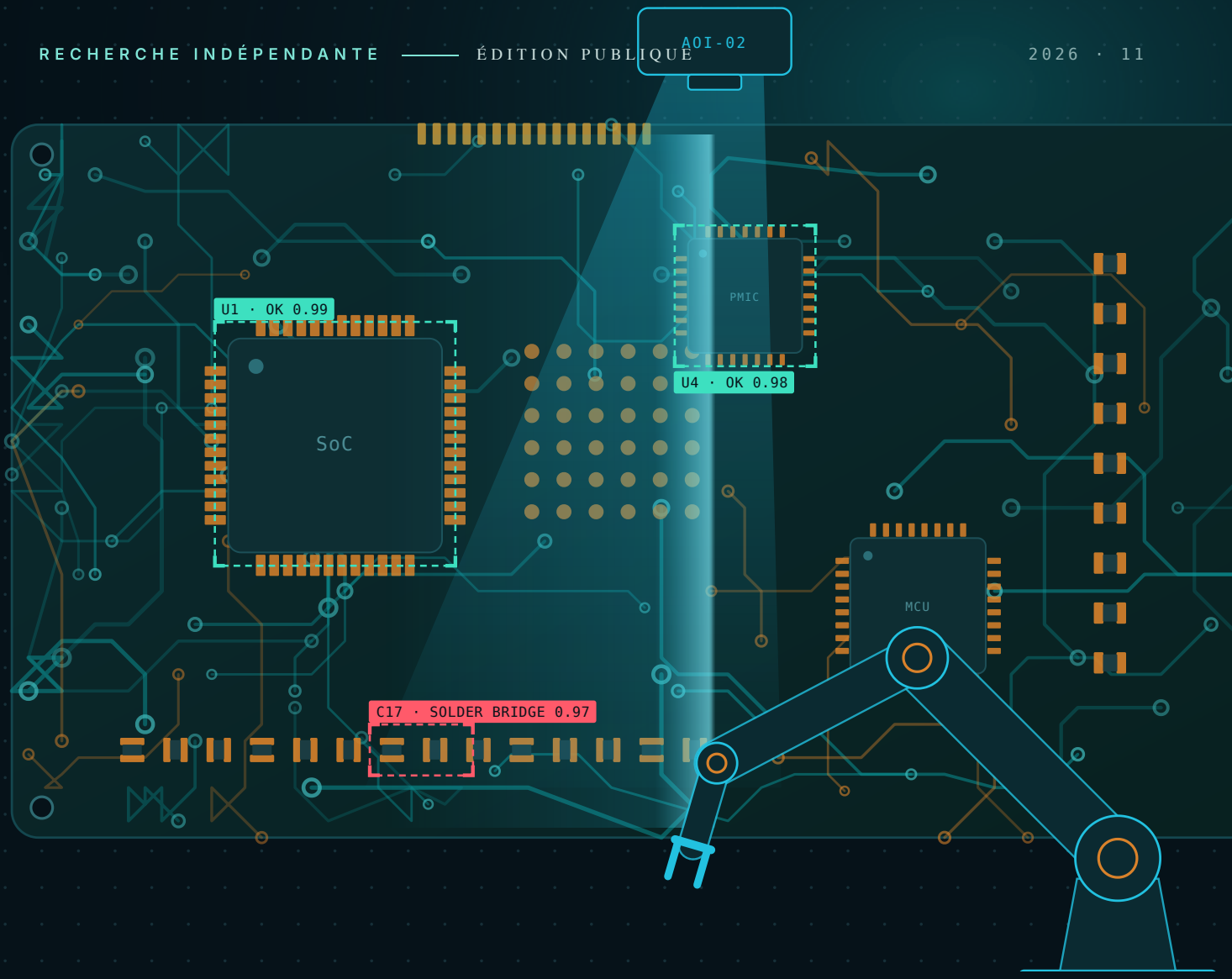


FABRICATION ÉLECTRONIQUE · INSPECTION PAR IA × ROBOTIQUE

Inspection visuelle par IA et bras robotisés dans les usines électroniques

Évaluation 2026

Configurations courantes · Faisabilité régionale · Déploiement sur site · Modèles enseignant–élève
Retour sur investissement · Concurrence · Ce que seul un pilote peut trancher

CONSTAT CLÉ

La revue IA vaut la peine d'être menée, de manière ciblée. Les bras robotisés incarnés peuvent attendre.

Tan Shuai
tanshuai.comi@tanshuai.com

Données au 11 novembre 2026
Chine continentale · Taïwan · Asie du Sud-Est · États-Unis

SOMMAIRE

Sommaire



◆ **RÉSUMÉ EXÉCUTIF** 5
Cela vaut la peine, mais de manière ciblée, maintenant et avec des partenaires

01 **La ligne aujourd'hui** 9
Ce que font tourner aujourd'hui les usines électroniques, et ce qui est techniquement faisable

02 **Sur site** 17
Faire tourner l'IA dans l'usine : salles blanches, coupures d'air, matériel et quantification

03 **Enseignant et élève** 26
Enseignant et élève : les modèles frontières peuvent pré-étiqueter les données d'inspection ; pour les bras robotisés, les vrais enseignants sont encore des personnes

04 **Coûts et retour sur investissement** 35
Coûts et retour sur investissement : la revue IA est rentabilisée en mois, l'alimentation machine en années, les bras de prélèvement jamais

05 **Marché et concurrence** 43
Marché, concurrence et comment les acheteurs paient

06 **IA incarnée** 52
IA incarnée : d'abord elle change la manière d'enseigner aux robots, plus tard leur manière de bouger

07 **Questions ouvertes** 59
Ce que les sources publiques ne peuvent répondre, et comment un pilote le règle

◆ **À propos de l'auteur** 62
Tan Shuai

Balises de niveau de confiance dans le texte

- **verified** deux sources indépendantes ou plus, ou deux documents primaires indépendants, concordent
- **single source** une seule source, y compris les documents primaires
- **vendor claim** publié par le vendeur du produit ou du service
- **estimate** calcul ou déduction de l'auteur à partir de données citées

Comment lire ce rapport

Les chiffres clés portent une balise de niveau de confiance. **verified** signifie que deux sources indépendantes ou plus concordent, ou que deux documents primaires indépendants concordent ; **single source** signifie une seule source (y compris les documents primaires) ; **vendor claim** signifie publié par le vendeur du produit ou du service, même largement repris ; **estimate** signifie mon propre calcul ou ma déduction à partir de données citées, avec les entrées sourcées. « Primaire » décrit le type de source (rapport annuel, prospectus, réglementation, norme, statistiques officielles) ; cela ne change pas la classe de confiance. Lorsque les sources divergent, les chiffres sont donnés côte à côte et non moyennés. Les conversions monétaires utilisent le cours pivot de la Banque populaire de Chine du 9 novembre 2026, 1 $ = 6,7188 RMB (Xinhua), et pour Taïwan le fixing de SMBC du même jour, 1 $ = 32,12 NT$ (SMBC). Les données de main-d'œuvre et de subventions continentales utilisent le delta de la rivière des Perles, où la fabrication électronique est la plus concentrée (données salariales officielles de Shenzhen), avec le delta du Yangtsé pour comparaison. Les affirmations issues d'entretiens sectoriels sont traitées comme des affirmations à vérifier, jamais comme des faits. Les questions auxquelles les sources publiques ne peuvent répondre sont rassemblées au chapitre 7.

FIGURES

Figures



01	Où l'IA s'insère aujourd'hui dans l'usine	7
02	Une ligne SMT en volume : les gens sont assis derrière l'AOI	10
03	Environ 97–99 % des alertes AOI sont de fausses alertes	11
04	Une caméra polyvalente n'est pas un AOI	13
05	Un déploiement isolé (air gap)	17
06	Les zones propres et ESD décident où le calcul peut siéger	18
07	Limites de vitesse par usage	20
08	Calcul sur site : mémoire, prix et ce qui est légalement achetable en Chine continentale	21
09	Une grosse station de travail ne fait pas un assistant partagé	22
10	Quantification : 8-bit est sûr, 4-bit dépend de la recette, 3-bit et moins est éliminé	23
11	Candidats enseignants : les scores génériques ont doublé en une génération	27
12	« Y a-t-il un défaut ? » reste difficile pour les modèles généraux	28
13	Deux chaînes enseignant–élève	29
14	Quelle taille doit avoir le jeu de référence	30
15	Latence de politique edge	32
16	Élèves robots : proches du niveau production seulement si le poste est structuré	33
17	Ce que coûte par an un agent d'inspection ou de manutention	36
18	L'échelle de prix	37
19	Rentabilité par région	39
20	Pourquoi l'extraction des cartes NG ne rentabilise pas	40
21	Droits américains sur les robots, AOI et ordinateurs edge	41
22	Les tailles de marché diffèrent de plus de dix fois	44
23	Ascendant : 40 000 lignes → 16 000 servies → ≈320 M RMB ponctuel	45
24	La fenêtre de retrofit est d'environ cinq ans	46
25	Le créneau de revue est encerclé sur trois côtés	46
26	La marge se situe dans le matériel industrialisé	48
27	Comment les acheteurs paient	49
28	Jours de ventes en attente	50
29	La pile d'IA incarnée	52
30	Ce que divulguent réellement les déploiements robot en usine	54
31	Main-d'œuvre nette remplacée par robot	54
32	Chronologie des normes et de la réglementation	56

RÉSUMÉ EXÉCUTIF

Cela vaut la peine, mais de manière ciblée, maintenant et avec des partenaires

CONSTAT CLÉ

Le verdict : pour les usines électroniques, ce qui est mature en 2026 et ce que les usines paieront, c'est la revue IA et l'inspection de niche sur les images de l'AOI, plus l'alimentation machine robotisée programmée, pas « un grand modèle qui regarde les cartes plus des robots incarnés qui les prélèvent ».

0.46–0.65 an
Retour sur investissement de la revue IA aux États-Unis, à Taïwan et à Shenzhen, à 50 000 RMB par ligne

≥95 % · <20 PPM
Couverture et taux d'échappement de la revue IA interne d'un grand EMS : le niveau à atteindre pour vendre aux EMS de premier rang

100–700 RMB M
Marché de rétrofit tiers adressable en Chine continentale, ponctuel, sur une fenêtre d'environ cinq ans

0 scores
Scores publiés aux benchmarks de défauts industriels pour les modèles phares de novembre 2026 : des enseignants, pas des inspecteurs

1.1–1.5 an
Retour sur investissement d'une cellule cobot d'alimentation machine sur deux équipes aux États-Unis, à Taïwan et à Shenzhen

0.003–0.14 FTE
Main-d'œuvre économisée par équipe par un bras robotisé de prélèvement sur une ligne SMT : jamais rentabilisé

Inspection
ROI

L'apprentissage profond supervisé appliqué aux images qu'un AOI existant produit déjà réduit les fausses alertes ou les heures de revue de 60 à 90 % (déclarations des vendeurs). DBG, un grand EMS, indique dans son rapport annuel que son système interne couvre plus de 95 % des composants signalés, avec des échappements inférieurs à 20 PPM et plus de 85 % de revue humaine en moins ; c'est le niveau qu'un produit tiers doit atteindre pour vendre aux EMS de premier rang. À 50 000 RMB par ligne, la revue IA est rentabilisée en moins d'un an en Chine continentale, à Taïwan et aux États-Unis (Shenzhen 0,65 an, Taïwan 0,53, États-Unis 0,46), les cellules cobots d'alimentation machine en 1,1–1,5 an et l'Asie du Sud-Est en 2–6 ans, tandis qu'un bras robotisé ajouté à une ligne SMT uniquement pour prélever les cartes défectueuses n'est jamais rentabilisé nulle part, car au-dessus de 90 % de rendement au premier passage, le prélèvement ne mobilise que 0,003–0,14 personne par équipe estimate.

Modèles
ROBOTS

Les grands modèles multimodaux ne peuvent toujours pas être l'inspecteur principal. Au 11 novembre 2026, aucun des modèles phares (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 et autres) ne publie de score sur un benchmark de défauts industriels ou de PCB, et dans le défi PCBA d'ACM MM 2026, les cinq meilleurs systèmes, entraînés sur 6 600 images réelles de ligne, n'ont obtenu que 77–82 sur « y a-t-il un défaut ». Les modèles frontières peuvent pré-étiqueter et trier. Pour les bras robotisés, les vrais enseignants restent les démonstrations humaines, l'apprentissage par renforcement sur robot réel et la simulation : les élèves atteignent 90–99,4 % sur des postes fixes et

structurés, mais seulement 30–80 % sur les convoyeurs en mouvement, les pièces jamais vues et les nouveaux produits ; un agent frontière a réalisé 104 des 108 tâches quasi statiques sur robot réel, à 280 secondes chacune.

Sur site
DÉPLOIEMENT

Le déploiement isolé (air gap) dans des usines propres et contrôlées est tout à fait pratique. Un atelier SMT ordinaire est généralement une zone protégée ESD plutôt qu'une salle blanche classée : un Jetson sans ventilateur ou un boîtier NPU national se place sur la ligne et les serveurs GPU vont en salle des serveurs. Selon les propres classifications d'exportation de NVIDIA, Jetson Thor et DGX Spark peuvent être vendus à une usine continentale ordinaire sans licence américaine ; DGX Station ne le peut pas. La précision par défaut est de 8 bits ; le 4 bits quantifié par le vendeur est utilisable sous conditions ; le 3 bits et en dessous est exclu.

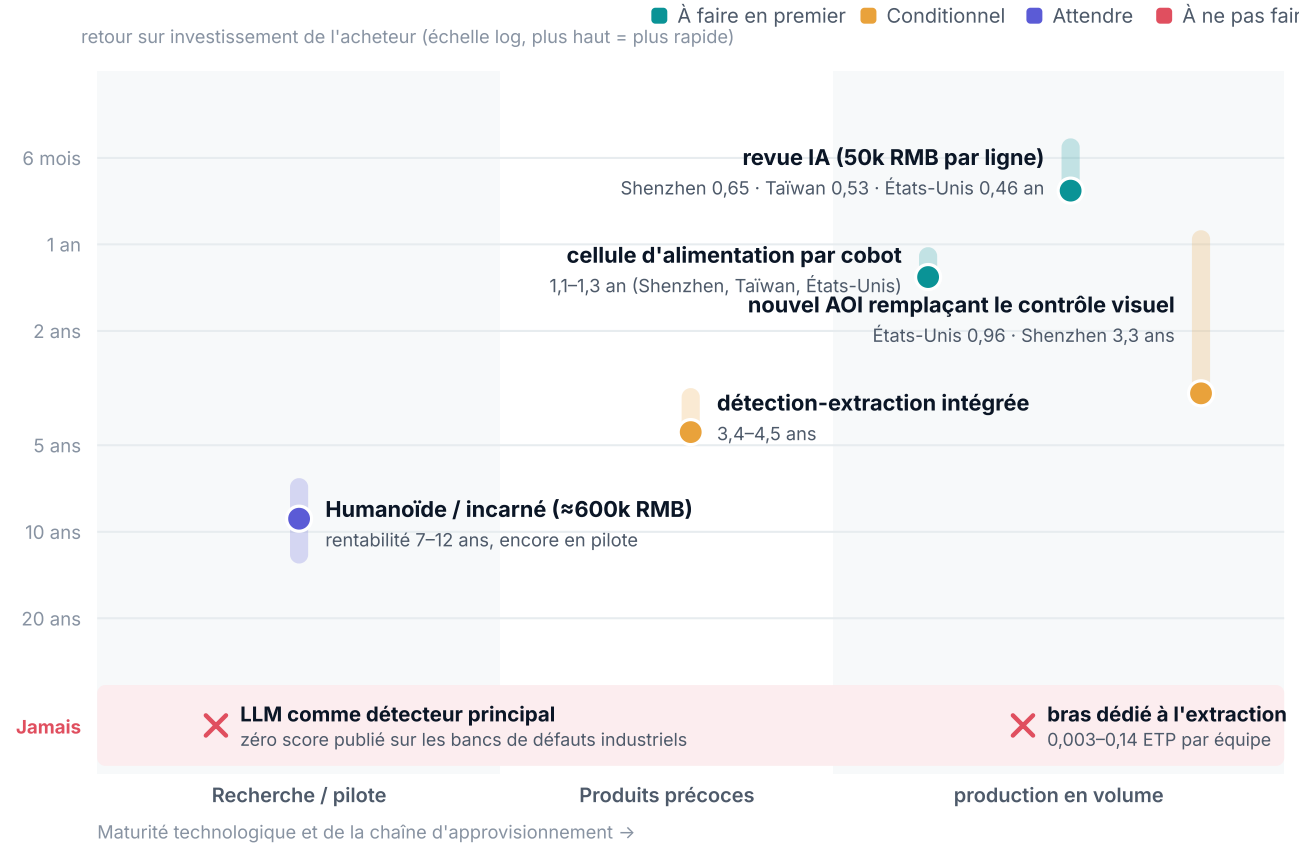
Marché
CONCURRENCE

Le marché est réel mais limité. La Chine continentale compte environ 35 000–46 000 lignes SMT et environ 26 000–66 000 AOI en ligne qui n'utilisent pas l'IA ; le marché adressable pour les rétrofits tiers est d'environ 100–700 millions de RMB ponctuels, sur une fenêtre d'environ cinq ans, et les fabricants d'AOI ont commencé à intégrer la revue dans les nouvelles machines estimate. La concurrence est déjà forte : le créneau de la revue est encerclé par les fabricants d'AOI, les équipes internes des EMS et les fournisseurs de plateformes, et la marge se situe dans le matériel industrialisé.

Question	Jugement	Chiffres clés (sources et niveau de confiance dans les chapitres)
① **Que font tourner aujourd'hui les usines électroniques courantes ?**	Une ligne SMT en volume est « SPI 3D + AOI avant et après refusion + rayons X par échantillonnage + ICT/FCT + réparation et inspection visuelle », avec une revue humaine derrière chaque AOI. L'IA entre dans la revue de trois manières : intégrée par les fabricants d'AOI, construite en interne par les grands EMS, ou ajoutée par des tiers. L'assemblage et la manutention utilisent des bras, des SCARA et des AMR avec perception apprise et mouvement programmé ; l'IA incarnée est encore au stade pilote	Takt ≈ 13–22 s par carte ; un ODM a budgété 30,6 M RMB d'AOI pour 18 lignes, ≈ 3–5 AOI par ligne ; ≈ 97–99 % des alarmes d'AOI sont de fausses alertes
② **Jusqu'où va la technologie, l'argent peut-il acheter zéro personne, et que peut utiliser chaque région ?**	L'inspection est de niveau industriel sous forme « détecteur dédié + revue IA + quelques personnes » ; même un argent illimité n'atteint pas zéro personne. Le continent n'utilise que des modèles nationaux ou des poids ouverts locaux ; les usines privées de Taïwan peuvent utiliser presque tout, mais les fonderies et les infrastructures critiques excluent en pratique les marques continentales ; l'Asie du Sud-Est a le moins de restrictions ; les États-Unis ont toute la pile plus les droits de douane et la certification	La revue IA réduit les fausses alertes ou les heures de revue de 60 à 90 % vendor claims ; une caméra 4K sur un panneau entier ≈ 78 µm par pixel contre 8–20 µm pour un AOI ; les bras fabriqués sur le continent paient ≈ 40 % de droits de douane américains cumulés, les instruments de type AOI ≈ 37,5 %
③ **Peut-il tourner sur site dans les salles blanches et les usines isolées, sur quel matériel, assez vite et assez précisément ?**	Oui ; les contraintes sont procédurales, pas liées au calcul. Un boîtier sans ventilateur ≤ 100 W sur la ligne, des serveurs GPU en salle des serveurs ; Thor pour les postes, un Mac Studio ou un appareil national pour les questions-réponses de l'usine ; pas de DGX Station sur le continent. La vitesse est correcte quand la taille du modèle correspond à la tâche ; garder le 8 bits, utiliser le 4 bits quantifié par le vendeur sous conditions, jamais le 3 bits	≈ 0,5 s par recadrage pour un modèle de vision 8B sur Thor estimate ; kit de développement Thor 59 104 RMB en Chine ; DGX Station 123–140 k $, 139 jetons/s au total pour 32 utilisateurs sur un modèle 743B

Question	Jugement	Chiffres clés (sources et niveau de confiance dans les chapitres)
④ **Le modèle phare le plus récent peut-il servir d'enseignant sur des dizaines de tours et être distillé en élèves locaux, y compris pour la fabrication flexible et les bras robotisés ?**	Inspection : pré-étiquetage et tri ; n'accepter automatiquement les étiquettes de l'enseignant que pour les classes avec « ≥ 299 défauts, zéro manqué » sur un jeu de référence. Robots : les enseignants restent les démonstrations humaines, le RL sur robot réel et la simulation, et les élèves n'approchent le niveau production que sur des postes structurés	Aucun modèle phare de novembre n'a de score de défaut industriel ; élèves à 97,8 % sur des connecteurs déjà entraînés, 29–67 % sur des inconnus, 37–59 % sur des convoyeurs en mouvement ; modèles frontières pilotant directement les articulations 0–5,5 %
⑤ **Combien cela coûte-t-il, et quel est le retour sur investissement par région ?**	La revue IA est rentabilisée en moins d'un an en Chine continentale, à Taïwan et aux États-Unis ; l'alimentation cobot en 1,1–1,5 an ; l'Asie du Sud-Est en 2–6 ans ; un bras de prélèvement sur une ligne SMT jamais	Boîtier de revue à 50 000 RMB par ligne : Shenzhen 0,65 an, Taïwan 0,53, États-Unis 0,46, Vietnam 2,2 estimate ; prix acceptable d'un module tiers ≈ 10–30 k RMB par ligne ponctuel ou 5–10 k RMB par an
⑥ **Les concurrents sont-ils déjà assez forts, et vaut-il mieux commencer maintenant ou attendre ?**	La demande n'est pas saturée, mais l'offre est encombrée ; l'IA des OEM limite la fenêtre de rétrofit à environ cinq ans. Entrer maintenant avec un produit étroit via des partenaires de distribution, sans affronter de front les fournisseurs de plateformes ; attendre les données de production de 2027–2028 avant de parier sur l'IA incarnée	≈ 26–66 k AOI en ligne sans IA sur le continent, en recul de 8–12 % par an estimate ; le principal fournisseur de vision IA a réalisé un chiffre d'affaires 2025 de 1,086 Md RMB et une perte nette de 991 M RMB ; le chiffre d'affaires 3C du comparable coté le plus proche a chuté de 79,9 % sur un an en janvier–mai 2026

FIG 01 **Où l'IA s'insère aujourd'hui dans l'usine** ***La revue IA des fausses alertes AOI et l'alimentation de machines par cobot sont matures et vite rentabilisées ; un bras d'extraction dédié et un LLM comme inspecteur principal n'en valent pas la peine.***



Source : Rentabilité du modèle de l'auteur (chapitre 4) ; l'axe de maturité est un jugement de l'auteur.

