes de brazos.

FIG 29 **La pila de IA corpórea** ***Los modelos grandes planifican sin conexión, las políticas pequeñas corren en la celda, el movimiento y la seguridad permanecen deterministas.***

Capa	Qué hace y hasta dónde llega la evidencia	Escala de tiempo · quién la controla
Planificación L4	**modelos de frontera (sin conexión)** escribir o modificar programas de robot sin conexión, cambios de serie, explicar averías reescribe programas ABB RAPID al 77–99,7 %, desplegado tras simulación bucle de control a solo 0,2–0,4 Hz: queda fuera del tiempo real	**segundos a minutos** gigantes de modelos
Habilidades / políti L3	**ACT, políticas de difusión, VLA pequeños** puestos estructurados con utillaje 90–99,4 % chips edge: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**bloques de acción 100–300 m** fabricantes de chips · fabrican
control de movimie L2	**Controlador de robot** trayectorias, servo, control de fuerza, detección de colisiones código determinista; las políticas aprendidas solo fijan objetivos	**milisegundos (el tiempo real exige >83 Hz)** fabricantes de robots
Seguridad L1	**PLC de seguridad, escáneres, límites de fuerza y velocidad** certificado según ISO 10218-1/-2:2025 y normas locales Sin ML en funciones de seguridad: evaluación por organismo notificado de la UE desde el 20 en 2027	**tiempo real estricto** robot certificado + integrador

Las pequeñas empresas ganan en integración y hardware de detección: meter percepción aprendida y políticas pequeñas en celdas robóticas c

Fuente: Síntesis del autor (capítulos 3 y 6); ISO 10218:2025; Reglamento de Máquinas de la UE.

A qué distancia del trabajo no estructurado, flexible y de lotes pequeños: estaciones estructuradas en 1–3 años, ensamblaje flexible 5+

Las cifras reales de fábrica han empezado a aparecer, pero son reportadas por vendedores, estrechas y de estaciones estructuradas. Los ocho humanoides de AgiBot en la línea de prueba final de tabletas de Longcheer registra-

ron 64,828 tareas al 99.99% durante más de 64 horas (AgiBot, Yicai Global) [vendedor y cliente], pero es un trabajo de atención que un brazo convencional con visión puede hacer. UBTech dice que su Walker S2 trabaja al 30–50% de la eficiencia humana (Irish Times/FT) single source. Los 11 meses de Figure 02 en BMW Spartanburg produjeron cifras de volumen (1,250+ horas, 90,000+ piezas), mientras que sus cifras de ciclo, éxito e intervención son objetivos, no resultados (Figure, BMW Group) verified, solo numeradores. Nucleus dice que alrededor del 40% de su vídeo de fábrica es teleoperado (Interesting Engineering) vendor claim, y una investigación de Reuters citó alrededor del 20% de la producción de un trabajador para algunos humanoides chinos (Notebookcheck) single source. Ningún despliegue publica tasa de éxito, tiempo de ciclo y tasa de intervención juntos, y ninguno ha sido auditado, y sin embargo esos tres son exactamente lo que necesita un cálculo de retorno.

Despliegue	Tasa de éxito	Takt	Intervenciones o teleoperación	Tiempo de actividad o intervalo de fallo	Ejecución en volumen
Figure 02 @ BMW Spartanburg	Solo objetivo (>99% por turno)	Solo objetivo (84 s)	Solo objetivo (0 por turno)	No revelado (parte superior del antebrazo como punto de fallo)	1,250 h
Agility Digit @ GXO	No revelado	No revelado	No revelado	No revelado	100,000 cajas
Apptronik Apollo @ Mercedes-Benz	No revelado	No revelado	No revelado	No revelado	No revelado
Agile ONE @ Audi	No revelado	No revelado	No revelado	No revelado	No revelado
Nucleus (planta sin nombre)	No revelado	No revelado	≈40% teleoperado (método no revelado)	No revelado	≈2 h de vídeo
AgiBot G2 @ Longcheer	99.99% de 64,828 tareas	18–20 s por paso	No revelado	Un fallo de comunicación, ≈2 min para recuperar	64+ h
Célula híbrida de Neuromeka	322/324	12.8% más lento que una persona	No revelado	No revelado	324 ciclos
Physical Intelligence π0.6 en un sitio de cliente	—	—	96.4% autónomo durante un turno	No revelado	Un turno

Fuentes: este capítulo y el capítulo 3 mayormente vendor claims.