Ce qu'il faut construire et ce qu'il ne faut pas construire

	Recommandation
Construire	La revue IA et la classification des défauts sur images AOI, pour les EMS de taille moyenne disposant d'AOI anciens ou de marques mixtes et sans mise à niveau IA du fabricant d'AOI ; l'inspection que l'AOI SMT couvre mal (composants traversants et soudure manuelle, vernis de tropicalisation, câbles, connecteurs et FPC, aspect final et vis) ; les cellules cobots de chargement/déchargement où chaque pièce est manipulée (montages de test, plateaux, emballage), avec l'inspection en complément ; un jeu de référence et un mode fantôme sur une ligne réelle comme base de chaque affirmation
Ne pas construire	Une plateforme générale « IA plus bras robotisé » ; un LLM multimodal comme inspecteur principal ; un bras robotisé sur une ligne SMT uniquement pour prélever les cartes NG (un aiguillage NG ou un tampon suffit) ; les humanoïdes et les robots-en-tant-que-service ; les modèles frontières dans une boucle de contrôle ou une fonction de sécurité ; des piles de mémoire GPU aux postes pour faire tourner des modèles « pleine taille » ; la vente directe aux fonderies, aux infrastructures critiques taïwanaises ou aux usines électroniques de défense américaines
Avec qui travailler	Les fabricants d'AOI et les distributeurs (canal et interfaces d'image ; les distributeurs ont apporté 46,68 % du chiffre d'affaires 2025 de Sinic-Tek) ; les fabricants de cobots (bras certifiés et fonctions de sécurité) ; les EMS de 10 à 40 lignes (expansion progressive et acomptes) ; les intégrateurs locaux en Thaïlande et ailleurs (pour respecter les règles de lien local du BOI)

La ligne aujourd'hui

01 Ce que font tourner aujourd'hui les usines électroniques, et ce qui est techniquement faisable

La ligne en volume standard : plusieurs AOI, rayons X par échantillonnage, revue centralisée, 13 à 22 secondes par carte

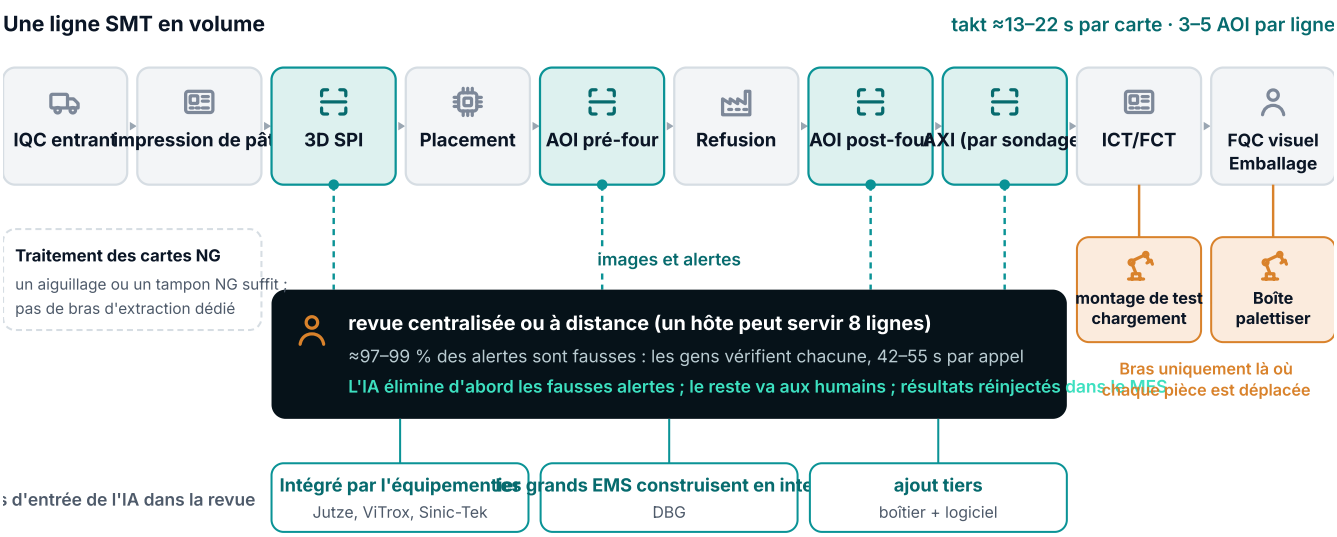
Les documents primaires déposés par les ODM cotés montrent à quoi ressemble réellement une ligne en volume en 2025–2026. Le prospectus de Hong Kong de Longcheer (données au 30 septembre 2025) liste ses contrôles qualité SMT comme « inspection par rayons X, inspection de la pâte à braser (SPI), inspection optique automatisée (AOI), test fonctionnel PCBA et inspection visuelle » ; son flux de processus fait tourner l'AOI avant refusion, l'AOI après refusion, les rayons X par échantillonnage, les cartes NG vers un poste de réparation SMT, puis un contrôle visuel final. Ses usines de Huizhou, Nanchang et du Vietnam comptent 34, 24 et 8 lignes SMT et 67, 60 et 13 lignes d'assemblage final (prospectus H de Longcheer) single source, primary filing. L'usine d'ordinateurs portables 2023 de Huaqin prévoyait 18 lignes SMT avec un budget AOI de 30,6 millions de RMB, soit environ 1,7 million de RMB par ligne, ce qui, aux prix nationaux des AOI 3D, représente trois à cinq AOI par ligne pour l'inspection avant et après refusion de cartes double face (prospectus A de Huaqin) primary filing + estimate. Une ligne en volume comporte donc généralement deux à quatre points de revue, et non un seul.

Les machines sont déjà très automatisées : Huaqin fait état de 96 % d'automatisation sur ses lignes SMT (prospectus H de Huaqin) single source. Les personnes restantes sont dans la revue, la réparation, l'inspection visuelle, le test et l'assemblage final, et aucune entreprise ne publie le nombre de réviseurs par ligne (chapitre 7).

Trois estimations indépendantes concordent sur un takt de **13 à 22 secondes par carte** (un dossier environnemental 2026 pour cinq lignes du Hebei (EIE du projet SMT du Hebei), les lignes d'ordinateurs portables de Huaqin et un cas iFactory), soit 7 à 11 secondes par face pour les cartes double face estimate. La difficulté d'une ligne à forte mixité n'est pas le volume mais l'introduction de nouveaux produits et le changement de série, et chaque changement de série est un nouveau départ à froid pour un modèle d'IA.

Côté logiciel, les propres postes de revue des fabricants d'AOI sont déjà centralisés : le logiciel de poste de réparation de Sinic-Tek permet « à un hôte de contrôler l'équipement de huit lignes SMT » (prospectus Sinic-Tek) single source, primary, 2022, et après le déploiement par Jabil du contrôle AOI à distance ViTrox V-ONE en 2026, un ingénieur pouvait gérer 5 à 8 AOI à distance contre 1 à 2 localement (AtlasPCB) vendor claim. Un produit de revue IA tiers se branche exactement sur cette station centrale, via l'export d'images et de résultats de l'AOI (IPC-CFX, IPC-HERMES-9852 ou le format propre du vendeur). Les AOI eux-mêmes intègrent l'IA rapidement : Yamaha indique que son jugement IA prend environ une seconde par appel, aide les inspecteurs après 30 images NG et devient de « très haute » qualité après 1 000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Une ligne SMT en volume : les gens sont assis derrière l'AOI*La main-d'œuvre se concentre aux postes de revue après l'inspection optique automatisée. C'est là que l'IA entre en jeu.***



Source : Documents publics et supports fournisseurs (Huaqin, DBG, Sinic-Tek et autres) ; compilé par l'auteur.

Étape	Produits et technologies courants	Fournisseurs représentatifs	Où en est l'IA
Après impression	SPI 3D (lumière structurée, moiré, un peu de triangulation laser)	Koh Young, Sinic-Tek, TRI	Jugement de pâte IA des OEM en production ; Jutze dit que certains modèles sont « entièrement IA »
Après placement et refusion	AOI 2D/3D avec LED multi-angles multicolores et lumière structurée, 8–20 µm par pixel, caméras 8–12 MP	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek et autres	Revue IA des OEM (Jutze en production, ViTrox V510Ai, Sinic-Tek avec Tencent) ; les grands EMS construisent les leurs (DBG) ; modules tiers
Joints à terminaison inférieure et cachés	Rayons X par échantillonnage ou en ligne (AXI)	—	Hikvision affirme que son modèle Guanlan se trompe sur les vides aux rayons X moins de 1 % du temps vendor claim
Revue et réparation	Station de revue centrale (un hôte, huit lignes), revue à distance, poste de réparation	Logiciels des fabricants d'AOI, ViTrox V-ONE	L'IA clôt les fausses alertes, les personnes traitent le reste
Inspection visuelle, IQC, FQC	Surtout manuel ; quelques machines d'aspect IA avec chargement robotisé	SmartMore et autres	Machines d'aspect en production ; pas de données publiques d'effectifs pour l'IQC/FQC
Manutention et alimentation machine	Bras six axes et cobots, SCARA, AMR, guidage vision 2D/3D	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite ; Mech-Mind, Hikrobot	Perception apprise plus mouvement programmé en production ; VLA et humanoïdes encore au stade pilote
Données et systèmes	MES, IPC-CFX, IPC-HERMES-9852	—	Résultats réécrits dans le MES ; le support sur les équipements anciens est inégal

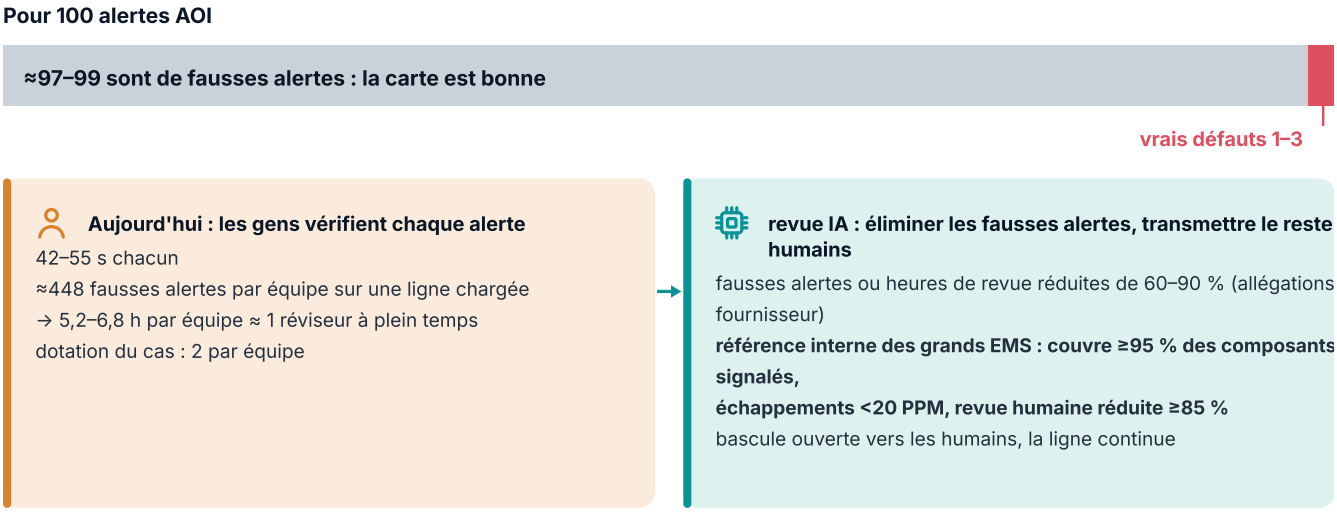
Sources du tableau : les documents ci-dessus et (Koh Young, rapport annuel 2025 de Jutze, rapport annuel 2025 de Sinic-Tek, ViTrox, Sohu sur Hengtianyi, Xinhua sur Hikvision).

Les personnes sont derrière l'AOI, car la plupart des alarmes sont fausses

La chaîne d'inspection d'une ligne SMT/PCBA courante est : SPI 3D après impression pour la pâte manquante, insuffisante ou excédentaire, les ponts et le décalage ; AOI après placement (facultatif) et après refusion pour les pièces manquantes, décalées, tournées, inversées ou « tombstone » et le volume de soudure ; rayons X lorsqu'il y a des BGA ou d'autres pièces à terminaison inférieure ; puis test en circuit ou à sonde mobile, test fonctionnel et inspection visuelle manuelle. Chaque AOI a un poste de revue et un poste de réparation derrière lui (prospectus Sinic-Tek, JLCPCB) verified. L'AOI après refusion est le plus courant et le plus rentable, mais il ne voit que la forme et la surface ; les joints, les vides et le mouillage sous un BGA exigent les rayons X ou un test électrique (JLCPCB).

La main-d'œuvre est dans la revue parce que la plupart des alarmes sont fausses. Un EMS britannique a enregistré **54 019 fausses alertes contre 544 défauts réels sur 917 références, soit environ 99 % des alarmes** (arXiv 2505.03576) single source, preprint ; à l'usine de Rastatt de Siemens, **97,5 % des cartes rejetées par l'AOI étaient bonnes** (étude de cas Siemens) vendor claim. Une ligne chargée voit environ 448 fausses alertes par équipe à 42–55 secondes chacune, soit 5,2–6,8 heures par équipe, environ un réviseur à temps plein (iFactory) vendor claim; hours are an estimate.

FIG 03 **Environ 97–99 % des alertes AOI sont de fausses alertes** ***Aujourd'hui, une personne confirme chaque appel ; la revue IA élimine d'abord les fausses et transmet le reste aux humains.***



Source : Étude de cas Siemens, iFactory (allégation fournisseur), rapport annuel DBG 2025 ; heures de revue estimées par l'auteur.

À partir de « heures de revue par équipe = cartes × fausses alertes par carte × secondes par appel », un réviseur peut couvrir environ 0,7 à 1 ligne en volume avec un AOI à base de règles. Une fois que l'IA supprime 60 à 90 % des fausses alertes, le seul temps de jugement permettrait 2 à 14 lignes par personne, mais les déplacements, la gestion du tampon NG et les passations de réparation limitent la couverture locale à deux à quatre lignes adjacentes ; seule la revue centralisée ou à distance s'approche du plafond théorique estimate; seconds per call come from one source. Comme les AOI courants balaient 60–80 cm² par seconde (guide AOI 2026 de Yunheng), une face de 20×15 cm prend 4–5 secondes ; l'IA n'a qu'à traiter les quelques régions signalées dans un takt, pas à rescanner la carte.

Ce qui fonctionne déjà en production : la classification supervisée sur les images propres de l'AOI

Le schéma éprouvé est la classification par apprentissage profond supervisé sur les images de l'AOI qu'une usine possède déjà. Le module complémentaire de Fujitsu sur l'AOI d'ADLINK à Taïwan a réduit les fausses alertes de **plus de 90 %** (cas Fujitsu) ; OKI utilise l'IA depuis juillet 2026 pour réduire la confirmation manuelle après AOI d'environ **80 %** dans sa propre activité EMS (Circuits Assembly) ; Siemens fait état de 60 % d'analyse manuelle en moins, 160

287 € économisés la première année et un retour sur investissement de huit mois (étude de cas Siemens) all vendor claims. Deux points sont communs à ces cas : presque aucun ne publie de taux d'échappement, et tous gardent des personnes sur les alarmes restantes. Le seul cas qui donne un chiffre d'échappement (de 0,8 % à 0,3 %, contre un 0,5 % contractuel) est incohérent en interne (iFactory). L'IA réduit les heures de revue, pas le poste de revue lui-même.

La revue IA est passée des modules tiers aux fonctionnalités OEM et aux systèmes internes, ce qui change l'étalon auquel un nouvel entrant est mesuré. Le rapport annuel 2025 de Jutze liste sa « mise à niveau AOI 3D » comme en production de volume et dit qu'elle « utilise des algorithmes d'IA pour supprimer la revue manuelle côté inspection » (rapport annuel 2025 de Jutze) ; Sinic-Tek développe un « système de revue manuelle assisté par IA multimodale » depuis 2019 et co-développe un AOI 3D haut-et-bas avec Tencent, visant une couverture des défauts de 99,5 %+ et une précision de reconnaissance de 98 % (rapport annuel 2025 de Sinic-Tek) vendor claims in primary filings. Les grands EMS construisent les leurs. Le rapport annuel 2025 de DBG liste le « jugement IA pour l'inspection visuelle » comme projet achevé qui « couvre plus de 95 % des composants signalés par l'AOI », avec un « taux d'échappement inférieur à 20 PPM » et « plus de 85 % de revue manuelle en moins », entièrement traçable (rapport annuel 2025 de DBG) single source, primary. C'est le seul chiffre d'usine électronique que j'ai trouvé qui donne ensemble couverture, échappements et réduction. **Un produit tiers vendu aux EMS de premier rang doit égaler la couverture de 95 %, les 20 PPM d'échappements et la réduction de 85 % de DBG** ; en dessous du premier rang, l'étalon est la réduction de 50 à 85 % que les vendeurs annoncent.

La détection d'anomalies entraînée uniquement sur des pièces bonnes est rapide (EfficientAD-S prend 2,2 ms par image (EfficientAD, WACV 2024)) mais faible sur les anomalies « logiques » comme les pièces manquantes ou erronées, et les meilleures méthodes atteignent en moyenne moins de 60 % d'AU-PRO sur le plus difficile MVTec AD 2 (MVTec AD 2) single source. Les pièces manquantes et erronées se vérifient mieux contre la CAO et la nomenclature (BOM).

Le plafond à tout prix : des couches de machines, très peu de personnes, mais pas zéro

Je n'ai trouvé aucune usine électronique qui a supprimé entièrement la revue manuelle et atteint zéro échappement. La pratique pour les produits IPC-A-610 Classe 3 (haute fiabilité) fait encore décider par des inspecteurs formés les cas limites que les seuils automatiques ne peuvent classer, et un assembleur de PCB qualifie « l'inspection automatisée peut remplacer la validation humaine » de « malentendu courant » (PCBCart) single source. Le niveau de référence humain doit aussi être corrigé : citant une étude de Sandia de 2012, un inspecteur visuel seul attrape au mieux environ **80 %** des défauts, deux en série environ 96 % (Instrumental) single source, cited by a vendor. L'affirmation défendable est « l'IA plus une petite boucle de revue humaine bat l'inspection entièrement manuelle », pas « l'IA remplace l'inspection ».