FIG 30 **Qué divulgan realmente los despliegues robóticos en planta** ***La mayoría divulga solo el volumen; ninguno divulga a la vez tasa de éxito, takt y tasa de intervención.***

	tasa de éxito	Takt	Intervenciones / teleoperación	Disponibilidad / averías	Volumen
Figure 02 @ BMW	solo objetivo >99 %/turno	solo objetivo 84 s	solo objetivo 0/turno	n/d	1.250 h
Agility Digit @ GXO	n/d	n/d	n/d	n/d	100.000 contenedores
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 (planta sin nombre)	n/d	n/d	≈40 % teleoperado*	n/d	≈2 h de vídeo
AgiBot G2 @ Longcheer	99.99%	18–20 s por paso	n/d	una avería, 2 min	64+ h
celda híbrida Neuromeka	322/324	12,8 % más lento	n/d	n/d	324 ciclos
π0.6 @ sitio del cliente	—	—	96,4 % autónomo	n/d	1 turno

* método no divulgado

Divulgado (sobre todo afirmaciones del proveedo… solo objetivos n/c

Fuente: Material público de BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence y otros.

La teleoperación importa más a la economía que la tasa de éxito. Personas netas sustituidas por robot por turno ≈ (ciclo humano ÷ ciclo robot) × tiempo de actividad × éxito, menos el tiempo de intervención y la teleoperación esti- mate, mi fórmula . Una estación estructurada estilo Neuromeka (12.8% más lenta, 99.4% de éxito, 95% de tiempo de actividad, 0.5 intervenciones de 2 minutos por hora) sustituye unas 0.82 personas; un humanoide al 40% de la velocidad humana, 90% de tiempo de actividad y dos intervenciones de 3 minutos por hora unas 0.26; al 20% de Reuters, unas 0.13; y si, como Nucleus, un operador lo teleopera el 40% del tiempo, unas 0.05, casi nada. A unos RMB 600k por humanoide y RMB 85k por persona al año en el continente, los tres últimos se amortizan a lo largo de décadas; solo la estación de brazo estructurada tiene un retorno realista (capítulo 4) estimate .

FIG 31 **Mano de obra neta sustituida por robot** ***Una celda de brazo estructurado sustituye ≈0,82 de trabajador; un humanoide teleoperado el 40 % del tiempo, ≈0,05.***



Fuente: Estimaciones basadas en fórmulas del autor.

Lejos de las estaciones estructuradas, los modelos generales siguen puntuando 40–90% en tareas nuevas: Gemini Robotics 2 alcanza el 74.2% en pick-and-place de laboratorio, el 78.9% en kitting de herramientas y el 44% en atar una bolsa (eesel/DeepMind); en RoboChallenge Table30, 30 tareas de robot real, π0.5 puntúa 43.7% y el mejor modelo abierto 50.3% (RoboChallenge, QbitAI).

Las previsiones externas difieren, pero ninguna apoya el uso generalizado en dos años. El IFR dice que la IA física general «aún no es de grado industrial» y que la mayoría de los humanoides son «todavía prototipos o pilotos»