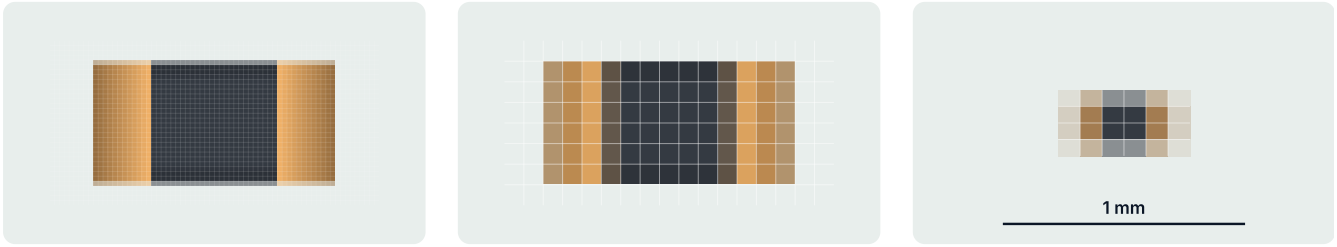
Sans limite de budget, le plafond est SPI 3D, AOI 3D, rayons X, revue IA et postes d'aspect IA empilés, avec des personnes uniquement pour les cas limites, les nouveaux produits et changements de série, les nouveaux types de défauts et la validation client. Les usines phares font état de 85 à 92 % d'inspection manuelle en moins (I-Connect007, 2021, WiAdvance) vendor claims ; mon estimation du plafond sans budget est une main-d'œuvre de revue à 8–15 % de celle d'aujourd'hui estimate. L'assemblage final reste intensif en main-d'œuvre : l'automatisation des boutons de l'iPhone 16 a été abandonnée en raison de taux de défauts élevés (9to5Mac) single source.

Une caméra haute résolution générale n'est pas un AOI

Les AOI de production imagent tuile par tuile à 8–20 µm par pixel, avec des caméras 8–12 MP, un éclairage programmable multi-angles multicolores et une 3D par lumière structurée (Koh Young, Saki) verified. Le secteur a noté depuis longtemps qu'une pièce 01005 ne couvre qu'environ 8×16 pixels à 25 µm par pixel, « possiblement insuffisant » pour un jugement fiable (Circuits Assembly, 2007). Par le même calcul estimate, une caméra 4K sur un panneau de 300 mm obtient environ **78 µm par pixel** : une pièce impériale 0402 fait environ 13×6 pixels, assez pour la présence et un grossier décalage mais pas pour la forme du joint, et une 01005 fait environ 5×3 pixels, pratique-

ment invisible. Atteindre 15 µm par pixel sur un panneau entier exigerait environ 330 mégapixels, ce qui explique pourquoi les AOI déplacent une caméra sur la carte. Quand quelqu'un dit « 0402 », vérifiez lequel il désigne : le 0402 métrique est le 01005 impérial, bien au-delà de toute caméra grand angle.

FIG 04 **Une caméra polyvalente n'est pas un AOI** ***Une caméra 4K sur un panneau de 300 mm donne ≈78 µm par pixel : un 0402 se réduit à ≈13×6 pixels et un 01005 disparaît.***



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

Caméra 4K sur un panneau de 300 mm : ≈78 µm/px
Le 0402 se réduit à ≈13×6 px : juste assez pour la présence et un décalage grossier

Même caméra sur un 01005
≈5×3 px : effectivement invisible

Source : Circuits Assembly ; calcul de l'auteur pour un panneau de 300 mm et un capteur 4K. Les grilles de pixels sont illustratives.

Sur les lignes qui ont déjà des AOI, le meilleur emplacement de l'IA est au-dessus des images de l'AOI. Seules les lignes sans AOI, ou les postes de fin de ligne, justifient un poste caméra fermé avec éclairage contrôlé ; une caméra haute résolution ouverte au-dessus de la ligne est la plus exposée à la lumière ambiante et aux ombres (CSDN).

Assemblage et manutention : perception apprise plus mouvement programmé en production

Le rapport 2026 de l'IFR dit que « la vision industrielle pour l'inspection et le guidage des robots est mature et de niveau industriel », tandis que « l'IA physique généraliste et l'apprentissage par simulation et expérience ne sont pas encore de niveau industriel » (résumé exécutif World Robotics 2026 de l'IFR) single source, primary . L'électronique est le plus grand acheteur de robots (142 235 unités installées dans le monde en 2025, 59 % de toutes les installations en Chine (IFR, communiqué IFR Chine)), pourtant la Chine n'a que 166 robots pour 10 000 travailleurs manufacturiers contre 307 aux États-Unis (IFR, avril 2026). Le schéma en production est perception apprise plus mouvement déterministe ; le mouvement appris de bout en bout n'est proche du niveau production que sur des postes structurés (chapitres 3 et 6).

Ce que chaque région peut utiliser

La Chine continentale s'appuie sur les modèles nationaux et les poids locaux ; à Taïwan, c'est l'identité du client qui décide ; aux États-Unis, les droits de douane et la certification.

Élément	Chine continentale	Taïwan (usines privées)	Asie du Sud-Est	États-Unis
Claude (tout canal)	● Non disponible	● Conditionnel : le fournisseur ne doit pas être majoritairement détenu ou contrôlé par des personnes ou entités de Chine continentale, de Hong Kong ou de Macao	● Idem	● Idem
Cloud GPT et Gemini	● Non disponible	● Disponible	● Disponible (listes non vérifiées pays par pays)	● Disponible

Élément	Chine continentale	Taïwan (usines privées)	Asie du Sud-Est	États-Unis
Modèles cloud chinois (Qwen, DeepSeek, Doubao, GLM, Kimi)	● Disponible (modèles déposés ; proposer un service exige un enregistrement)	● Légal, mais l'hébergement continental passe rarement l'examen de sécurité des clients ; interdit dans le secteur public, les entreprises d'État et les infrastructures critiques	● Légal (nœuds internationaux)	● Légal, examen strict
Poids ouverts locaux d'origine chinoise (Qwen, GLM, DeepSeek)	● Premier choix ; un déploiement privé en usine n'exige pas de dépôt de modèle ou d'algorithme (lecture de cabinets d'avocats)	● Disponible (secteur public restreint)	● Disponible	● Conditionnel : examen de l'origine
Poids ouverts locaux d'origine non chinoise (par ex. Gemma 4)	● Disponible	● Disponible	● Disponible	● Disponible
Modèles de vision classiques non génératifs	● Disponible (hors règles sur l'IA générative)	● Disponible	● Disponible	● Disponible
Jetson Thor/T5000/T4000, DGX Spark	● Disponible : NVIDIA les classe 5A992.c, pas de licence américaine pour une usine civile ordinaire ; Spark a une SKU Chine ; prix canal environ 40–60 % au-dessus du tarif américain	● Disponible	● Disponible	● Disponible
RTX 5090, RTX PRO 6000, DGX Station	● Non disponible : les deux premiers sont 4A090.a, DGX Station 5A992.z ; les alternatives légales sont 4090D, 5090 D V2, RTX PRO 5000 72 Go, RTX 6000D, L20 (4A994.l)	● Disponible	● Disponible, mais les acheteurs dont le siège est en Chine ou détenus en dernier ressort depuis la Chine comptent comme chinois ; la Malaisie exige des permis d'exportation et de transbordement	● Disponible
Mac Studio M5 Ultra 512 Go	● Disponible (boutique Apple Chine)	● Disponible	● Présumé disponible	● Disponible
Huawei Ascend et autres accélérateurs chinois	● Disponible	● Risque élevé	● Risque élevé	● Inutilisable en pratique
Bras fixes, AOI, caméras industrielles fabriqués sur le continent	● Disponible	● Importable ; depuis le 1er décembre 2025, les autorités peuvent agir directement sur les marques continentales, et les agences, entreprises d'État, certaines fondations et fournisseurs d'infrastructures critiques peuvent être restreints ou exclus	● Disponible	● Légal, mais droits cumulés d'environ 40 % sur les bras, 37,5 % sur les instruments de type AOI et les caméras, 25 % sur les ordinateurs edge
AMR et manipulateurs mobiles fabriqués à l'étranger	● Disponible	● Disponible	● Disponible	● Les nouveaux modèles exigent une approbation conditionnelle du ministère américain de la Guerre

Élément	Chine continentale	Taïwan (usines privées)	Asie du Sud-Est	États-Unis
Caméras Hikvision et Dahua	● Disponible	● Usage privé autorisé, secteur public interdit	● Disponible	● Nouvelles autorisations FCC bloquées ; interdites pour les sous-traitants fédéraux
Sécurité et certification des robots	● GB 11291.1-2011 et GB 11291.2-2013 obligatoires, en révision depuis avril 2026 ; certification CR volontaire ; GB/T 39404-2026 couvre la cybersécurité des contrôleurs	● Clôture par défaut ; la collaboration sans clôture doit satisfaire CNS 14490 ou ISO 10218 avec une évaluation écrite	● Selon le client et la destination d'exportation, confirmé par écrit avant expédition	● L'équipement électrique doit être « approuvé » (en pratique une marque NRTL ou une évaluation sur site) ; UL 1740 exige que les cellules robotisées satisfassent ISO 10218-2 depuis le 11 août 2025
Fonderies et usines de produits contrôlés	—	● Les contrats d'achat de TSMC exigent SEMI E187 ; l'équipement continental est retiré des fonderies 2 nm	—	● Les données techniques ITAR incluent les « photographies » ; politique de refus pour la Chine ; les fournisseurs contrôlés par le continent ne peuvent en pratique pas servir les usines électroniques de défense

Sources du tableau : (régions prises en charge par Anthropic, pays pris en charge par OpenAI, disponibilité de Gemini, mesures provisoires de la CAC sur l'IA générative, données de classification d'exportation NVIDIA, eCFR 15 CFR 774, NVIDIA Chine DGX Spark, Apple Chine, Covington sur les nouvelles règles BIS, Greenberg Traurig, juin 2026, Fulcrum, Tom's Hardware sur les consignes Ascend, liste d'importation BOFT de Taïwan, historique de la loi taïwanaise sur la gestion de la cybersécurité, règles d'examen taïwanaises pour les produits menaçant la cybersécurité nationale, Taipei Times, USITC HTS chapitre 99, Latham & Watkins, IPVM, plateforme des normes nationales chinoises, GB 11291.2, mise à jour Intertek UL 1740, eCFR 22 CFR 126.1).

Chine continentale : une pile complète, mais ni les modèles frontières fermés ni les machines individuelles les plus puissantes ne peuvent entrer. Les modèles cloud nationaux déposés, les poids ouverts locaux (Qwen sous Apache-2.0, GLM et DeepSeek sous MIT), les modèles de vision classiques et les AOI, caméras et cobots nationaux sont tous utilisables. Les propres classifications d'exportation de NVIDIA listent le kit Jetson AGX Thor, les modules T5000/T4000 et DGX Spark en 5A992.c, de sorte qu'une usine EMS ou PCBA ordinaire n'a pas besoin de licence américaine, tandis que les RTX 5090 et RTX PRO 6000 (4A090.a) et DGX Station (5A992.z, TPP 58 484) ne peuvent être achetés sur le continent (classifications d'exportation NVIDIA, eCFR) primary; the legal reading is mine and not legal advice. RS China liste le kit Thor à 59 104 RMB taxes comprises, soit environ 1,4 fois le tarif américain (RS China).

Quatre cabinets d'avocats s'accordent à dire qu'un déploiement privé en usine pour l'usage propre de l'usine n'exige aucun dépôt de modèle ou d'algorithme, tandis qu'un service hébergé multi-locataires se situe dans une zone grise (Covington, FPF, JunHe) verified: consistent readings, no official ruling. GB 11291.1-2011 et GB 11291.2-2013 restent la référence obligatoire de sécurité des robots et sont en révision depuis le 28 avril 2026 ; GB/T 39404-2026 sur la cybersécurité des contrôleurs est entrée en vigueur le 1er novembre 2026 (plan de révision 20262542, GB/T 39404-2026) single source, primary. Jusqu'à l'arrivée des nouvelles GB 11291 (2028–2029 estimate), concevoir selon ISO 10218:2025 et conserver les dossiers de conformité GB 11291.

Taïwan : les usines privées peuvent utiliser presque tout ; c'est l'identité du client qui décide. Les API frontières, les poids ouverts et l'équipement continental sont tous légaux (data.gov.tw). Mais depuis l'entrée en vigueur de la loi modifiée sur la gestion de la cybersécurité le **1er décembre 2025**, les autorités peuvent agir directement sur les produits qui sont des « marques continentales » et les restreindre dans les organismes publics, les entreprises d'État et les fournisseurs d'infrastructures critiques (base de données des lois et règlements, règles d'examen) verified ; la liste des fournisseurs d'infrastructures critiques est confidentielle ; et TSMC a inscrit SEMI E187 dans ses contrats d'achat (CNA) et retirerait l'équipement continental de ses fonderies 2 nm (TrendForce) single source. Les EMS privés, l'assemblage et les petites usines PCBA sont ouverts ; les fonderies, les services publics de parcs et les clients liés au gouvernement sont fermés en pratique. Un fournisseur contrôlé par le continent devrait concevoir

selon E187, livrer une nomenclature logicielle (SBOM), proposer des mises à jour hors ligne et ne jamais se connecter à distance depuis le continent.

Asie du Sud-Est : le moins de restrictions ; les problèmes sont le retour sur investissement lent et les maisons mères chinoises. Tout est utilisable, mais la directive américaine du 31 mai 2026 exige une licence pour exporter des articles informatiques avancés vers des entités dont le siège est en Chine ou détenues en dernier ressort depuis la Chine, même lorsque le destinataire est dans un pays tiers (Greenberg Traurig) single source, et la Malaisie a ses propres permis pour les puces avancées (Fulcrum). La demande se déplace ici : TrendForce s'attendait à ce que la Thaïlande, le Vietnam et la Malaisie approchent 50 % de la capacité SMT des ODM de serveurs en 2026 (TrendForce) single source, a 2023 forecast.

États-Unis : toute la pile, au prix des droits de douane, de la certification et des données contrôlées. Au chapitre 99 du tarif américain, un bras robotisé fabriqué sur le continent paie environ 40 % de droits cumulés, les instruments de type AOI et les caméras environ 37,5 % et les ordinateurs edge 25 % ; fabriqué en Malaisie ou en Indonésie environ 10–12,5 %, au Vietnam, en Thaïlande ou aux Philippines 12,5–15 %, à Taïwan 10 %, les ordinateurs à 0 % hors continent (USITC HTS chapitre 99, Great Lakes Customs Law) primary schedule + estimate. L'avis de l'USTR du 10 novembre a prolongé 178 anciennes exclusions de la Section 301 jusqu'au 9 janvier 2027, aucune ne couvrant les robots, les instruments d'inspection, les caméras ou les ordinateurs (Federal Register 2026-22925) verified.

L'OSHA exige que l'équipement électrique du lieu de travail soit « approuvé » (en pratique une marque NRTL, ou une évaluation sur site d'environ 3 000 à 20 000 $+ par cellule), et depuis le 11 août 2025, UL 1740 soumet les cellules robotisées entières à ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). L'électronique de défense est fermée en pratique aux fournisseurs contrôlés par le continent : l'ITAR traite les photographies de production comme des données techniques et applique une politique de refus à la Chine (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Une cellule destinée aux États-Unis devrait utiliser un bras fixe, des caméras sans restriction et des modèles locaux ou en cloud américain, avec un assemblage final en Asie du Sud-Est là où cela change réellement l'origine.

À tout budget, le plafond diffère selon la région mais n'atteint nulle part zéro personne estimate. Le continent peut empiler SPI 3D, AOI 3D, rayons X, revue IA et une équipe IA interne pour ramener la main-d'œuvre de revue à 8–15 %, mais sans modèles frontières fermés ni DGX Station, les questions-réponses LLM à l'échelle de l'usine doivent tourner sur des poids ouverts locaux et du calcul national (un modèle pleine taille pour de nombreux utilisateurs exige au moins deux serveurs Ascend à huit cartes). Taïwan et l'Asie du Sud-Est peuvent techniquement utiliser le meilleur de tout, y compris les API frontières et DGX Station, limités par l'identité du client, les marques continentales et les licences des maisons mères chinoises. Les États-Unis n'ont aucune limite technique, au prix des droits de douane, de la certification NRTL et de la séparation du personnel pour les données contrôlées. Partout, des personnes restent pour les cas limites, les nouveaux produits et changements de série, les nouveaux types de défauts et la validation client.

Un produit vendu dans plusieurs régions a donc besoin de deux piles de modèles : une pile continentale sur modèles nationaux déposés ou poids ouverts locaux, et une pile étrangère sur API frontières plus un modèle local d'origine non chinoise. La politique d'Anthropic effective le 12 novembre 2026 juge par propriété et contrôle, de sorte qu'un fournisseur contrôlé par le continent ne peut pas bâtir sur Claude même via une filiale à l'étranger, une plateforme cloud ou un revendeur (Anthropic) single source, primary.

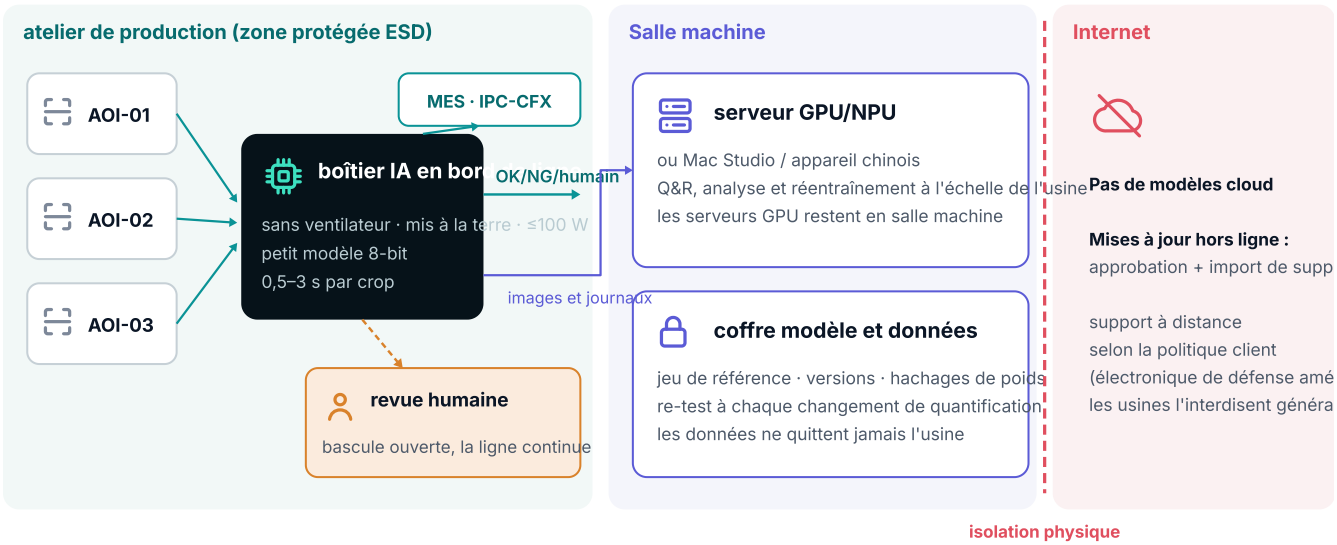
Sur site

02 Faire tourner l'IA dans l'usine : salles blanches, coupures d'air, matériel et quantification

Dans une usine isolée, la contrainte est le processus, pas le calcul

L'inspection visuelle entièrement hors ligne est courante ; les contraintes sont procédurales. Le guide 2024 du MIIT chinois sur la sécurité des systèmes de contrôle industriel demande un zonage, des pare-feu ou des diodes de données, une maintenance à distance limitée dans le temps et auditée, des supports scannés et des correctifs testés (texte du guide) single source ; les clients semi-conducteurs ajoutent SEMI E187 et E188, que TSMC a inscrits dans ses contrats d'achat en 2023 (SEMI, TXOne) verified. Les résultats reviennent au MES via IPC-CFX et l'identité de la carte circule via IPC-HERMES-9852 (norme Hermes). Le piège courant est la licence qui suppose un réseau : le mode hors ligne de Roboflow met en cache les poids sur un bail de 30 jours (Roboflow). Un produit hors ligne a besoin d'une activation hors ligne, de mises à jour signées sur supports scannés, d'une nomenclature logicielle, d'un contrôle USB, de journaux d'audit locaux et d'un adaptateur CFX.

FIG 05 **Un déploiement isolé (air gap)** ***Un petit boîtier en ligne décide ; la salle machine répond aux questions ; les pannes basculent vers les humains et les données ne quittent jamais l'usine.***



Source : Synthèse de l'auteur (chapitre 2).

Les modèles frontières fermés n'entrent pas, en pratique, dans la salle des serveurs d'une usine. Gemini n'est proposé isolé que via Google Distributed Cloud pour les gouvernements et les grandes entreprises (Google Cloud) ; Foundry Local de Microsoft n'ouvre les grands modèles qu'aux « clients qualifiés » et ne nomme aucun modèle GPT (Microsoft) ; Claude n'est proposé que dans des régions cloud gouvernementales, sans déploiement sur site (Claude for government) ; aucun des trois n'est proposé en Chine continentale. L'équipement d'usine doit être conçu autour des poids ouverts ; les modèles frontières peuvent au plus étiqueter ou expliquer des images au siège, hors ligne, lorsque le client autorise la sortie d'images anonymisées.

Le matériel de salle des serveurs doit être traité comme du matériel de bureau : le Mac Studio est homologué pour 10–35 °C et le DGX Spark pour 5–30 °C, avec des signalements d'arrêts thermiques en inférence soutenue (Apple,

forum NVIDIA) verified. La ligne reçoit des boîtiers industriels sans ventilateur, comme un PC industriel Jetson T4000 homologué pour −25 à 60 °C (AAEON) vendor claim. Comme une heure d'arrêt coûte à une grande usine de 36 000 à 2,3 millions de dollars (Siemens True Cost of Downtime 2024) vendor survey, chaque étape d'IA sur la ligne doit **échouer en mode ouvert** : en cas de panne, elle revient au processus AOI-plus-personnes existant.

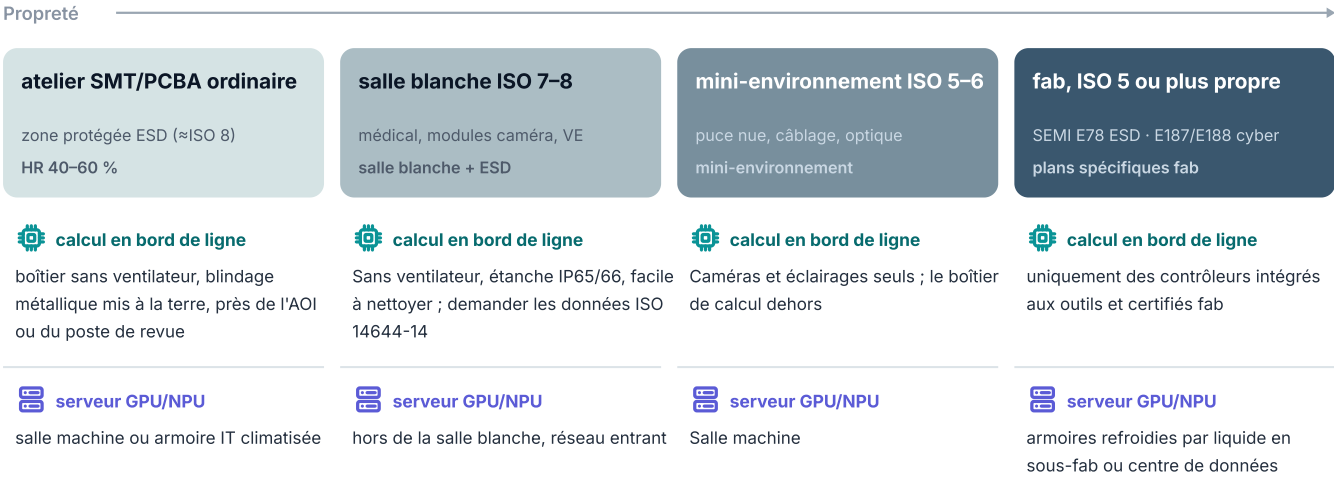
Zones propres et ESD : un atelier SMT ordinaire est une zone protégée ESD, pas une salle blanche classée

Ce qui limite l'emplacement du calcul, c'est l'électricité statique et le flux d'air, pas la classe de propreté. La plupart des zones d'assemblage SMT ne sont « pas formellement contrôlées en particules », et les usines de classe mondiale « peuvent atteindre ou approcher ISO 8 » (panel d'experts Circuitnet) single source. Les PCBA médicales et les modules caméra se situent généralement en ISO 7, la puce nue et l'alignement optique en ISO 5–6 local (PCBSync) single source ; les constructeurs de salles blanches annoncent ISO 7–8 pour le SMT en général (Angstrom Technology) vendor claim.

L'exigence universelle est le contrôle ESD : ANSI/ESD S20.20-2021 exige des conducteurs isolés sous 35 V et maintient les isolants à champ fort à au moins 30 cm des dispositifs sensibles (ANSI, ESDA, Desco) verified. Les fonderies placent le calcul dans la sous-fonderie ou le centre de données, dans des baies refroidies par liquide conformes à SEMI S2/S8/F47 (Prodrive) vendor claim.

L'entrée d'un équipement en zone propre est testée pour l'émission de particules selon ISO 14644-14 ; les robots détiennent de tels certificats (Fraunhofer IPA), mais je n'ai trouvé aucun PC industriel, boîtier edge AI ou serveur GPU qui en possède un. CE, FCC et IEC 60068 montrent qu'un boîtier tolère son environnement, pas qu'il peut entrer dans une zone protégée ESD ; le coordinateur ESD du client le vérifie. Les serveurs GPU ne vont jamais sur la ligne : un DGX Station a tiré 1 292 W au GPU en test (StorageReview).

FIG 06 **Les zones propres et ESD décident où le calcul peut siéger** ***Un atelier SMT ordinaire est une zone protégée ESD, pas une salle blanche. Les serveurs GPU et NPU restent en salle machine.***



Source : ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, normes SEMI et supports fournisseurs ; compilé par l'auteur.

Zone	Environnement typique	Calcul sur la ligne	Serveurs GPU/NPU
Atelier SMT/PCBA ordinaire (grand public, industriel)	Zone protégée ESD selon S20.20 ou IEC 61340-5-1, généralement pas une salle blanche classée (équivalent ≈ ISO 8) ; 40–60 % HR	Boîtier sans ventilateur dans un châssis métallique mis à la terre (Jetson T4000/Thor, boîtier NPU chinois) à côté de l'AOI ou du poste de revue ; moniteurs et habillages plastiques hors de la zone ou ≥ 30 cm des cartes	Salle des serveurs ou baie informatique climatisée

Zone	Environnement typique	Calcul sur la ligne	Serveurs GPU/NPU
Médical, modules caméra, électronique EV (salle blanche ISO 7–8)	Salle blanche plus protection ESD	Sans ventilateur, étanche IP65/66, lisse et facile à essuyer ; demander les données ISO 14644-14 ou un certificat IPA	Hors de la salle blanche, connecté en réseau
Puce nue, liaison filaire, postes optiques (ISO 5–6 local)	Mini-environnement	Uniquement caméras et lumières à l'intérieur ; boîtier de calcul à l'extérieur	Salle des serveurs
Salle blanche de fonderie (ISO 5 et plus propre)	Plans ESD et d'attraction électrostatique spécifiques à la fonderie, équipement conforme à SEMI E78 ; cybersécurité selon SEMI E187/E188	Uniquement des contrôleurs intégrés à l'outil et qualifiés fonderie ; pas de PC complémentaires avec ventilateurs	Baies refroidies par liquide en sous-fonderie ou centre de données de la fonderie

Le tableau est ma synthèse des sources ci-dessus estimate. Pour un produit, cela signifie un boîtier côté ligne sans ventilateur sous environ 100 W, un châssis métallique relié à la terre de protection, une entrée 12–24 V CC, un fonctionnement à large température au flux d'air réel de l'armoire, et un rapport d'émission de particules ISO 14644-14 prêt pour les clients en ISO 7 et plus propre. Toute conception où « un serveur GPU sert quatre lignes » place le serveur en salle des serveurs et ne laisse qu'un petit boîtier ou une caméra sur la ligne.

Limites de vitesse par usage : la précision est plus difficile que la vitesse

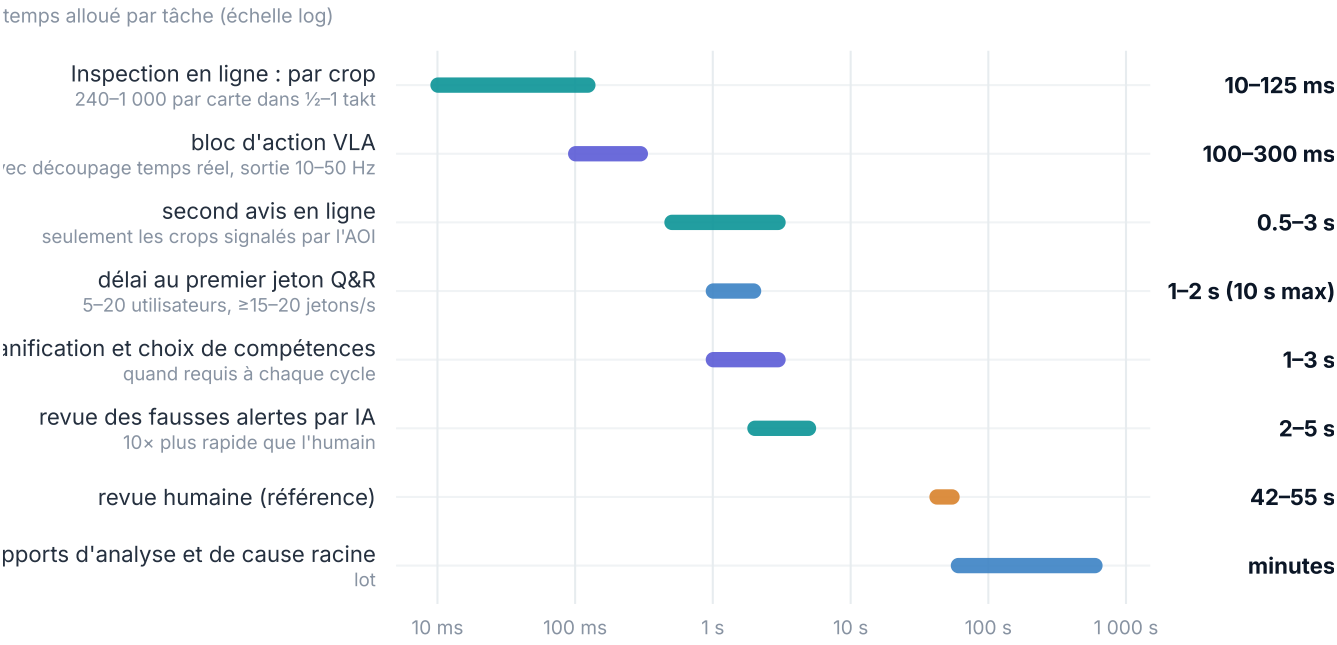
Les usines et les vendeurs ne publient presque jamais de cibles de latence ou de disponibilité de l'IA. Les seuils ci-dessous dérivent du takt, du temps de revue humaine, des cadences de contrôle robotique et de la recherche sur l'interaction estimate :

Usage	Limite de vitesse	Limite de précision	Disponibilité
Inspection en ligne de chaque pièce	Dans ½–1 takt : 240–1 000 recadrages dans un takt de 10–30 s, 8–100 recadrages par seconde	Échappements pas pires que « AOI plus personnes » d'aujourd'hui (exemple contractuel < 0,5 %)	Échec en mode ouvert, ne jamais arrêter la ligne ; cible ≥ 99,9 % (pas de SLA public)
Second avis en ligne sur les seuls recadrages signalés par l'AOI	Signaux (généralement ≤ 5–20 par carte) dans ½ takt : 0,5–3 s par recadrage	Idem ; entraîné sur les images propres de la ligne	Revient à la revue humaine
Revue IA des fausses alertes	2–5 s par appel, plus de dix fois plus rapide que les 42–55 s d'une personne	Échappements pas pires qu'un réviseur humain	La dégradation ajoute des personnes, n'arrête jamais la ligne
Contrôle robotique (VLA)	100–300 ms par bloc d'action avec segmentation en temps réel, sortie 10–50 Hz	≥ 99 % de réussite de tâche pour battre le mouvement programmé	Système de sécurité indépendant du modèle
Planification robotique et choix de compétence	1–3 s au besoin à chaque cycle ; ≤ 1 min si uniquement sur exceptions ou changements de série	La sortie doit être du code vérifiable ou un identifiant de compétence	Revient au programme fixe
Questions-réponses (SOP, requêtes MES)	Premier jeton idéalement ≤ 1–2 s, au plus 10 s ; ≥ 15–20 jetons/s par utilisateur ; 5–20 utilisateurs simultanés	Citations exactes ; fidèle sur un long contexte	Niveau IT de bureau (≈ 99–99,5 %)

Usage	Limite de vitesse	Limite de précision	Disponibilité
Analyse et rapports de cause racine	Les minutes conviennent (lot)	Le plus haut ; fidèle sur un long contexte	Niveau IT de bureau

Physical Intelligence a mesuré 97 ms d'inférence pour un VLA de classe 3B (environ 108–139 ms avec réseau et prétraitement), constaté que la segmentation en temps réel maintient le débit inchangé jusqu'à 200 ms de délai supplémentaire, et réalisé des tâches fines comme allumer une allumette ou brancher un câble réseau à plus de 300 ms (Physical Intelligence) vendor research ; la recherche sur l'interaction fixe la limite pour garder le fil de la pensée à une seconde et pour l'attention à dix (NN/g). La précision est la partie difficile : le Qwen3-VL-8B ouvert atteint en moyenne seulement 56,8 en zero-shot sur cinq jeux de données d'anomalies industrielles (IndusAgent) single source. Aucun matériel ne sauve un modèle vision-langage qui n'a pas été ajusté sur les images propres de la ligne.

FIG 07 **Limites de vitesse par usage** ***L'inspection en ligne se mesure en millisecondes, la Q&R en secondes, l'analyse en minutes. Les humains prennent 42–55 s par appel.***



Source : Synthèse de l'auteur (chapitres 1–2).

Matériel : Thor pour le poste, un Mac Studio ou un appareil national pour les questions-réponses de l'usine, et pas de DGX Station sur le continent

Appareil	Mémoire et bande passante	Prix (novembre 2026)	À quoi il sert	Chine continentale
Kit de développement Jetson AGX Thor / T5000 / T4000	128 Go LPDDR5X, 273 Go/s ; 40–130 W	3 499 $ au lancement → 5 499 $ (juillet 2026) ; T5000 4 999 $, T4000 2 999 $; kit de développement RS China 59 104 RMB taxes comprises	Modèles de vision 2–8B re-vérifiant les recadrages ; VLA de classe 3B en temps réel	Oui (5A992.c, pas de licence pour les usines civiles)
DGX Spark	128 Go, 273 Go/s	3 999 $ → 4 699 $ → 6 950 $; le modèle Chine ASUS GX10 à même puce 32 999 RMB sur JD (oct. 2025)	MoE 8–35B multi-utilisateurs	Oui (5A992.c, modèle Chine, canaux autorisés)

Appareil	Mémoire et bande passante	Prix (novembre 2026)	À quoi il sert	Chine continentale
RTX PRO 6000	96 Go GDDR7	≈ 8 565 $ → 13 250 $ (juin 2026) → ≈ 16 005 $ en moyenne eBay en novembre	27–35B FP8 multi-utilisateurs ; MoE de classe 120B en 4 bits	Non (4A090.a)
Mac Studio M5 Ultra 512 Go	512 Go unifiés, 1,2 To/s ; 480 W max soutenus	À partir de 22 299 $ aux États-Unis ; ≈ 173–186 k RMB estimé en Chine	DeepSeek 671B en 4 bits pour 1 à 3 utilisateurs	Oui
DGX Station GB300	748 Go (252 Go HBM3e à 7,1 To/s + 496 Go LPDDR5X à 396 Go/s) ; 1 292 W mesurés au GPU	123–140 k $ prix public en novembre (≈ 830–940 k RMB avant droits) ; 95–117 k $ en août	Forte concurrence sur les MoE 120–235B qui tiennent en HBM ; seulement quelques utilisateurs sur les modèles 433–743B	Non (5A992.z, TPP 58 484)
Huawei Atlas 300I Duo	96 Go LPDDR4X, 408 Go/s	13 300 RMB chez un marchand JD	Revue par lots à faible volume ; peu de modèles pris en charge	Oui (national)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 Go)	8 × 64 Go HBM	Non obtenu ; DeepSeek complet en exige deux en W8A8, quatre en BF16	Questions-réponses et analyse multi-utilisateurs à l'échelle de l'usine	Oui (national)
RTX 5090 D V2 (modèle Chine)	24 Go	16 499 RMB au lancement (août 2025), en hausse depuis avec les prix mémoire	Modèles de vision 8B, petits VLA	Oui

Sources : (NVIDIA China Thor, lancement NVIDIA Thor, CNX sur les hausses de prix, données de classification d'exportation NVIDIA, RS China, Implicator, IThome sur ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, spécifications Apple, TechRadar, IThome sur les prix Chine du M5 Ultra, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, matrice de support vLLM-Ascend, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). Les prix continentaux du Mac Studio 512 Go sont extrapolés des primes de configuration américaines et sont de faible confiance estimate.

FIG 08 **Calcul sur site : mémoire, prix et ce qui est légalement achetable en Chine continentale** ***Les classifications à l'export décident de la liste restreinte avant même les benchmarks.***



Source : Classifications d'exportation NVIDIA, Apple, RS China, JD.com et prix de revue (novembre 2026).

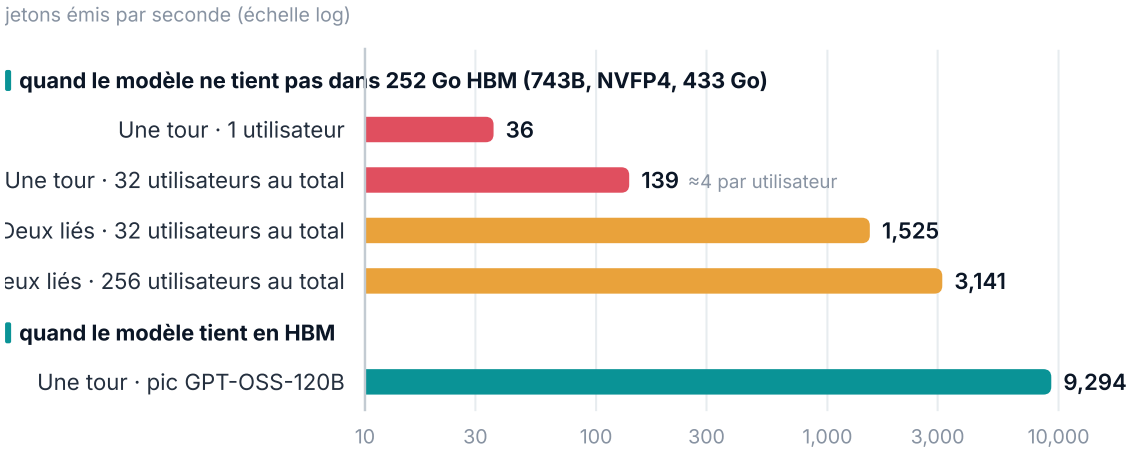
Jetson Thor rend le déploiement au niveau du poste réalisable. Les tests TensorRT Edge-LLM de NVIDIA montrent Qwen3-VL-8B (NVFP4) préremplissant 2 048 jetons en 104,5 ms et décodant 44,4 jetons/s sur Thor (TensorRT

Edge-LLM) vendor claim ; pour un recadrage AOI et une réponse de 20 jetons, cela fait environ 0,5 s avec le modèle 8B et 0,2 s avec le 2B, soit 8,5–21 recadrages par seconde en lot estimate. C'est la longueur de sortie, pas le nombre de paramètres, qui fixe la latence, donc le décodage contraint vers un jeton de classe ou un court JSON fait partie de la conception. GR00T N1.7 (3B) tourne à 25,1 Hz sur Thor (Jetson AI Lab) vendor claim. Un Orin NX prend environ 1,8 s par recadrage et un Atlas 300I Duo a un p95 au-dessus de 23 s, apte seulement au travail par lots (CSDN) single source. Sur le continent, budgétez les produits NVIDIA à 40–60 % au-dessus du tarif américain estimate.

Un Mac Studio 512 Go rend un modèle pleine taille possible sur une seule machine, pour quelques utilisateurs. Le M5 Ultra (août 2026) offre jusqu'à 512 Go à 1,2 To/s à partir d'environ 22 299 $ (Apple, TechRadar) ; DeepSeek 671B en 4 bits (≈ 380–404 Go) décode 17–20 jetons/s sur un M3 Ultra (Pinggy) verified et environ 25–29 estimé sur le M5 Ultra. Les longs prompts sont le point faible : un prompt de récupération de 10 000 jetons prend environ 17–70 s estimate, au-delà de la limite d'interaction de 10 secondes.

DGX Station est la machine individuelle la plus puissante, mais indisponible sur le continent et cinq à six fois le prix d'un Mac. Le GB300 livré a **748 Go**, dont seulement 252 Go de HBM rapide (Wccftech, Tom's Hardware, août 2026) verified. Les tests indépendants montrent environ 9 300 jetons de sortie/s sur GPT-OSS-120B, qui tient en HBM, mais un modèle 743B qui déborde en LPDDR5X n'a donné à 32 utilisateurs que 139 jetons/s au total, environ 4 chacun ; deux unités liées ont atteint 1 525 (StorageReview, août 2026, StorageReview, oct. 2026) single source. Les prix publics ont grimpé à 123–140 k $ en novembre (CyberPowerPC, monitors.com). Sur le continent, les options réalistes pour le même travail sont le Mac Studio, les serveurs CPU+GPU de type KTransformers (KTransformers) vendor test ou les serveurs Ascend nationaux.