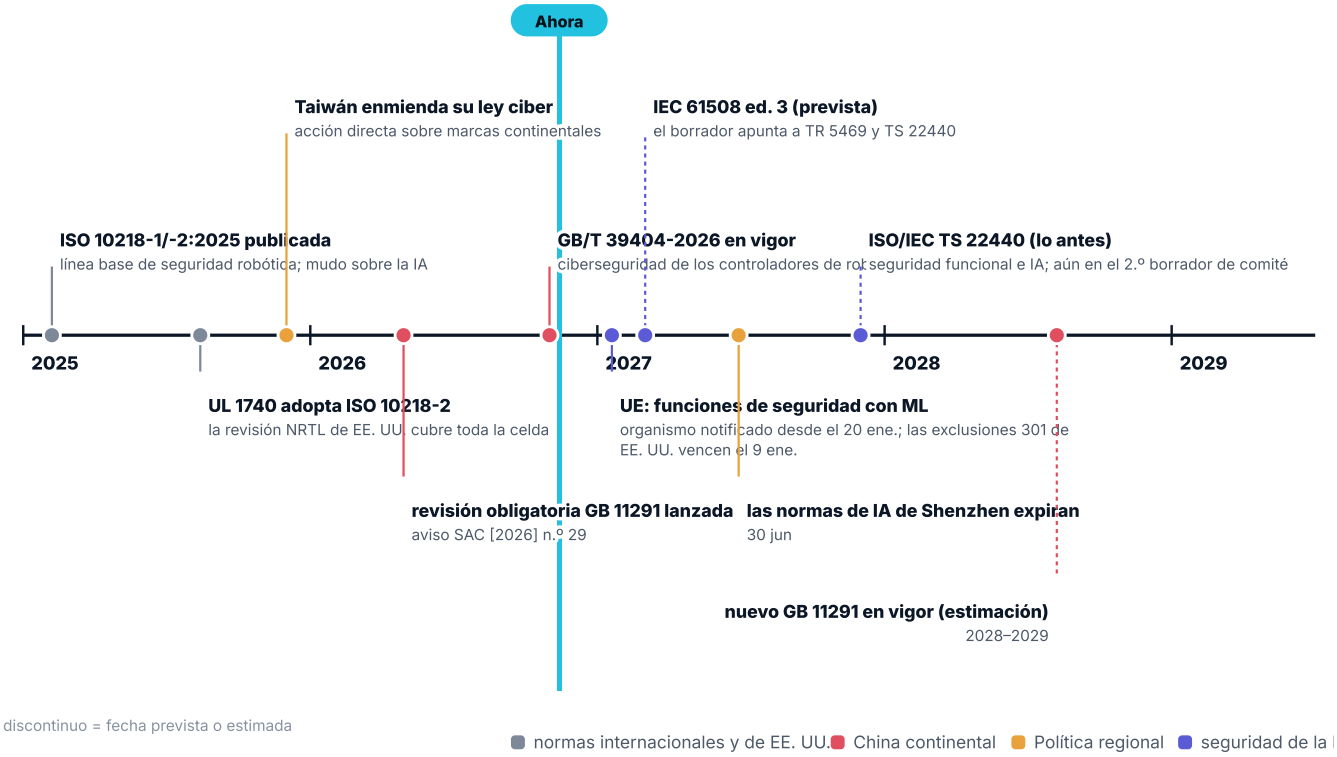
(IFR); Interact Analysis estima que solo alrededor del 10% de los humanoides producidos en 2025 entraron en uso real, con el punto de inflexión comercial «después de 2032» (automation.com); UBS ve un «momento VE» de los humanoides como improbable antes de 2030 (Yicai Global); los expertos chinos estiman de dos a tres años más antes del uso a escala de fábrica (Gasgoo) single sources. Mi juicio estimate: en plantas de electrónica y automoción bien dotadas con el vendedor en el sitio, las **tareas de estación estructurada** (atención, clasificación OK/NG, inserción con utillaje) alcanzarán ≥99% a velocidad cercana a la humana en unos 1–3 años; el **kitting de piezas mixtas** con VLA generales en unos 3–5; los **arneses de rutas libres y el ensamblaje flexible de varios pasos** en más de 5. El retorno de los humanoides no funciona hoy: un Walker S2 a unos RMB 600k (10jqka) single source, al 30–50% de la eficiencia humana durante dos turnos, equivale a unos 0.6–1.0 trabajadores, un retorno simple de 7–12 años a costes laborales continentales antes de integración, mantenimiento y tiempo de inactividad estimate.