FIG 09 **Une grosse station de travail ne fait pas un assistant partagé** ***Quand le modèle ne tient pas en mémoire à large bande, 32 utilisateurs simultanés n'obtiennent qu'environ 4 jetons par seconde chacun.***



Source : StorageReview (août et oct. 2026) ; ServeTheHome. Tests indépendants d'une seule source.

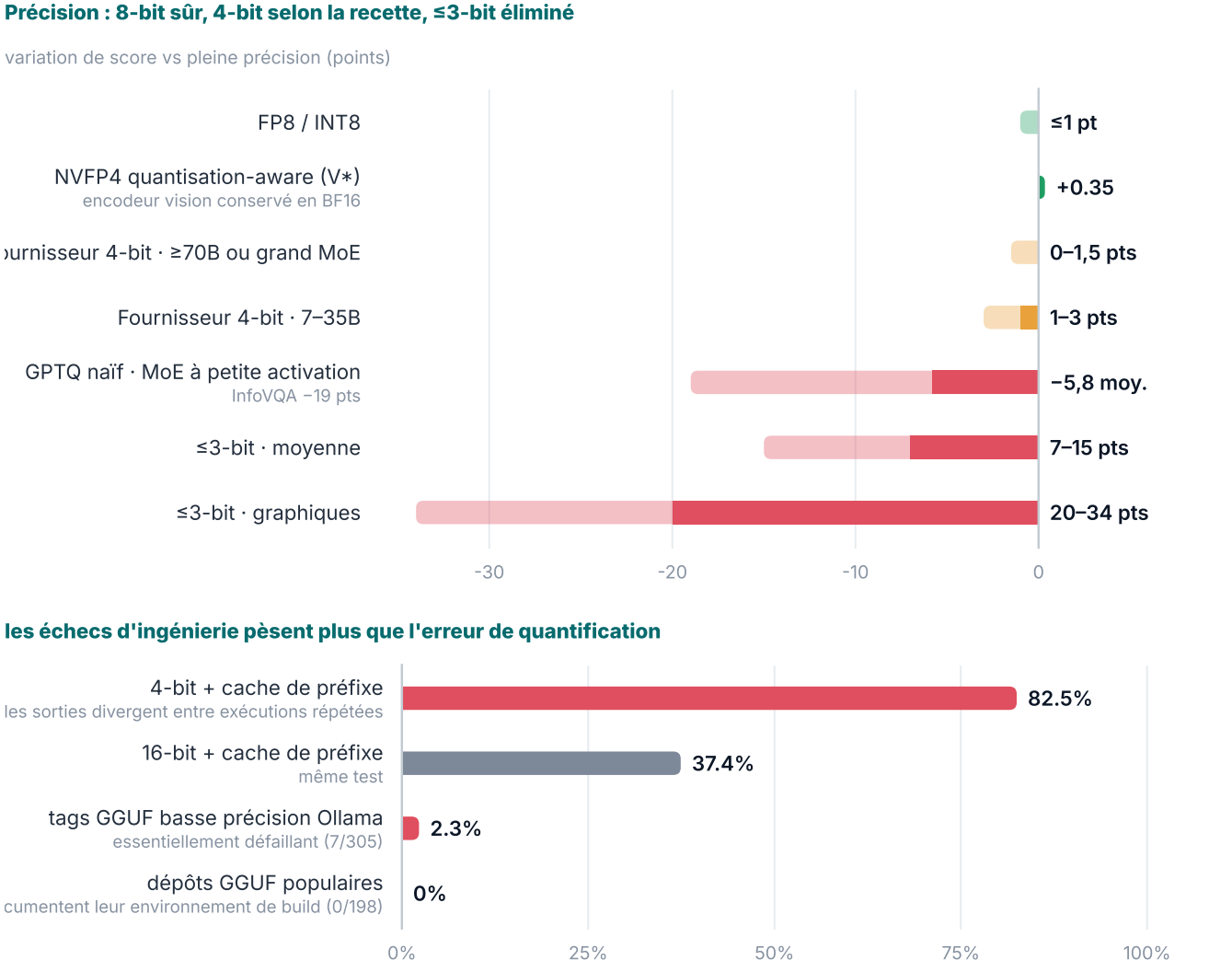
Les pénuries de mémoire ont déplacé les mêmes pièces de 20 à 100 % en un an (le DGX Spark est passé de 3 999 $ à 6 950 $ (TechRadar)), donc les devis d'équipement devraient être indexés plutôt que fixes.

Les accélérateurs nationaux font tourner des modèles pleine taille ; la maintenance logicielle est le goulot d'étranglement. L'Ascend 910B a 64 Go de HBM, mais vLLM-Ascend marque DeepSeek V3/R1 comme « plus maintenu » et ne prend en charge que quelques modèles sur des cartes plus petites (vLLM-Ascend) single source, primary, et chaque nouveau modèle prend un à deux mois à adapter (Caijing/Anue) ; livrez une version de modèle figée et validée. Servir DeepSeek 671B complet à 20–100 utilisateurs exige au moins deux serveurs Atlas 800I A2 à huit cartes (China Telecom Cloud) verified, pourtant les appels d'offres publics d'appareils LLM en Chine n'ont totalisé que 27,5 millions de RMB en janvier–septembre 2025 (C114) single source : les usines paient l'intégration et le service, pas une boîte de calcul.

Quantification : le 8 bits est sûr, le 4 bits dépend de la recette, le 3 bits et en dessous est exclu

Le 8 bits (FP8 ou INT8) est essentiellement sans perte pour les modèles vision-langage (Red Hat, Hugging Face) verified. En 4 bits, la perte dépend de la taille et de la recette : la version AWQ 72B de Qwen perd moins d'un point sur MMMU tandis que la 7B perd 2,8 (Qwen AWQ) verified, et les tâches riches en détails chutent davantage (le GPTQ 4 bits naïf coûte à Qwen3-VL-30B-A3B 19 points sur InfoVQA (VEQ)). À 3 bits et en dessous, la perte moyenne est de 7 à 15 points et les graphiques perdent 20 à 34 (VLMQ) verified ; au-delà de 64 k jetons, le 4 bits perd jusqu'à 59 %, pire sur les entrées non anglaises (arXiv 2505.20276). Gardez l'encodeur de vision en FP16/BF16.

FIG 10 **Quantification : 8-bit est sûr, 4-bit dépend de la recette, 3-bit et moins est éliminé** ***Les erreurs d'ingénierie dans la pile de service coûtent plus de précision que la quantification elle-même.***



Source : Fiche modèle NVIDIA Nemotron, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 et autres.

Deux avertissements comptent davantage. Premièrement, **au 11 novembre 2026, aucune étude ni fiche de modèle n'a mesuré la quantification sur des images de défauts PCB/SMT**, donc revalidez sur le jeu de référence. Deuxièmement, les erreurs de pile de service sont souvent plus grandes que l'erreur de quantification : le même modèle a obtenu environ 30 points de moins sous Ollama que sous l'implémentation de référence (SiliconBench) ; 7 des 305 balises officielles de modèles de code d'Ollama étaient essentiellement cassées bien que les fichiers soient intacts (arXiv 2609.05881) ; aucun des 198 dépôts GGUF populaires n'a enregistré sa compilation (arXiv 2610.19498) ; et avec la réutilisation du cache de préfixe, 82,5 % des épisodes ont divergé en 4 bits contre 37,4 % en 16 bits (arXiv

2609.04748) single sources. Le 4 bits quantifié par le vendeur est quasi sans perte sur les tâches de détail (NVIDIA Nemotron 3.5) vendor claim, mais les VLA en 4 bits perdent 1,5–4,3 points sous perturbation et tournent plus lentement sur Thor (LoopQ) single source. Gérez un modèle quantifié comme un nouvel instrument de mesure : épinglez les hachages de poids, enregistrez la compilation, réinitialisez le cache de préfixe et revalidez après tout changement de précision, de moteur ou de gabarit.

Précision	≥ 70B dense ou ≥ 100B MoE	7–35B	≤ 4B ou tâches riches en détails	Verdict
FP8 / INT8	Sûr (≤ 1 point)	Sûr	Sûr ; le défaut pour l'inspection	Utiliser
4 bits AWQ/GPTQ (recette vendeur, encodeur vision FP16)	0–1,5 point	1–3 points ; GPTQ naïf sur petit MoE actif −5,8 en moyenne, −19 sur InfoVQA	2–3 points ; valider d'abord les tâches de détail	Conditionnel
NVFP4 (recette vendeur)	≈ 0,3–1 point sous FP8 (données vendeur)	Compilations quantifiées quasi sans perte sur V*, OCRBench	Pas de données publiques	Conditionnel (préférer les compilations quantifiées)
GGUF Q4_K_M	Non testé	Forte variance sur les tâches de domaine	Pas pour les petits textes ou défauts	Pas pour le jugement de détail
GGUF Q5_K_M et au-dessus	Présumé sûr	Quasi sans perte dans deux tests	Le plancher pour le détail sur llama.cpp	Utiliser
≤ 3 bits (y compris les modèles 1T compressés à 1,8–2 bits sur une machine)	Pas pour les décisions	Non	Non	Non

Quelle configuration pour quel travail

Configuration	Inspection en ligne	Revue IA des fausses alertes	Planification et contrôle robotique	Questions-réponses et analyse
Détecteurs dédiés ou modèles d'anomalie (classe YOLO, RF-DETR, PatchCore), INT8/FP16, tout boîtier edge NPU ou GPU	● Oui : le juge principal, 2–12 ms par recadrage	● Oui (classifieur ajusté sur la ligne)	● Perception seulement	—
Modèles de vision 2–4B, NVFP4/AWQ/FP8 avec encodeur vision FP16, sur Thor/T4000 ou DGX Spark	● Conditionnel : re-vérifier les recadrages signalés par l'AOI à 0,2–0,3 s chacun	● Conditionnel, sous 1 s après ajustement	● Squelette VLA (VLA 2–3B à ≈ 40–130 ms par bloc)	● Faible
Modèles de vision 7–9B (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 ou 4 bits vendeur, sur Thor, Spark ou une carte 24–32 Go	● Conditionnel : ≈ 0,5 s par recadrage signalé	● Oui : 0,5–1 s par appel, meilleur rapport qualité-prix	● Planification et choix de compétence (1–3 s), pas dans la boucle de contrôle	● Questions-réponses SOP de base pour un ou quelques utilisateurs
MoE 26–35B (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● Non (trop lourd)	● Oui, comme explicateur ou pour les cas difficiles	● Planification	● Oui (8 utilisateurs sur un Spark, ≈ 0,1 s premier jeton)
MoE 106–241B en 4 bits sur un Mac Studio, RTX PRO 6000 ou Spark	● Non	● Non	● Planification hors ligne	● Oui, pour un ou quelques utilisateurs

Configuration	Inspection en ligne	Revue IA des fausses alertes	Planification et contrôle robotique	Questions-réponses et analyse
DeepSeek 671B ou Kimi K2.x 1T en 4 bits sur un Mac 512 Go, serveur KTransformers ou DGX Station	Non	Non	Génération de programmes hors ligne uniquement	Conditionnel : 1–3 utilisateurs et prompts courts ; les longs prompts prennent 16–170 s
Tout en ≤ 3 bits ; GGUF Q4_K_M pour le détail sur petits modèles ; Ollama pour de nombreux utilisateurs	Non	Non	Non	Non

Sources : les données TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research et KTransformers ci-dessus, et (tests DGX Spark, tests RTX PRO 6000, Hardware Corner sur Atlas, Ultralytics YOLO26, RF-DETR) **community and vendor tests, mostly single source** .

Donc, aux deux inquiétudes habituelles : **il n’est pas obligé d’être lent** si le modèle correspond au travail (les modèles de vision petits et moyens sur Thor ou Spark respectent chaque budget en ligne et de revue ; les modèles 671B–1T restent hors ligne), et **la quantification n’est pas obligée de le rendre faux** avec FP8/INT8 par défaut, ou un 4 bits de recette vendeur au-dessus de 7B validé sur le jeu de référence. Les risques plus grands sont la précision industrielle zero-shot d’environ 57 points des modèles généraux et les erreurs de pile de service. Donnez aux demandes d’inspection des tailles de lot fixes et leur propre file (la température 0 n’est pas reproductible bit à bit en lot (documentation vLLM)), gardez un secours chaud N+1, et faites échouer en mode ouvert chaque étape de la ligne.

Enseignant et élève

03 Enseignant et élève : les modèles frontières peuvent pré-étiqueter les données d'inspection ; pour les bras robotisés, les vrais enseignants sont encore des personnes

Candidats enseignants de novembre 2026 : la vision générique a doublé, les scores de défauts industriels manquent

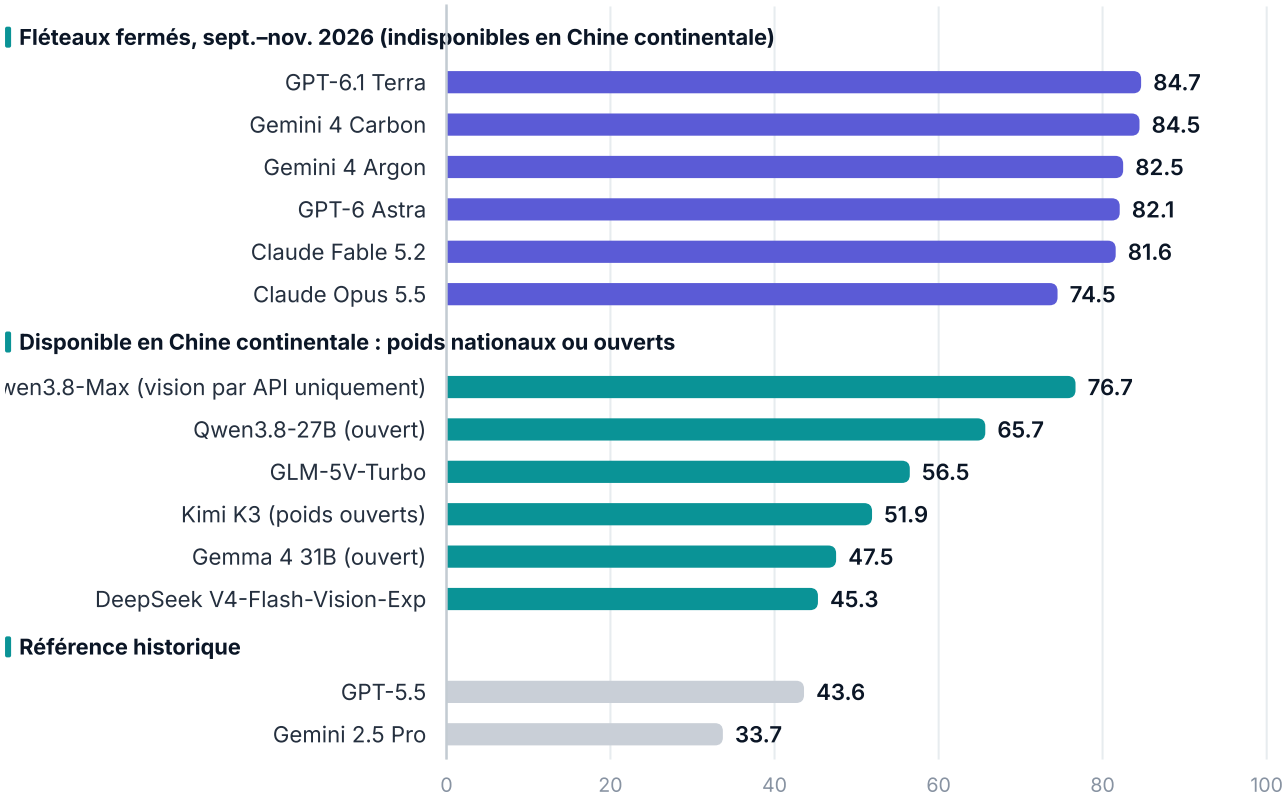
Chaque modèle phare fermé qui pourrait servir d'enseignant en novembre 2026 accepte des images : Claude Fable 5.2 (disponible le 9 novembre, 8 $/40 $ par million de jetons) et Opus 5.5 (4 $/20 $) (Anthropic, aperçu des modèles Claude) verified ; GPT-6 Astra (10 $/50 $) et GPT-6.1 Terra (1 $/5 $) (tarifs OpenAI) ; et Gemini 4 Argon et Carbon, ouverts par étapes (Google Carbon). Utilisables en Chine continentale : Qwen3.8-Max (vision via API) et des modèles ouverts comme Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp et Gemma 4 (blog Roboflow, Hugging Face).

L'écart clé doit être dit clairement : aucun de ces modèles ne publie de score sur MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB ou tout jeu de données PCB/SMT. MMAD n'a toujours pas de classement (MMAD GitHub), et les deux plus récents articles d'anomalies industrielles (août–septembre 2026) utilisent encore Gemini 2.5 Pro et Claude Sonnet 4 comme leurs plus récents niveaux de référence généraux (InspectorGPT, Anomaly-LR). Une recherche en texte intégral sur arXiv le 11 novembre n'a trouvé aucune occurrence de « Fable 5.2 » ou « Gemini 4 Carbon » dans ce contexte (API arXiv) single source, absence of evidence. Les substituts les plus proches :

- Le modèle frontière le plus récent avec un score MMAD est GPT-5.5 : 69,95 % sur « défaut ou non » et 76,20 % en moyenne sur sept tâches sur un sous-ensemble de 500 images, contre 95,24 % et 86,65 % pour des experts humains (AnomalyClaw) single source ; GPT-4o a obtenu 68,63 % et 74,92 % en 2024 (MMAD, ICLR 2025), soit environ 1,3 point de progrès en deux ans.
- Sur LDU-Bench, de vraies images de revue SEM de lithographie CI et le seul benchmark électronique avec des modèles 2026, le tri bon-ou-mauvais est correct (macro F1 0,932 pour GPT-5.4, 0,884 pour GLM-5V-Turbo, 0,876 pour Claude Opus 4.6, 0,504 pour Gemma 4 31B), mais la classification de 11 morphologies de défauts n'atteint que 0,20–0,41 (LDU-Bench) single source.
- Sur les évaluations de vision indépendantes de Roboflow (mises à jour le 10 novembre 2026), le mAP@50 de détection d'objets est de 84,5 pour Gemini 4 Carbon, 84,7 pour GPT-6.1 Terra et 81,6 pour Fable 5.2, contre 33,7, 43,6 et 38,6 pour Gemini 2.5 Pro, GPT-5.5 et Claude Opus 4.8 (Roboflow Vision Evals) single source, independent evaluator, private datasets.

FIG 11 **Candidats enseignants : les scores génériques ont doublé en une génération*Mais aucun des fléteaux 2026 n'a publié de score sur un banc de défauts industriels.***

détection générique Roboflow mAP@50



Bancs de défauts industriels (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB) : zéro score publié pour ces fléteaux 2026
les scores de détection génériques ont doublé en une génération, mais ce n'est pas un jugement de défaut ; testez sur votre propre jeu de référence

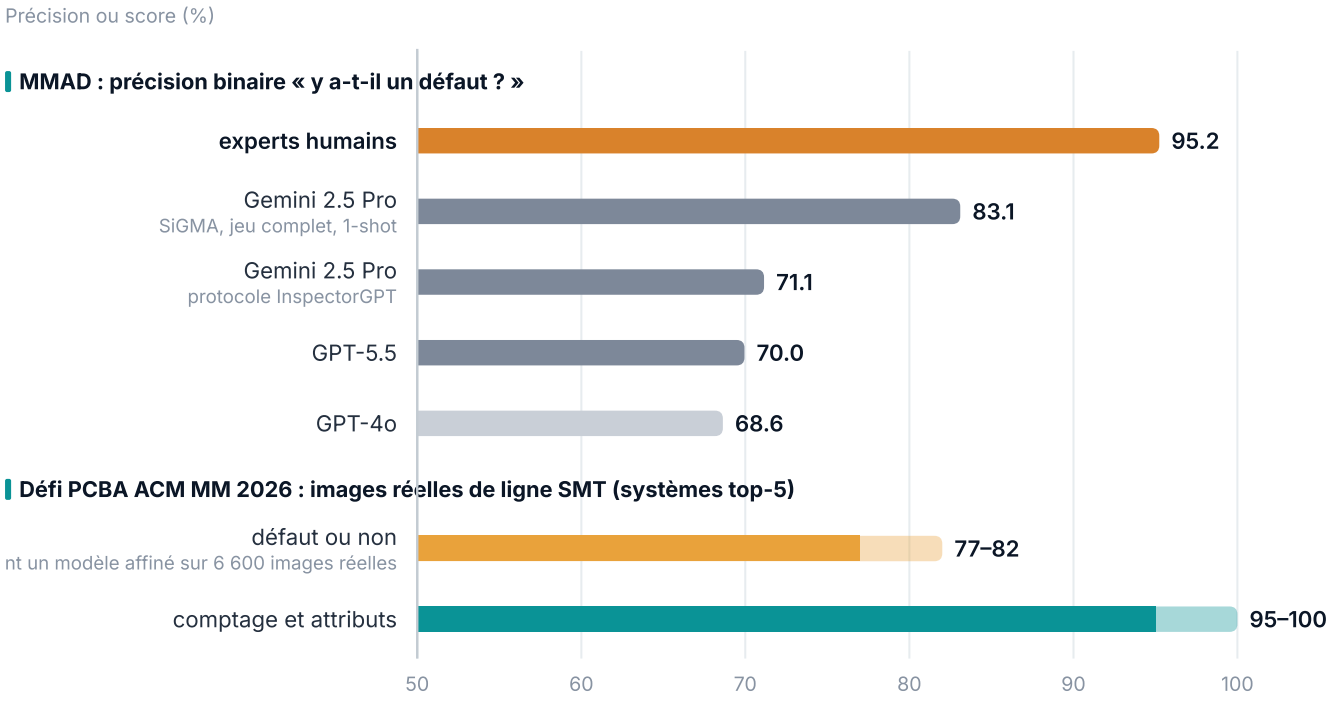
Source : Roboflow Vision Evals ; notes de version des modèles ; pages MMAD, MVTec AD 2 et autres bancs (au 11 nov. 2026).

Candidat enseignant (nov. 2026)	Prix par M de jetons (entrée/sortie)	Détection Roboflow mAP@50 / comptage	Benchmarks de défauts industriels	Utilisable en Chine continentale
Claude Fable 5.2	8 $ / 40 $	81,6 / 86,5	Aucun	Non
Claude Opus 5.5	4 $ / 20 $	74,5 / 80,6	Aucun	Non
GPT-6 Astra	10 $ / 50 $	82,1 / 80,2	Aucun	Non
GPT-6.1 Terra	1 $ / 5 $	84,7 / 80,2	Aucun	Non
Gemini 4 Carbon	2 $ / 10 $ (promotionnel)	84,5 / 90,1	Aucun	Non
Qwen3.8-Max (vision via API uniquement)	2 $ / 6 $	76,7 / 81,1	Aucun	Oui
Qwen3.8-27B (ouvert)	Local	65,7 / 64,9	Aucun	Oui
GLM-5V-Turbo	1,20 $ / 4,00 $	56,5 / 48,6	Tri LDU-Bench 0,884	Oui
Gemma 4 31B (ouvert)	Local	47,5 / 51,4	Tri LDU-Bench 0,504	Oui (poids locaux)
Référence : GPT-5.5 / GPT-4o	—	43,6 / 68,0 (GPT-5.5)	MMAD binaire 69,95 % / 68,63 %	Non

Candidat enseignant (nov. 2026)	Prix par M de jetons (entrée/sortie)	Détection Roboflow mAP@50 / comptage	Benchmarks de défauts industriels	Utilisable en Chine continentale
Experts humains	—	—	MMAD binaire 95,24 %	—

Les articles plus récents testent encore la génération précédente, mais pointent dans le même sens. Sur le jeu MMAD complet, les experts humains obtiennent 95,2 % sur défaut-ou-non et Gemini 2.5 Pro 83,1 % (SiGMA), pourtant le même modèle n'a obtenu que 71,14 % dans un autre article (InspectorGPT) : le protocole de test déplace le score autant qu'une génération de modèle. Dans le défi VAND 4.0 de CVPR 2026, les modèles vision-langage prêts à l'emploi ont plafonné à 61,88 AP contre 90,84 pour les méthodes supervisées, et les organisateurs ont conclu qu'ils « ne peuvent pas encore remplacer les classifieurs affinés » (VAND 4.0) single source. Dans le défi PCBA d'ACM MM 2026, les cinq premiers ont obtenu 76,92–82,47 sur défaut-ou-non tandis que le type de composant et le comptage de broches atteignaient 95–99,5 (arXiv 2609.21276) single source : reconnaître est bien plus facile que juger.

FIG 12 **« Y a-t-il un défaut ? » reste difficile pour les modèles généraux** ***Le meilleur modèle général accuse plus de 12 points de retard sur les experts humains sur MMAD ; les systèmes affinés font mieux sur des images de ligne réelles.***



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

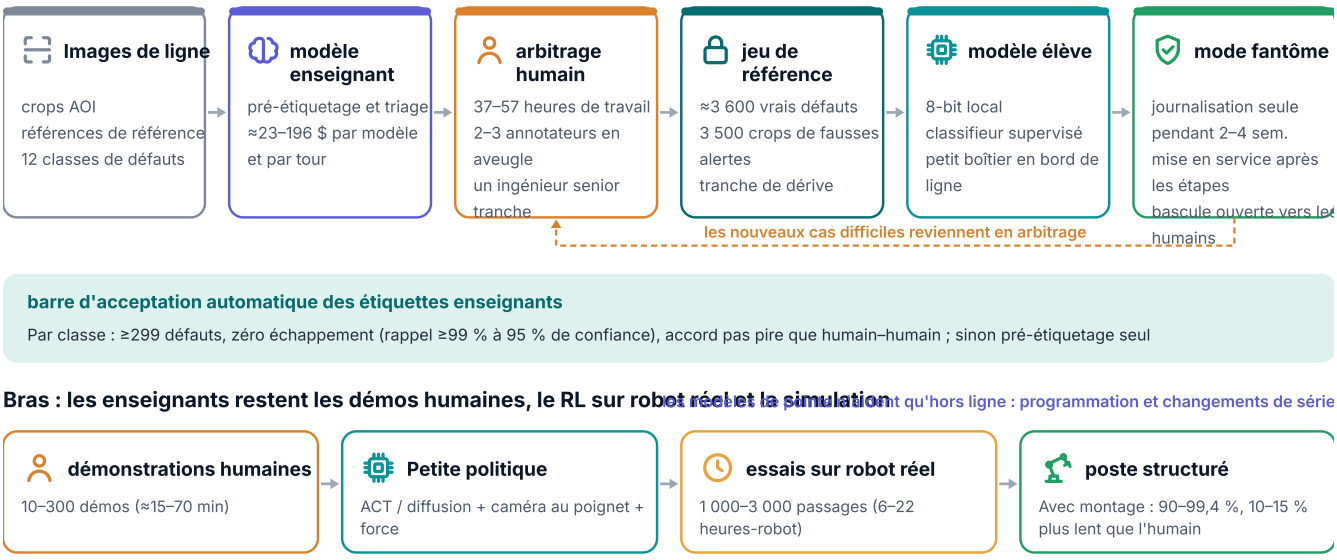
Trois conclusions s'ensuivent. Voir et compter ont doublé en une génération, mais juger l'anomalie avance lentement, dépend fortement du protocole et comporte des biais systématiques ; ce que gagnent les nouveaux modèles ne peut être mesuré que sur les images propres d'une usine. Dans leur distribution d'entraînement, les spécialistes 7B affinés battent encore les modèles généraux (Anomaly-LR, UniPCB). Et les données réelles de ligne effondrent les scores : sur des panneaux électroniques assemblés, l'AUROC au niveau image est tombé de 97–98 sur MVTec AD à 31,5–78,6 (arXiv 2606.01992) single source. Pour les entreprises continentales, l'enseignant le plus puissant disponible, Qwen3.8-Max, se situe environ 8 points sous la frontière mondiale.

Faire tourner des dizaines ou des centaines de tours : l'argent n'est pas le problème, la qualité des étiquettes l'est

L'argent n'est pas le problème. Aux tarifs de novembre 2026, 1 000 images d'un mégapixel coûtent environ 3 $ (GPT-6.1 Terra) à 32 $ (GPT-6 Astra) par tour, et cent tours sur 100 000 images environ 32 000–135 000 $ avant remises de volume estimate (docs vision Claude, guide vision OpenAI, Roboflow Vision Evals). Pour des étiquettes simples, une personne coûte environ autant que Terra ; l'avantage de l'enseignant est la vitesse et l'échelle, pas le prix. Les NDA des clients gardent généralement les images de produit dans l'usine, donc le jeu de référence est étiqueté sur site.

Le vrai obstacle est la qualité, et davantage de tours ne corrige pas les erreurs systématiques. Roboflow fait tourner chaque tâche trois fois et les scores ne bougent que de ±0,1–1,4 point (Roboflow), donc un modèle répète surtout ses erreurs et le vote majoritaire ne peut pas le sauver. Les tours n'aident que lorsque chacun ajoute de l'information : une ou deux bonnes images de référence ajoutent environ 3–4 points, plus n'ajoutent rien (AD-Copilot) ; faire argumenter un modèle contre une bonne référence sur plusieurs tours ajoute 6,23 points d'AUROC inter-domaines pour GPT-5.5 et 2,95 sur MMAD, culminant après environ trois appels (AnomalyClaw) single sources . « Un élève entraîné sur des étiquettes d'enseignant approche les étiquettes humaines » n'a été montré que sur des images naturelles : un ResNet-101 entraîné sur des étiquettes Qwen3-VL de COCO et Objects365 a atteint 95,7–99,4 % du résultat des étiquettes humaines, et les auteurs disent que les données de domaine ne sont pas testées (arXiv 2602.20972) single source .

FIG 13 **Deux chaînes enseignant–élève** ***Pour l'inspection, un modèle de pointe pré-étiquette et les humains arbitrent ; pour les bras robot, les vrais enseignants restent les démonstrations humaines.***



Source : Synthèse de l'auteur (chapitre 3) ; coûts et tailles d'échantillons sont des estimations.

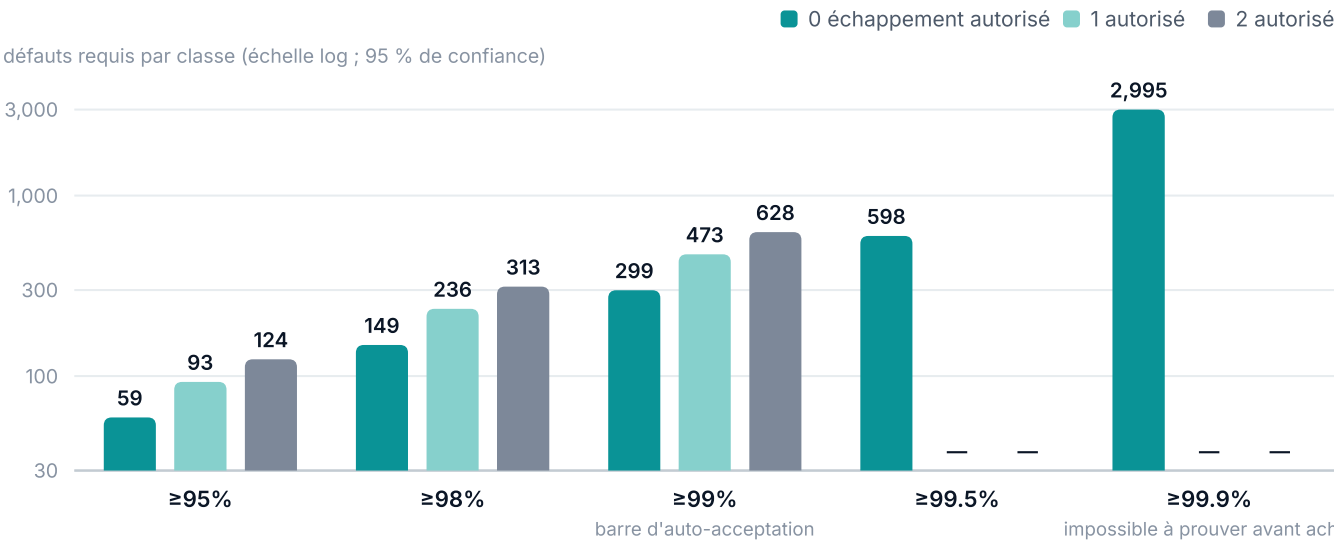
Un protocole réalisable estimate :

1. **Collecter le jeu de référence depuis les propres archives AOI de l'usine et les cartes NG conservées**, stratifié en 12 vraies classes de défauts (manquant, décalé, tombstone, polarité, mauvaise pièce, pont/court-circuit, soudure insuffisante, soudure excédentaire, bille de soudure, broche soulevée/coupure, pièce endommagée, corps étranger) et quatre causes de fausses alertes (ombre, flux, sérigraphie, tolérance), plus une « tranche de dérive » d'au moins 300 défauts d'une seconde ligne ou d'un éclairage, d'une couleur de carte ou de références différents.

2. **Étiqueter en aveugle avec deux ou trois inspecteurs formés à l'IPC-A-610**, un ingénieur senior décidant et les rayons X, le test électrique ou les coupes transversales tranchant les échantillons ambigus. N'accepter une définition de classe que si l'accord entre inspecteurs atteint κ ≥ 0,75 (≥ 0,90 pour les classes critiques) ; l'accord humain sur 11 morphologies de défauts électroniques dans LDU-Bench n'était que de 0,727, donc certaines classes devront être fusionnées ou redéfinies. L'accord d'attribut de style AIAG (efficacité ≥ 90 %, manqué ≤ 2 %, fausse alarme ≤ 5 %, κ > 0,75 ; > 0,90 pour la mise en production) est un minimum raisonnable (Fabrico, iFactory).
3. **Le dimensionner avec des bornes binomiales exactes (Clopper–Pearson) unilatérales à 95 %** (intervalles de confiance binomiaux, règle de trois :

Rappel à prouver	0 manqué autorisé	1 autorisé	2 autorisés
≥ 95 %	59 défauts	93	124
≥ 98 %	149	236	313
≥ 99 %	299	473	628
≥ 99,5 %	598	—	—
≥ 99,9 %	2 995	—	—

FIG 14 **Quelle taille doit avoir le jeu de référence*Prouver 99 % de rappel avec zéro échappement exige 299 défauts par classe ; 99,9 % ne peut pas être prouvé avant achat.***



Source : Intervalles de confiance binomiaux exacts (95 %).

Soixante défauts par classe suffisent pour un écran succès/échec ; 300 sont nécessaires pour décider de l'acceptation automatique. Le jeu complet, environ 3 600 défauts plus 3 500 recadrages de fausses alertes, prend environ 37–57 heures de travail à étiqueter et arbitrer, et faire tourner un candidat enseignant dessus coûte 23–196 $; le vrai coût est deux à six semaines de collecte, fixées par la classe la plus rare estimate. Envoyez des recadrages AOI à composant unique, pas des cartes entières, avec la cible d'au moins 42–48 pixels de large (arXiv 2610.09313).

1. **N'accepter automatiquement les étiquettes de l'enseignant que pour une classe sans aucun manqué sur ≥ 299 défauts de référence**, un accord modèle–vérité pas pire que l'accord humain–humain, et au moins 50 % des fausses alertes clôturées. Les classes critiques comme les courts-circuits, les pièces manquantes, la polarité et les mauvaises pièces exigent 99,5–99,9 % (598–2 995 défauts) ; sinon, tout ce que l'IA ne peut clôturer avec une haute confiance va à une personne. Un taux d'échappement sous 0,1 % des défauts ne peut être prouvé

avant achat (2 995 défauts sans manqué ; 0,01 % en prendrait 29 956), donc la ligne garde un échantillonnage humain des appels clôturés par l'IA, l'ICT/FCT en aval comme filet et le SPC sur les échappements. Les étiquettes vérifiées sont ensuite distillées dans un détecteur ou classifieur conventionnel et revalidées à chaque nouveau produit.

Les conditions d'utilisation et les règles de données décident aussi de l'enseignant. Anthropic interdit d'utiliser ses services pour entraîner des modèles concurrents (conditions commerciales Anthropic) ; OpenAI autorise les sorties pour entraîner des classifieurs internes non vendus à des tiers (conditions commerciales OpenAI) ; DeepSeek autorise explicitement la distillation (conditions DeepSeek) primary terms. Aucune ne définit un « modèle concurrent », donc l'usage commercial exige une revue juridique, et les règles de données des clients gardent généralement les images sur site. Le défaut le plus sûr est un enseignant à poids ouverts exécuté localement, avec les modèles frontières cloud uniquement lorsque le client l'autorise par écrit.

Élèves d'inspection : petits et rapides ; les goulots d'étranglement sont l'optique, les données et la validation

L'élève n'est pas le goulot d'étranglement : YOLO26n prend 3,80 ms par image en INT8 sur un Orin Nano Super à 249–399 $ (Ultralytics) vendor test, NVIDIA rapporte 98,51 % sur trois classes de défauts PCB après adaptation d'un modèle de fondation sur environ 600 images (NVIDIA) vendor claim, et les défauts synthétiques n'ajoutent que 0,9–2,5 points de mAP (arXiv 2605.04635). Les contraintes sont l'optique, les données de ligne, la validation et les licences : Ultralytics YOLO est AGPL-3.0 ou sous licence entreprise (Ultralytics), donc les produits fermés ont besoin d'une licence ou d'un détecteur Apache-2.0 comme RF-DETR.

Élèves robots : les postes structurés sont proches du niveau production ; le travail flexible et dynamique ne l'est pas

En novembre 2026, il y a trois sortes d'enseignants robots, et une seule peut devenir un élève local : les VLA ouverts et affinables (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Les modèles robots fermés (π0.6/π0.7 via une API bêta, Gemini Robotics 2 pour testeurs de confiance) ne peuvent pas être déployés sur site (The AI Insider, DeepMind). Les modèles généraux frontières sont lents et médiocres en manipulation fine : dans le test d'Artificial Analysis, ils ont obtenu 46–69 avec 28–45 % de réussite de préhension et environ 30 minutes par tâche, « bien trop lents et coûteux pour un usage en production » (AA-Robotics) ; GPT-6 Astra n'a réussi un emplacement de précision que 2 fois sur 20 (RoboCurve) ; et piloter directement les articulations donne 0–5,5 % sur LIBERO à 0,2–0,4 Hz contre environ 83 Hz nécessaires (Anthropic) verified across three sources. Le 64,9 % de Fable 5.2 sur le benchmark simulé RoboDojo est un test interne (Anthropic) vendor claim.

Les études tierces depuis octobre concordent : environ 33–44 % sur LIBERO-10 pour les trois agents leaders (arXiv 2610.30051) ; des agents revendiquant faussement le succès sur 24–39 % des sous-tâches (arXiv 2611.08209) ; et 104 des 108 tâches quasi statiques sur robot réel pour Opus 5.5, mais à 280 secondes chacune et sans insertion ni assemblage (AMP1) single sources. Les modèles frontières appartiennent au hors ligne : écrire des programmes, orchestrer des tâches et rédiger des critères de succès.

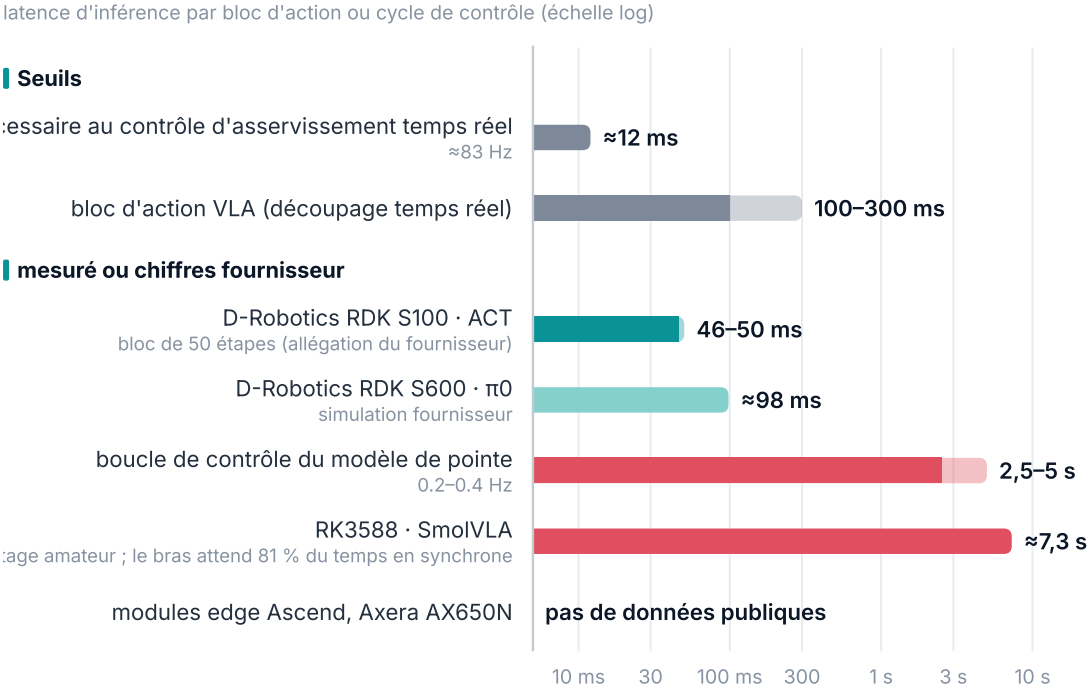
La qualité des élèves a un net avantage : bonne dans la distribution d'entraînement, s'effondrant en dehors. π0.7 « dépasse souvent 90 % » sur les tâches vues et atteint 60–80 % sur les inédites (article π0.7) vendor research ; ContactDP atteint 97,8 % sur des connecteurs entraînés mais 28,6–66,7 % sur des inédits (ContactDP) ; de légers changements de caméra ou de pose font chuter 95 % sous 30 % dans LIBERO-Plus (LIBERO-Plus, CVPR 2026) ; la préhension apprise sur convoyeurs en mouvement plafonne à 58,89 % (ForeTime-VLA) single sources, un travail que la poursuite par vision classique fait en production depuis des années.