Dónde está atascado

Eslabón	Evidencia en 2026
Datos	El sector estima que se necesitan al menos 10 millones de horas de datos reales de alta calidad y solo se habían recogido unas 500,000 a principios de 2026 (Gasgoo); Reuters informa de unas 500,000 horas frente a unos 100 millones necesarios (Notebookcheck); un teleoperador novato necesita unos 300 intentos por movimiento utilizable; las plantas no compartirán datos de proceso; las políticas entrenadas en simulación caen bruscamente en robots reales
Fiabilidad	Estaciones estructuradas 98–99.99% (vendor claims, fallo sin definir); modelos generales 40–90% en tareas nuevas; conectores no vistos 29–67%
Velocidad	Humanoides al 30–50% de la velocidad humana (UBTech), algunos alrededor del 20% (Reuters); agentes de frontera unos 280 s por tarea cuasi estática sobre robot real; bucles de control LLM a 0.2–0.4 Hz frente a los ≈83 Hz necesarios (Anthropic)
Coste	Walker S2 promedia unos RMB 600k; retorno de 7–12 años estimate
Certificación de seguridad	ISO 10218-1/-2:2025 hace la seguridad funcional más explícita y absorbe los requisitos colaborativos, pero no dice nada sobre la IA (The Robot Report); ISO/IEC TS 22440 (seguridad funcional e IA) seguía siendo un segundo borrador de comité en noviembre de 2026 (ISS, ISO Open Data); la guía IEC 61508-3 desaconseja la IA por encima de SIL 1 (Analog Devices); bajo el Reglamento de Máquinas de la UE, las funciones de seguridad con aprendizaje automático necesitan evaluación por organismo notificado desde el 20 de enero de 2027 (EUR-Lex)
Manos y herramientas	Las manos diestras manejan unos 30 N (≈3 kg) por dedo frente a unos 10 kg de una mano humana (Gasgoo); los cambios de herramienta siguen siendo manuales
Integración	Lecturas erróneas de sensores, desajustes de protocolo y enlaces MES en el sitio; tasas de intervención casi nunca reveladas

La implicación de seguridad es directa: las normas ni prohíben las políticas aprendidas ni las reconocen. ISO 10218-1/-2:2025 no dice nada sobre la IA; ISO/IEC TS 22440 (seguridad funcional e IA) sigue en borrador de comité y es improbable antes de finales de 2027 o 2028 estimate; ISO/IEC TR 5469:2024 es solo guía, recomendando respaldos no-IA, restricciones de supervisión y voto redundante; IEC 61508 edición 3 se espera a principios de 2027 (ISO Open Data, BSI, 61508 Association) verified. Ninguna norma publicada ofrece todavía una vía aceptada para certificar la percepción con aprendizaje automático como función de seguridad. Las reglas regionales también se mueven: China está revisando GB 11291 (en vigor hacia 2028–2029) (plan de revisión 20262543); UL 1740 ha llevado ISO 10218-2 a la revisión NRTL estadounidense desde agosto de 2025; la UE exige evaluación por organismo notificado de las funciones de seguridad con aprendizaje automático desde el 20 de enero de 2027; y ningún humanoide está certificado para uso de fábrica sin vallas mientras ISO 25785-1 siga en borrador (i-SCOOP). La arquitectura defendible: el aprendizaje automático toma solo decisiones de proceso («qué pieza», «es buena esta placa»), mientras que una capa de seguridad determinista certificada (típicamente PL d / Categoría 3) y un PLC de seguridad deciden velocidad, fuerza y paradas. Trata cualquier afirmación de «IA de cámara para seguridad sin vallas» como no probada hasta que un certificado PL d o SIL 2 nombre el componente de aprendizaje automático.

FIG 32 **Cronología de normas y regulación** ***Una vía de certificación para la IA dentro de funciones de seguridad no se asentará antes de 2027–2029.***



Fuente: ISO, IEC, UL, EUR-Lex, SAC y bases de datos regulatorias nacionales.

Qué capa es la oportunidad: los modelos pertenecen a los gigantes, las cajas de borde a los fabricantes de chips, las empresas pequeñas a la integración y el hardware de sensado

La capa de modelo necesita capital y datos y pertenece a los gigantes (AgiBot entrenó GO-2 con «decenas de miles de horas» (iFeng Tech); los modelos más fuertes son cerrados o solo en la nube). La capa de cajas de borde pertenece a los fabricantes de chips: Jetson Thor es el estándar de facto, y en el continente las placas RDK de D-Robotics (D-Robotics) y los coprocesadores Rockchip (CSDN) compiten, mayormente sin resultados de VLA publicados.

Las empresas pequeñas tienen dos aperturas. Una es el hardware de sensado que compra fiabilidad directamente: el tacto eleva el éxito de inserción de ContactDP del 58.7% al 97.8%, y los sensores de fuerza/par y las cámaras de muñeca hacen un trabajo similar; los kits de recolección de datos también se venden (10jqka), aunque las manos, los sensores y los brazos están todos cayendo de precio. La otra es la capa de integración entre los fabricantes de brazos y las empresas de modelos, poseyendo los datos de tarea, la envolvente de seguridad y el enlace MES de cada estación, que los fabricantes de brazos no venden como producto hoy. Los competidores caen en cuatro grupos: líderes de brazos (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot y otros), empresas de visión más brazo (Mech-Mind, Hikrobot, SmartMore), empresas de IA corporeizada (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) y entrantes de chips de automoción.