Les preuves proches de la production viennent toutes de postes structurés, de démonstrations humaines et de correction sur robot réel. Neurometka a gardé une machine à états et un mouvement pré-enseigné et n'a appris que la préhension, l'insertion de câble et la soudure, chacune à partir de moins de 20 minutes de données : 322 sur 324 en production, 12,8 % plus lent qu'une personne (Neurometka) single source, vendor. Fraunhofer a atteint 92,8 % sur

quatre connecteurs à partir de 68 minutes de démonstrations (Fraunhofer) ; HIL-SERL de Berkeley a réussi 100 fois sur 100 l'insertion de RAM, SSD et USB après 1–2,5 heures de RL sur robot réel (HIL-SERL) ; AgiBot a tourné à 18–20 secondes par étape dans un takt de 20 secondes sur une ligne de test de tablettes (IThome) vendor claim .

La distillation achète la vitesse, pas la généralisation : π0.5 tourne à 0,70 Hz sur Jetson Orin brut et 10,30 Hz après ordonnancement asynchrone (EagleVLA) ; la distillation de couches le fait passer d'environ 3 Hz à environ 10 Hz (Shallow-π) verified . Un élève n'hérite que de ce que l'enseignant a couvert ; chaque nouveau produit exige encore des données de tâche.

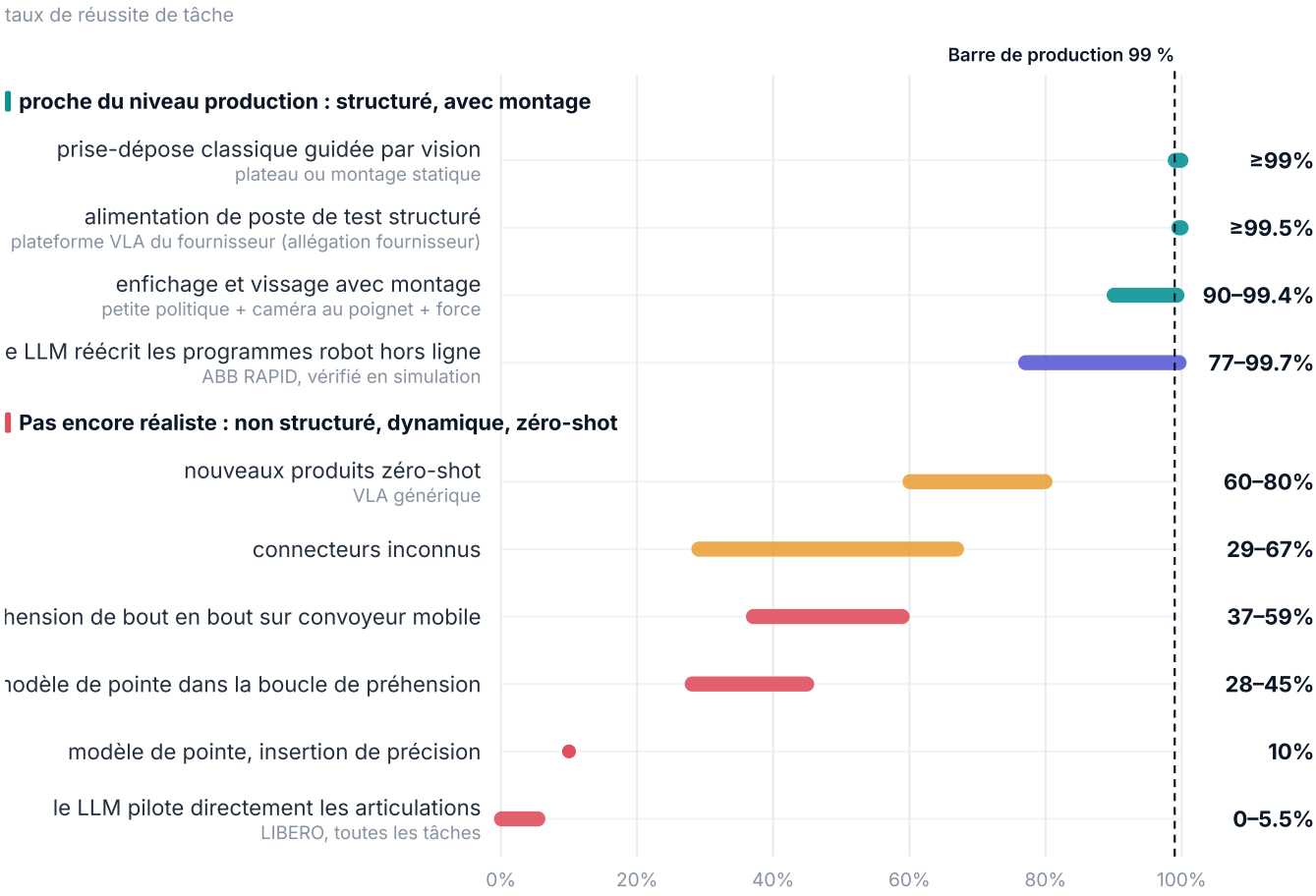
FIG 15 **Latence de politique edge** ***Une petite politique ACT tourne en ≈50 ms sur une carte edge ; un VLA porté sur une puce amateur prend 7,3 s.***



Source : Documentation et forum D-Robotics (chiffres fournisseur) ; ports communautaires GitHub ; recherche Anthropic.

Sur les puces edge chinoises, la seule politique apprise mesurée est une politique ACT sur le RDK S100 de D-Robotics à 46–50 ms par bloc de 50 étapes, sans taux de réussite publié (docs D-Robotics RDK S) vendor claim ; un portage amateur de SmolVLA sur un RK3588 prend environ 7,3 s par bloc (GitHub). La quantification n'achète pas la vitesse gratuitement : aucune compilation 4 bits de π0.5 n'a été plus rapide que le 16 bits à taille de lot 1 (LoopQ). Les démonstrations sont coûteuses (Reuters a rapporté environ 300 tentatives par mouvement utilisable pour un téléopérateur novice (Notebookcheck)), et la validation coûte plus : prouver un taux d'échec ≤ 0,1 % à 95 % de confiance exige environ 3 000 succès consécutifs, 17–22 heures-robot estimate . Aucune politique de manipulation apprise n'a de certification de sécurité.

FIG 16 **Élèves robots : proches du niveau production seulement si le poste est structuré** ***Les postes avec montage atteignent 90–99 %+ ; les tâches non structurées, dynamiques et zéro-shot ne sont pas encore réalistes.***



Source : Articles et supports fournisseurs cités au chapitre 3 ; certains chiffres sont des allégations fournisseur.

Tâche électronique	Élève réalisable	Succès attendu	Vitesse vs une personne	Coût par changement de série	Verdict 2026–2028
Pick-and-place à forte mixité, kitting, alimentation machine (plateau ou montage statique)	Perception apprise plus mouvement programmé ou planifié	≥ 99 % avec guidage vision classique	Comme les postes conventionnels	Heures	**Faisable maintenant**
Insertion de connecteurs fixes, FFC et câbles, vissage, soudure	Machine à états + petite politique BC/diffusion/ACT + caméra au poignet + force/couple (idéalement toucher) + correction humaine	90–99,4 %	10–15 % plus lent	10–300 démos (≈ 15–70 min) + 1 000–3 000 passages de validation (6–22 heures-robot)	**Faisable, intensif en R&D ; des mois pour le premier poste**
Alimentation de poste de test structuré et tri OK/NG	Plateforme VLA de vendeur + jumeau numérique + RL sur robot réel	≥ 99,5 % **vendor claim, conflicting**	18–20 s par étape, dans un takt de 20 s	≤ 4 h de réentraînement + 15 min de calibration	**Faisable sur une plateforme de vendeur**
Programmation de changement de série	Le LLM écrit ou modifie des programmes robots hors ligne, vérifiés en simulation	77–99,7 % sur des réécritures ABB RAPID	Aucun effet sur le takt	Minutes à heures par programme	**Faisable comme outil d'ingénierie**

Tâche électronique	Élève réalisable	Succès attendu	Vitesse vs une personne	Coût par changement de série	Verdict 2026–2028
Prélèvement appris de bout en bout sur convoyeurs en mouvement	Élève VLA	37–59 % ; ≤ 37 % sur tapis rapides	—	≥ 458 démos par configuration	**Pas avant 2028 ; utiliser la poursuite classique plus la perception apprise**
Nouveaux produits sans nouvelles données (zero-shot)	VLA générique	60–80 % ; connecteurs inédits 29–67 %	—	—	**Pas réaliste**
LLM frontière dans la boucle pilotant le bras	Agents GPT-6, Claude, Gemini	Préhension 28–45 % ; insertion de précision 10 %	≈ 30 min ou 280 s par tâche contre ≈ 20 s	—	**Pas réaliste**
Faisceaux de câbles à routage libre et assemblage flexible multi-étapes	—	Aucune preuve de production 2026	—	—	**Pas avant 2028**

Sources : les études ci-dessus et (Université de Passau sur la réécriture RAPID, GraspVLA, Komax) **mostly single source and vendor data** . En 2026–2028, la valeur d'un grand modèle pour les robots est celle d'un a priori pré-entraîné et d'un programmeur hors ligne, pas de quelque chose qui « enseigne » un élève. Chaque cas proche de la production que j'ai trouvé (Neuromeka, AgiBot chez Longcheer, Fraunhofer, ContactDP) repose sur des démonstrations humaines, la détection de force ou de toucher, la simulation ou des jumeaux numériques et le RL supervisé. Aucun ne repose sur des étiquettes d'un modèle frontière.

Coûts et retour sur investissement

04 Coûts et retour sur investissement : la revue IA est rentabilisée en mois, l'alimentation machine en années, les bras de prélèvement jamais

Coût de la main-d'œuvre : les États-Unis sont environ 5,5 fois Shenzhen, Taïwan environ 1,5 fois, la majeure partie de l'Asie du Sud-Est moins de la moitié

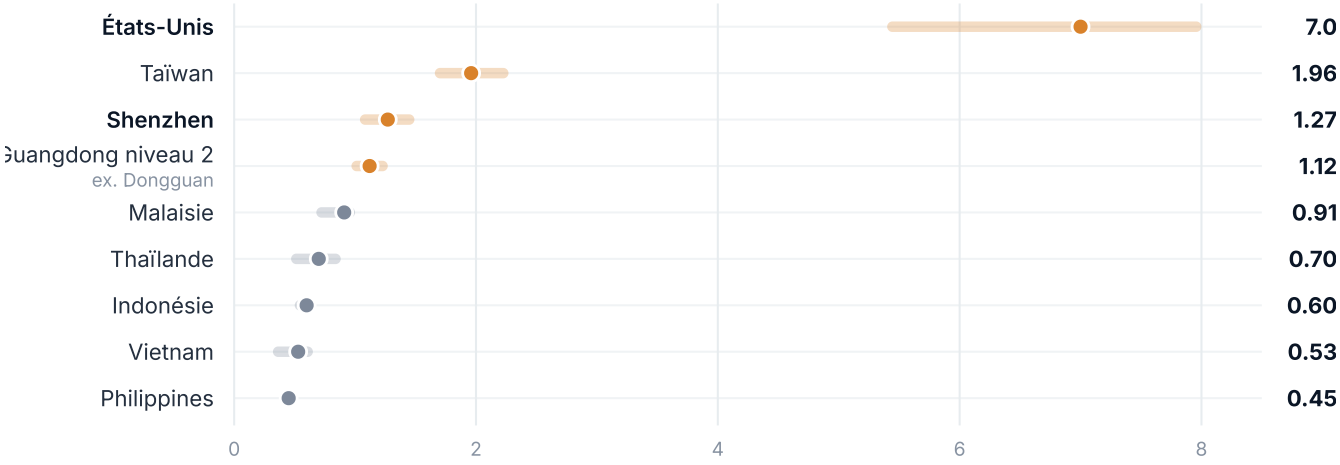
Le retour sur le remplacement de la main-d'œuvre commence par ce que coûte la main-d'œuvre. La guidance salariale officielle de Shenzhen place la médiane 2024 d'un inspecteur qualité manufacturier à **74 074 RMB** par an ; les salaires représentent 86,92 % du coût total de la main-d'œuvre dans la fabrication électronique, donc le coût chargé est d'environ 85 000 RMB. L'enquête 2025 de la JETRO place le coût annuel d'un ouvrier d'usine chinois à 12 667 $, environ 85 100 RMB, à 0,2 % près (guidance salariale de Shenzhen, JETRO 2025) verified .

Région	Coût annuel chargé d'un inspecteur ou manutentionnaire (bas / base / haut)	Base
Chine continentale, Shenzhen	70 k / **85 k** / 100 k RMB (10,4 k / **12,7 k** / 14,9 k $)	Médiane officielle × 1,15 ; concorde avec la JETRO verified
Zones salariales Guangdong tier-2 (Dongguan etc.)	65 k / **75 k** / 85 k RMB	Médiane Guangdong 2023 × 1,15 plus deux ans d'augmentations estimate
Delta du Yangtsé (Suzhou, Kunshan, Wuxi)	Comme Shenzhen	Aucune base pour une main-d'œuvre d'inspection moins chère ; le salaire minimum le plus élevé du Jiangsu est environ 5,6 % au-dessus de celui de Shenzhen estimate
Taïwan	16,6 k / **19,6 k** / 22,7 k $ (534 k / **630 k** / 729 k NT$)	Enquête salariale par profession du ministère du Travail plus ≈ 1,18 de charges légales ; JETRO comme haut de fourchette estimate
Malaisie	6,8 k / **9,1 k** / 10,0 k $	JETRO et salaires minimums estimate
Thaïlande	4,7 k / **7,0 k** / 8,8 k $	Idem
Indonésie	5,0 k / **6,0 k** / 6,3 k $	Idem
Vietnam	3,2 k / **5,3 k** / 6,5 k $	Idem
Philippines	4,0 k / **4,5 k** / 4,9 k $	Idem
États-Unis	54 k / **70 k** / 80 k $	Médiane d'inspecteur BLS mai 2025 de 48 570 $ ÷ salaires à 65,5 % de la rémunération estimate

Sources : (JETRO 2025, guidance salariale du Guangdong, bureau RH de Wuxi, enquête salariale par profession de Taïwan, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **Ce que coûte par an un agent d'inspection ou de manutention*À plein chargé, les États-Unis sont ≈5,5× Shenzhen et Taïwan ≈1,5×.***

coût annuel chargé par agent CQ ou manutentionnaire (10k $; point = base, barre = bas–haut)



Source : Guidance salariale officielle de Shenzhen, ministère du Travail de Taïwan, JETRO, BLS ; estimations de l'auteur.

Trois corrections établissent la base. Premièrement, Shenzhen a relevé son plancher de cotisation retraite de 4 775 à 5 521 RMB par mois en juillet 2026 (Shenzhen Bendibao) single source, donc un inspecteur de Shenzhen en 2026 coûte plus probablement 80–90 k RMB. Deuxièmement, un poste n'est pas une personne : deux équipes de 12 heures exigent environ 2,0 personnes (six jours par semaine) à 2,4 (toute l'année) dans le delta de la rivière des Perles, environ 2,7–3,1 en Asie du Sud-Est sous plafonds d'heures supplémentaires, et 4,2–4,5 pour du 24×7 aux États-Unis et à Taïwan estimate. La revue IA permet généralement à un réviseur de couvrir deux à quatre lignes, donc l'économie réaliste est **0,5–1,5 personne à temps plein par ligne, pas une personne par ligne et par équipe**. Troisièmement, les acheteurs à Taïwan, au Vietnam et en Malaisie ne cherchent pas seulement le coût : dans l'enquête de la JETRO, 61,1 % des entreprises à Taïwan, 48,2 % au Vietnam et 44,9 % en Malaisie ont dit que le recrutement était devenu plus difficile au cours des deux dernières années, contre 18,6 % en Chine continentale (JETRO 2025) single source.

Prix des équipements : quatre bases qu'il ne faut pas moyenner

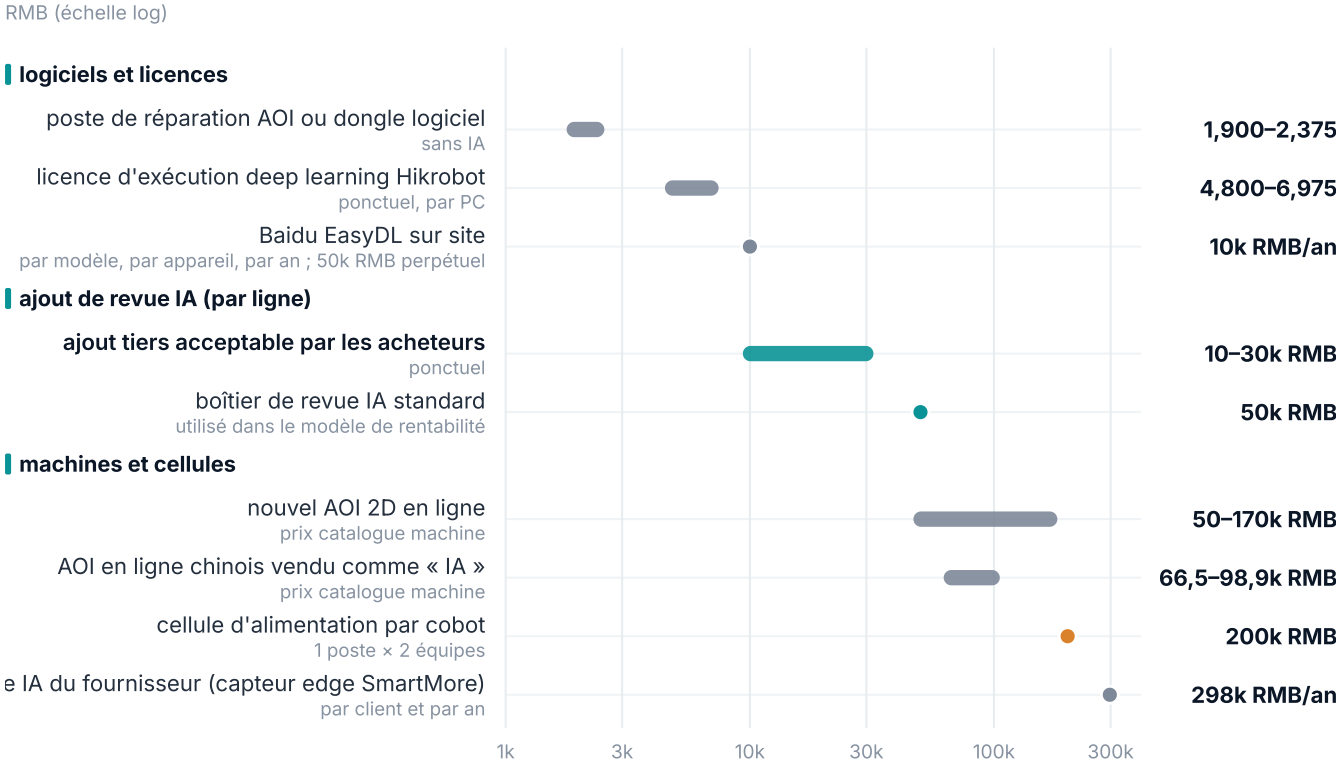
Les prix des AOI divergent surtout parce qu'ils mesurent des choses différentes. Les annonces en ligne (novembre 2026, neuf, occasion et location mélangés) tournent autour de 17–80 k RMB pour la 2D hors ligne ou de bureau, 50–170 k RMB pour la 2D en ligne et 80–475 k RMB pour la 3D en ligne, jusqu'à 700 k RMB pour la 3D coréenne (AOI en ligne 1688, AOI hors ligne 1688, AOI 3D 1688) single source, list prices. Un appel d'offres de novembre 2025 a plafonné un AOI 3D plus un SPI 3D à 800 k RMB (avis d'appel d'offres). Le prospectus de Sinic-Tek implique 428–512 k RMB par AOI 3D en 2019–2022, environ 340–470 k RMB selon son rapport 2025 (prospectus Sinic-Tek, rapport annuel 2025 de Sinic-Tek) primary + estimate. Un vendeur étranger estime 3 000–20 000 $ pour la 2D de bureau, 30–70 k $ la 2D en ligne, 50–90 k $ la 3D d'entrée et 100–200 k $+ la 3D haut de gamme, la même machine en ligne à 30–50 k $ en Chine (Averroes) vendor estimate from a seller of competing AI software.