Aceptación del comprador: ningún número sin denominador

Como ningún vendedor publica tasas de intervención, intervalos de fallo o conteos de éxito a largo plazo, estos pertenecen a los términos de aceptación del contrato. Estadísticamente, 3,000 ciclos de producción consecutivos con cero fallos demuestran una tasa de fallo de como mucho 0.1% al 95% de confianza; 10 fallos en 3,000 demuestran solo ≤0.57%; demostrar ≤1%, ≤0.5%, ≤0.1% y ≤0.01% con cero fallos requiere 299, 598, 2,995 y 29,956 ciclos esti-mate, binomial exacto. Con ese estándar, el 322/324 de Neuromeka tiene un intervalo del 95% a dos lados de 97.8–

99.9% y no demuestra ≥99%; los aproximadamente 6 fallos de AgiBot en 64,828 tareas implican 99.98–100%, pero solo si «fallo» se define como lo define el comprador estimate. Estos son los requisitos de aceptación que yo pondría a cualquier célula de brazo robótico o corporeizada; los umbrales se basan en los propios objetivos de Figure en BMW y en los casos de Neuromeka y AgiBot, y son juicio, no una norma estimate:

Elemento	Requerir	Umbral de aprobación sugerido
Definiciones	Definiciones firmadas de ciclo, éxito, fallo (éxito parcial, re-agarre, caída, fuga de calidad), intervención (toque, teleoperación remota, reinicio, e-stop) y cómo se calcula la autonomía	Las definiciones vienen antes de cualquier métrica; ningún porcentaje sin fórmula
Intervenciones y teleoperación	Por hora-robot, por tipo y duración; tiempo medio entre intervenciones; minutos de teleoperación como parte del tiempo de ejecución; operadores por robot	Estaciones estructuradas ≤0.5 intervenciones por hora-robot (≥2 h de separación); cero teleoperación durante la aceptación
Éxito a largo plazo	≥3,000 ciclos de producción consecutivos durante ≥3 turnos, ≥2 números de pieza y variación normal de iluminación y suministro; fallos clasificados; cota inferior exacta	Cota inferior ≥99.5% para estaciones en línea que paran la línea, ≥99.0% sin conexión; fugas de calidad reportadas por separado
Tiempo de ciclo	P50/P90/P99 por pieza incluyendo recuperación, contra ≥30 ciclos humanos cronometrados en la misma estación	≤1.15 × P50 humano, o dentro del takt de línea
Tiempo de actividad y fallos	Horas de ejecución, registro de inactividad, fallos por componente, tiempo de cambio de pieza	≥95% de tiempo de actividad durante ≥2 semanas; MTTR ≤20 min con repuestos en el sitio
Recuperación	Inyección de fallos con guion: caídas, oclusión, cambio de iluminación, parada de transportador, pérdida de MES o red, pérdida de energía, e-stop	≥90% de recuperación autónoma, de vuelta a producción en 2 min
Cambio de modelo	Horas, demostraciones, habilidad del operador y coste de recolección de datos por número de pieza o utillaje nuevo	Estaciones estructuradas: ≤1 turno y ≤100 demos por número de pieza nuevo
Seguridad	Brazo y célula según ISO 10218-1/-2:2025, evaluación de riesgos de la célula, reglas locales (GB 11291 en el continente; CNS 14490 más evaluación escrita en Taiwán; ANSI/A3 R15.06-2025 y UL 1740 en EE. UU.)	Ninguna función de seguridad implementada por aprendizaje automático; una capa de seguridad certificada (típicamente PL d / Cat 3) limita velocidad y fuerza; sin robots con patas sin vallas hasta que ISO 25785-1 se publique y evalúe
Cómputo y comportamiento sin conexión	Hardware de inferencia, latencia P99, comportamiento sin red, versionado de modelos, revalidación tras cada actualización	La inferencia crítica para el takt no depende de la nube; reaceptación completa tras cualquier cambio de modelo, cuantización o motor
Evidencia	Vídeo sin editar y registros brutos de toda la ventana de aceptación; presenciado por el equipo de calidad del cliente o un tercero	Sin vídeos destacados ni resúmenes solo del vendedor
Términos comerciales	Hitos de pago ligados a métricas de aceptación; penalizaciones de disponibilidad cuando se factura por hora-robot; reglas de propiedad de datos y acceso remoto	El pago final y la retención se liberan solo tras la aceptación de 3,000 ciclos y ≥4 semanas de funcionamiento estable
Economía	Personas netas sustituidas = ratio de velocidad × tiempo de actividad × éxito − tiempo de intervención − parte de teleoperación	Rechazar cualquier propuesta cuya sustitución neta no pueda sostener su retorno afirmado

Fuentes: (Figure, Neuromeka, Yicai Global, i-SCOOP); los umbrales son mis sugerencias.

Preguntas abiertas

07 Lo que el registro público no puede responder, y cómo lo resuelve un piloto

Las preguntas que las fuentes públicas no pueden responder caen en cinco grupos: por cuánto se vende realmente la revisión con IA y cuánto pagarán los compradores; cómo empaquetan los fabricantes de AOI la IA y actualizan las máquinas antiguas; qué antigüedad tienen los AOI instalados y cuántos ya ejecutan IA; cuántas personas y horas ocupa realmente la revisión por línea; y cómo convierten los pilotos y cómo pagan los compradores. Encima de esas viene cómo rinden realmente los modelos más nuevos, las compilaciones cuantizadas y los brazos robóticos bajo las condiciones de una planta dada. Cada una puede convertirse en una prueba con un umbral claro. Los umbrales de abajo son reglas de decisión que derivé de la evidencia de este informe estimate; recalcula una vez que lleguen los primeros datos de planta.

Pregunta abierta	Cómo resolverla	Qué aspecto tiene un buen resultado
Precio y disposición a pagar por la revisión con IA	Una solicitud de cotización estándar (por ejemplo cuatro líneas con AOI 2D antiguos y 3D más nuevos mezclados, dos turnos, unas 2,000 placas por turno, 12 clases de defectos, solo en local, integración MES o IPC-CFX) enviada a varios vendedores terceros; una encuesta de sensibilidad al precio de cuatro preguntas a 20–30 plantas de EMS	La mayoría de las cotizaciones dentro de RMB 40–120k en pago único o 20–40k al año por cuatro líneas; al menos la mitad de las plantas aceptan RMB 10–30k por línea con retorno en 12 meses
Empaquetado de IA de los fabricantes de AOI y actualizaciones para máquinas antiguas	Entrevistas estructuradas cortas con una docena de fabricantes y distribuidores de AOI	Si la revisión con IA es estándar, opcional o ausente en los modelos 2025–26; si las máquinas anteriores a 2020 pueden actualizarse y a qué precio
Antigüedad de los AOI instalados y parte que ejecuta IA	Encuesta en el sitio de unas 30 plantas de EMS, estratificada por tamaño y región, con listas de equipo revisadas bajo NDA	Una mediana de edad de AOI ≥5 años y ≥60% sin IA significa que el fondo de reacondicionamiento es real
Mano de obra y horas de revisión	Entrevistas y 2–4 semanas de registros de alarmas de AOI, revisión, reparación y turnos en dos o más plantas, con muestreo de trabajo	≥1 revisor dedicado por línea por turno y ≥3 horas de revisión por línea por turno, con al menos 0.5 FTE por línea que la planta está dispuesta a redistribuir
Rendimiento de la propia revisión con IA	Conjunto dorado más modo en sombra en una línea real (capítulo 3): acuerdo de datos en la semana 0, línea base y etiquetado en las semanas 1–2, ejecución en sombra en las semanas 2–4, ejecución asistida opcional en las semanas 4–8	Falsas alarmas u horas de revisión reducidas ≥50–60%; cotas de fuga por clase no peores que la línea base humana; cero fallos en clases críticas; latencia p95 ≤½ takt; fallo-abierto automático a personas
Calidad de etiquetado del profesor de los modelos de frontera	Ejecutar dos o tres candidatos en el mismo conjunto dorado, tres redacciones de prompt cada uno, una buena imagen de referencia, recortes de un solo componente ≥42–48 px	Aceptación automática solo para clases con cero fallos en ≥299 defectos y acuerdo de nivel humano; de lo contrario solo pre-etiquetado y clasificación por prioridad
Compilaciones cuantizadas	Cambiar solo la precisión; fijar hashes de pesos, plantilla, preprocesado, decodificación, motor y tamaño de lote; deshabilitar o restablecer la caché de prefijo; comprobar que la compilación de referencia reproduce los resultados del vendedor; puntuar pares en el conjunto dorado y la rodaja de deriva	Ningún fallo nuevo en clases críticas; ≤1 fallo nuevo en 1,000 defectos; resolución caída ≤2 puntos; ≥99.5% de decisiones idénticas en tres ejecuciones; de lo contrario volver a FP8/INT8

Pregunta abierta	Cómo resolverla	Qué aspecto tiene un buen resultado
Hardware de línea, limpieza y ESD	≥2 unidades de cada caja candidata sometidas a estrés al doble de la tasa de alarmas esperada; remojo de 72 horas con el flujo de aire real del armario; comprobación de ESD por el coordinador del cliente; prueba ISO 14644-14 para clientes en ISO 7 y más limpio	p95 ≤½ takt; sin cuelgues ni limitación térmica durante 72 horas; visto bueno de ESD; informe de partículas en la clase correcta
Términos de pago, reglas de piloto y conversión	Cláusulas de pago de contratos recientes de equipo de inspección en las plantas encuestadas; entrevistas sobre reglas de piloto, criterios de aceptación y tiempo de cualificación	Alrededor de 30/60/10 con efectivo en 60 días desde la aceptación; pilotos ≤8–12 semanas con un responsable de visto bueno nombrado; ≥1 de cada 3 pruebas convirtiendo; un marco que cubra ≥5–10 líneas en el primer año
Datos, TI y responsabilidad	Entrevistas y revisión de contrato con asesor legal	Una caja local puede conectarse y acceder a los datos; responsabilidad limitada a defectos de software
Cumplimiento regional antes de salir al extranjero	Taiwán: estado escrito del cliente bajo la Ley de Gestión de Ciberseguridad, autoevaluación E187; EE. UU.: resolución aduanera sobre cómo se clasifica una célula, cotizaciones NRTL, comprobación ITAR; Sudeste Asiático: vinculación local BOI y elegibilidad de la desgravación malasia, matriz del cliente; continente: alcance CCC para PC y fuentes de alimentación empaquetadas	El cliente no es un organismo controlado, y el retorno se mantiene ≤2 años con arancel y certificación incluidos
Aceptación de un robot o célula corporeizada	La tabla de aceptación de 12 elementos del capítulo 6: definiciones firmadas, ≥3,000 ciclos de producción consecutivos durante ≥3 turnos y ≥2 números de pieza, inyección de fallos con guion, ≥30 ciclos humanos cronometrados, presenciado	Cota inferior ≥99.5% en línea (≥99.0% sin conexión); ciclo ≤1.15 × humano; ≤0.5 intervenciones por hora y cero teleoperación; ≥95% de tiempo de actividad; MTTR ≤20 min; ≥90% de recuperación
Políticas ACT y VLA en chips de borde chinos	Cuatro semanas: portado y micro-benchmark (≥1,000 inferencias, distribución de latencia, potencia, remojo de 60 minutos); simulación en bucle cerrado con latencia medida inyectada; A/B sobre robot real contra una GPU; resistencia de 8–24 horas (≥3,000 ciclos)	Fragmentos ACT p95 ≤100 ms, VLA ≤300 ms; éxito sobre robot real dentro de 3 puntos de la GPU y ciclo ≤1.2 × humano; de lo contrario mantener la política en Jetson o una GPU
Precios de cobot, AMR e integración	Cotizaciones modelo por modelo de varios distribuidores de cobot, y cotizaciones de célula completa de dos o tres integradores para carga de utillajes de prueba (una estación, dos turnos)	Célula completa ≤RMB 200k y retorno ≤1.5 años en dos turnos en Shenzhen
Conteos de líneas SMT en Taiwán, Sudeste Asiático y EE. UU.	Datos de IPC (antes New Venture Research) o bases instaladas de distribuidores de equipo locales	Suficientes líneas aprovechables por región para justificar la ingeniería de aplicación local

De todas estas, cuatro deciden los números más sensibles de este informe: el precio que pagarán los compradores, la mano de obra que la revisión con IA realmente libera, si la revisión con IA realmente no falla más que las personas, y con qué rapidez vuelve el efectivo. Vale la pena resolverlas antes de cualquier compromiso significativo, por una planta que considera una compra o por un proveedor que considera un producto.

Una secuencia de piloto que funciona para ambos lados

Para una planta, y para un vendedor que quiere una referencia creíble, el orden es el mismo:

1. **Semanas 0–2: acordar datos y definiciones.** Firma el acuerdo de datos, define las clases de defectos y qué cuenta como fuga, y registra una línea base: falsas alarmas por placa, segundos de revisión por llamada, fugas

encontradas aguas abajo.

2. **Semanas 1–4: construir el conjunto dorado** a partir del archivo del AOI y las placas NG conservadas, etiquetado a ciegas y arbitrado (capítulo 3).
3. **Semanas 2–4 (o más): modo en sombra.** El modelo registra sus decisiones mientras las personas deciden como antes; comparar por clase.
4. **Semanas 4–8: ejecución asistida en turnos aleatorios,** con el modelo resolviendo solo las clases que pasaron la puerta y todo lo demás yendo a personas.
5. **Aceptación, por escrito, antes de escalar:** los resultados por clase, la latencia, el comportamiento de fallo-abierto, y quién posee modelos y etiquetas. Escala línea a línea, y revalida en cada producto nuevo, versión de modelo, cuantización o cambio de motor.

Limitaciones de este informe

Esta es una investigación documental. No realicé entrevistas primarias con plantas para esta edición; gran parte de los datos de línea provienen de vendedores y no se han auditado de forma independiente. Varios sitios web gubernamentales, bases de datos de licitaciones y plataformas de reclutamiento no pudieron accederse, de modo que las tasas de subvención de actualización de Shenzhen de 2026, la guía salarial del delta del río Yangtsé, los precios de transacción de la revisión con IA, los revisores por línea y las tasas de conversión de pilotos son estimaciones o analogías; cada una aparece arriba como una pregunta abierta. Los tipos de cambio son la paridad central del Banco Popular de China del 9 de noviembre de 2026 ($1 = RMB 6.7188) y, para Taiwán, la fijación de SMBC del mismo día ($1 = NT$32.12); el dong vietnamita usa datos de mayo de 2026. Todos los umbrales son reglas de decisión derivadas de la evidencia citada, no normas.

Sobre el autor

Tan Shuai

Tan Shuai es tecnólogo y fundador con más de veinte años de trabajo entre chips, hardware y software de sistema, y 20 patentes concedidas. Su trabajo lo ha llevado por Silicon Valley, Nueva York, Chicago, Taiwán, la Gran Área de la Bahía (Shenzhen y Hong Kong), Vietnam y Malasia, con un largo foco en la fabricación avanzada y las cadenas de suministro. Hoy trabaja como investigador y desarrollador de IA, centrado en cómo se usa realmente la IA en la planta de producción.

Le agrada conversar con personas que trabajan en IA en la fabricación, ya sea montando un equipo, buscando un socio o comparando notas: i@tanshuai.com · tanshuai.com

Notas y descargo de responsabilidad

Este informe es una investigación documental independiente basada en documentos públicos disponibles hasta el 11 de noviembre de 2026. Se proporciona solo con fines informativos y no es asesoramiento de inversión, legal, fiscal ni de ingeniería. Los nombres de productos, precios, lanzamientos de modelos, clasificaciones de exportación, aranceles y regulaciones cambian rápidamente; verifica cualquier cosa de la que dependas contra la fuente primaria. Las cifras marcadas como vendor claim o estimate deben leerse como tales. Los nombres de empresas y productos pertenecen a sus propietarios y se usan solo con fines de identificación; ninguna empresa revisó ni patrocinó este informe.

© 2026 Tan Shuai. Puedes compartir este informe sin cambios y con atribución.

FABRICACIÓN ELECTRÓNICA · INSPECCIÓN CON IA × ROBÓTICA

Inspección visual con IA y brazos robóticos en fábricas de electrónica: evaluación 2026

Cada cifra clave lleva su fuente y una etiqueta de confianza: verified, single source, vendor claim o estimate. Lo que el registro público no puede responder se recoge en el capítulo 7, junto con la forma en que un proyecto piloto resuelve cada pregunta.

Tan Shuai
tanshuai.comi@tanshuai.com

EDICIÓN PÚBLICA · DATOS AL 11 DE NOVIEMBRE DE 2026