La revue IA elle-même n'a aucun prix public en RMB ; les vendeurs chiffrent au cas par cas après une évaluation gratuite sur site (Sohu sur Hengtianyi), et les recherches d'annonces e-commerce, d'appels d'offres, d'articles de vendeurs et de médias taïwanais n'ont trouvé aucun devis public (1688, 51SMT). Les points d'ancrage : environ 27 k € par ligne impliqués par le cas Siemens estimate ; 21 924 $ catalogue pour le module réseau de revue à distance d'un distributeur américain (Manncorp) vendor list price ; environ 298 k RMB par client et par an pour les capteurs edge AI de SmartMore en 2025 (36Kr) single source. Un AOI 2D en ligne national vendu comme « IA » est listé à 66,5 k RMB (AOI en ligne 1688), et un tampon NG pour réinspection manuelle environ 50 k RMB (revue AOI 1688). Le pla-

fond pour un module de revue tiers est donc d'environ 10–30 k RMB par ligne, ou un abonnement annuel similaire ; au-dessus, l'acheteur préférerait remplacer la machine par un AOI national avec IA intégrée estimate.

Point d'ancrage	Prix (RMB)	Ce qu'il achète	Source
Poste de réparation AOI ou dongle logiciel (anciens TRI, Zhenhuaxing)	≈ 1 900–2 375	Clé de licence, sans IA	1688 list price
Runtime de vision par apprentissage profond général (dongle Hikrobot VisionMaster)	≈ 4 800–6 975 ponctuel par PC	La plateforme ; le client construit et maintient les modèles	1688 list price
Déploiement sur site Baidu EasyDL	10 000 par modèle par appareil et par an, ou 50 000 perpétuel	Service d'inférence packagé	docs Baidu AI 2022 price page
AOI en ligne national vendu comme « IA »	66 500–98 900 par machine	Remplacement de machine	1688 list price
Capteurs edge AI SmartMore	≈ 298 000 par client et par an	Solution de vendeur sur plusieurs machines	36Kr single source

FIG 18 **L'échelle de prix** ***De quelques milliers de RMB de licence logicielle à 200–300k RMB pour une cellule robot ou une offre IA fournisseur.***



Source : Annonces 1688, documentation Baidu AI (tarif 2022), 36Kr, entretien industriel (non vérifié).

Un abonnement annuel de style Baidu, avec un serveur d'inférence pour une à quatre lignes et un à quatre modèles chacune, implique un abonnement de 2,5–40 k RMB par ligne et par an, en réalité environ 5–10 k RMB ; à 20 % d'une licence perpétuelle, abonnement et achat s'équilibrent la cinquième année, une structure à copier estimate. Les acheteurs sont habitués à des frais logiciels quasi nuls : la totalité du chiffre d'affaires « autre » de Sinic-Tek (pièces détachées, service, location, accessoires) représentait 7,65 % des ventes 2025, environ 2 300–3 100 RMB par machine installée et par an (rapport annuel 2025 de Sinic-Tek) primary + estimate. Un abonnement IA de 5–10 k RMB par ligne et par an représente deux à cinq fois ce que les acheteurs dépensent aujourd'hui en logiciel AOI,

donc il doit être justifié par les heures de revue qu'il libère (environ 75–90 k RMB par personne à temps plein et par an), pas par comparaison avec les frais logiciels existants.

Pour les bras robotisés, le prix moyen des cobots en Chine est tombé à environ 58 600 RMB en 2025 (GGII) estimate, et la moyenne six axes de Dobot a chuté de 19,1 % cette année-là à mesure que les modèles moins chers gagnaient des parts (Huxiu) verified, même si les prix catalogue des mêmes modèles ont augmenté de 5–15 % en 2026 (Sina Finance). Aux États-Unis, un FANUC CRX-10iA était listé à 51 434 $ (Vention) et les compilations placent un UR7e à 30–48 k $ (cobotcost) estimate. Le bras ne représente que 40–50 % d'une cellule américaine (Standard Bots) ; en Chine, caméras, pinces, sécurité et ingénierie atteignent facilement deux à quatre fois un bras à 40 k RMB, donc budgétez le coût d'ingénierie absolu plutôt qu'un multiple estimate.

Retour sur investissement : revue en mois, alimentation en années, prélèvement jamais

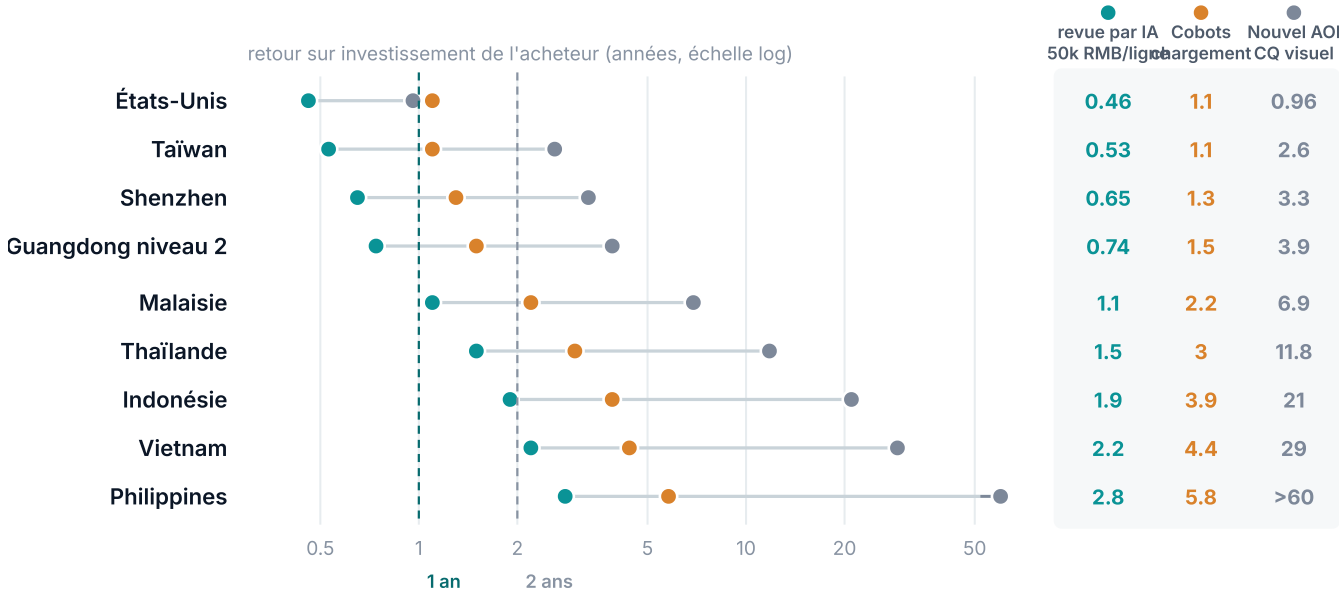
Retour simple = investissement ÷ (personnes économisées × coût de main-d'œuvre chargé − coût de fonctionnement annuel), pour une ligne à deux équipes sans subventions ni droits de douane. L'équipement taïwanais est au prix de 1,3 fois la Chine, le Vietnam, la Malaisie et la Thaïlande à 1,15 fois, l'Indonésie et les Philippines à 1,2 fois, et les États-Unis aux prix étrangers estimate.

Scénario	Investissement (Chine / États-Unis)	Personnes économisées	Coût de fonctionnement annuel
(a) Boîtier de revue IA ou boîtier caméra IA	50 k RMB / 30 k $ par ligne	1,0 par ligne	15 %
(b) Nouvel AOI 2D en ligne remplaçant les contrôles visuels manuels (0,5 réviseur par équipe reste)	200 k RMB / 60 k $	1,0	12 %
(c) Cellule cobot de chargement (un poste × deux équipes)	200 k RMB / 140 k $	2,0	8 %
(d) Cellule intégrée détection-et-prélèvement (inspection 1,0 + prélèvement 0,07)	300 k RMB / 200 k $	1,07	8 %
(e) Un bras robotisé ajouté uniquement pour prélever les cartes NG	150 k RMB / 100 k $	0,07	8 %

Région	(a) Revue IA	(b) Nouvel AOI	(c) Cellule cobot	(d) Détection-et-prélèvement	(e) Bras de prélèvement
Shenzhen	0,65 an (0,54–0,80)	3,3 (2,6–4,4)	**1,3** (1,1–1,6)	4,5 (3,6–5,9)	Jamais
Guangdong tier-2	0,74	3,9	1,5	5,3	Jamais
Taïwan	0,53	2,6	1,1	3,6	Jamais
Malaisie	1,1	6,9	2,2	9,2	Jamais
Thaïlande	1,5	11,8	3,0	15	Jamais
Indonésie	1,9	21	3,9	25	Jamais
Vietnam	2,2 (1,6–4,5)	29	4,4	34	Jamais
Philippines	2,8	> 60 (coût de fonctionnement ≈ économies)	5,8	≈ 100	Jamais

Région	(a) Revue IA	(b) Nouvel AOI	(c) Cellule cobot	(d) Détection-et-prélèvement	(e) Bras de prélèvement
États-Unis	0,46	**0,96**	**1,1**	3,4	Jamais

FIG 19 **Rentabilité par région** ***La revue IA se rentabilise en mois aux États-Unis, à Taïwan et à Shenzhen et en 1–3 ans en Asie du Sud-Est ; l'alimentation par cobot en années.***



Un bras robot juste pour extraire les cartes NG ne rentabilise jamais, nulle part
Avec un rendement de premier passage >90 %, l'extraction ne prend que 0,003–0,14 ETP par équipe

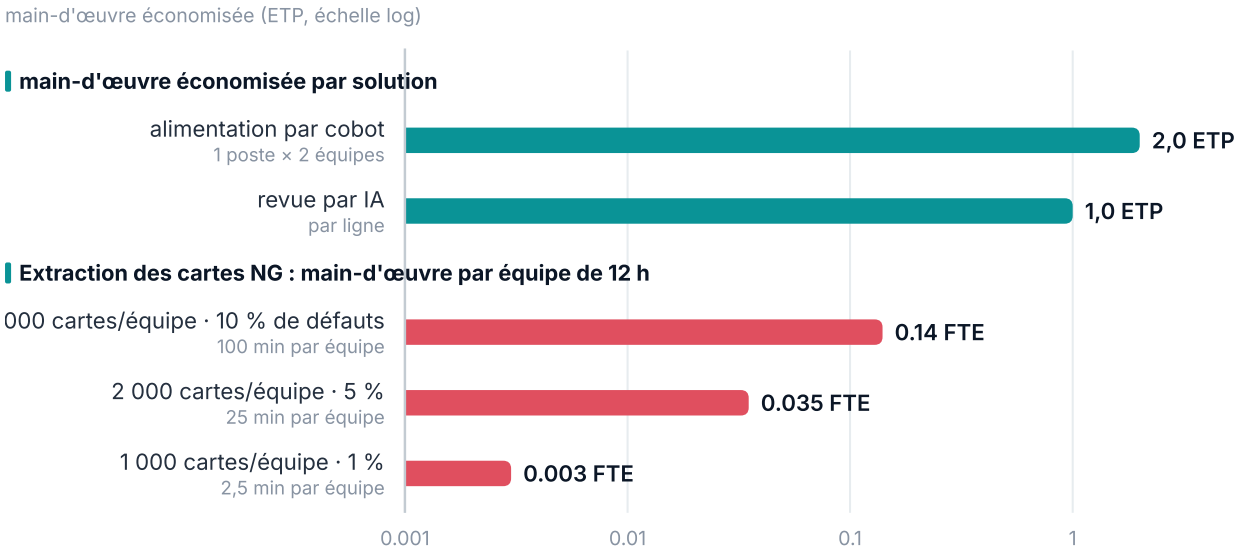
Source : Modèle de l'auteur : boîtier de revue IA 50k RMB par ligne ; cellule d'alimentation cobot à 200k RMB ou 140k $, deux équipes.

Toutes les valeurs sont des estimations à partir des entrées ci-dessus. Sensibilités : divisez par deux les personnes économisées et chaque retour double environ ; faites tourner une équipe et (b)–(d) doublent ; l'amortissement de 100 % la première année aux États-Unis raccourcit le retour d'environ 20 %, tandis que les droits de douane sur l'équipement chinois, japonais ou coréen l'allongent d'environ 10–40 % ; une subvention locale de 15 % en Chine le raccourcit d'environ 15 %. Je n'ai trouvé aucune enquête indépendante sur les seuils de retour des acheteurs ; les vendeurs disent que l'AOI se rembourse « généralement en 9–18 mois » (Averroes) et le cas Siemens revendique huit mois vendor claims. J'utilise un et deux ans comme lignes de filtrage. Le modèle omet les gains de qualité (garantie, retours) et les coûts de rotation et de formation, donc il est conservateur.

La partie contre-intuitive de « détecter et prélever » est que la main-d'œuvre est dans la revue, pas dans le prélèvement des quelques vraies cartes NG hors de la ligne. À 15 secondes de manutention par carte NG estimate :

Cartes par équipe	Taux de défauts	Temps de prélèvement par équipe de 12 heures	Personnes
1 000	1 %	2,5 min	0,003
2 000	5 %	25 min	0,035
4 000	10 %	100 min	0,14

FIG 20 **Pourquoi l'extraction des cartes NG ne rentabilise pas*Au-dessus de 90 % de rendement de premier passage, l'extraction ne prend que 0,003–0,14 de personne par équipe.***



Source : Estimations de l'auteur.

Le cas de base (2 000 cartes par équipe, 5 % défectueuses, deux équipes) représente 0,07 personne, soit environ 6 000 RMB par an à Shenzhen, moins que le coût de fonctionnement annuel d'environ 12 000 RMB du bras, donc cela n'est jamais rentabilisé nulle part ; même le pire cas (4 000 cartes par équipe, 10 % défectueuses) prend environ 13 ans à Shenzhen et 9 aux États-Unis estimate. Si l'inquiétude est d'arrêter la ligne pour prélever, un aiguillage NG ou un convoyeur tampon standard le résout sans bras. Les gens du métier disent « au-dessus de 90 % de rendement au premier passage, arrêtez la ligne et sortez la carte à la main », et **c'est juste pour le prélèvement, mais pas pour l'inspection**. La conception produit sur une ligne SMT est donc « inspecter plus marquer ou aiguiller », avec des bras robotisés réservés aux endroits où chaque pièce est manipulée (chargement de montage de test, manutention de plateaux et de magasins, emballage), où l'inspection devient un complément quasi gratuit et où le scénario (c) s'applique.

Subventions, impôts et droits de douane déplacent le retour de 15–25 %, mais pas le classement

Chine continentale : le soutien abaisse surtout le coût de financement ; les machines individuelles sont rarement éligibles directement. Une subvention nationale de prêt paie 1,5 point d'intérêt jusqu'à deux ans dans 14 chaînes prioritaires (Yicai) verified, environ 9 000 RMB sur une machine à 300 k RMB ; la prime de modernisation du Guangdong jusqu'à 20 % exige un projet enregistré d'au moins 7 millions de RMB de nouvel équipement et exclut Shenzhen (règles du Guangdong), et Dongguan a réellement versé 4,42 % en juillet 2026 (allocation de Dongguan) single source, primary. Un dispositif d'inspection IA n'est approuvé que comme équipement « avec logiciel de support » dans le projet enregistré de l'acheteur, donc le cas de base suppose aucune subvention.

Taïwan : les incitations vont à l'acheteur. En vertu de l'article 10-1 de la loi sur l'innovation industrielle, modifiée en mai 2025 pour inclure l'IA, un acheteur peut créditer 5 % des dépenses 2025–2029 en nouvelles machines intelligentes ou IA (y compris l'inspection basée CNN/DNN) contre son impôt sur le revenu de l'année, ou 3 % par an pendant trois ans (loi, règles de crédit) verified ; les subventions côté fournisseur sont fermées aux capitaux continentaux (MOEA) single source.

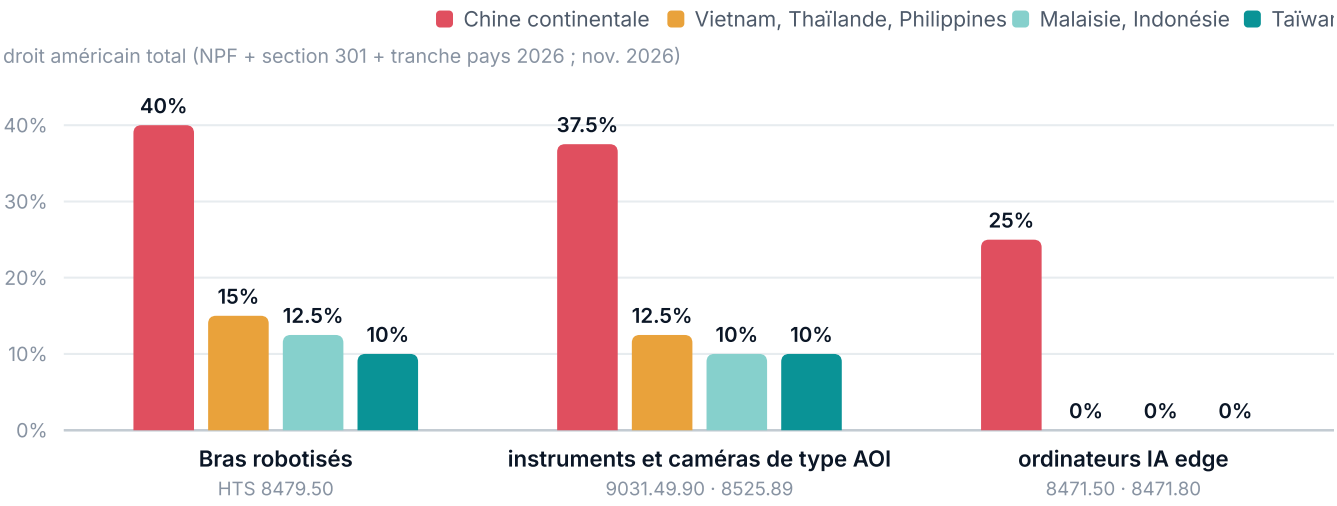
Asie du Sud-Est : seules la Malaisie et la Thaïlande récompensent l'achat d'automatisation. L'abattement d'impôt de 200 % sur l'automatisation en Malaisie sur les premiers 10 millions de RM vaut environ 24 % du capex pour un acheteur rentable (MIDA, PwC Malaisie) verified ; le BOI thaïlandais accorde une exonération fiscale de trois ans jusqu'à 50–100 % d'un investissement de modernisation, le plafond supérieur s'appliquant lorsque des fournisseurs

d'automatisation thaïlandais sont impliqués (BOI) single source, primary, donc entrer en Thaïlande fonctionne mieux avec un intégrateur local. Le Vietnam, l'Indonésie et les Philippines récompensent la construction d'usines plutôt que l'achat d'automatisation (Vietnam Law Magazine, PwC Philippines).

États-Unis : l'amortissement raccourcit le retour ; les droits de douane l'allongent. Les actifs achetés après le 19 janvier 2025 bénéficient d'un amortissement permanent de 100 % la première année, soit environ 21 % de bouclier fiscal fédéral pour un acheteur rentable (avis IRS 2026-11) single source, primary ; l'enquête Section 232 sur les robots et machines industrielles s'est terminée le 28 octobre 2026 par des négociations plutôt que de nouveaux droits de douane, à réévaluer sous 180 jours (Maison-Blanche, Mayer Brown) verified. L'assemblage final en Asie du Sud-Est réduit la surtaxe sur les bras et les instruments de type AOI d'environ 37,5–40 % à 10–15 % (chapitre 1).

Région	Principaux instruments	Effet sur le retour de l'acheteur estimate
Chine continentale	Subvention de prêt de 1,5 point jusqu'à deux ans ; primes de modernisation provinciales et municipales (le Guangdong exige des projets enregistrés ≥ 7 M RMB ; Dongguan a réellement versé 4,42 %) ; mesures Shenzhen 2026–2029 (taux non publics)	Environ 0–5 % pour une machine individuelle ; 4–15 % dans un grand projet
Taïwan	Crédit de la loi §10-1, 5 % ou 3 % × 3 ans	Environ 5 % plus court, si l'acheteur paie assez d'impôt
Malaisie	Abattement d'impôt de 200 % sur l'automatisation (premiers 10 M RM)	Environ 15–25 % plus court pour les acheteurs rentables
Thaïlande	Exonération fiscale BOI de trois ans jusqu'à 50–100 % de l'investissement	Environ 15–25 % plus court pour les acheteurs rentables hors congé fiscal
Vietnam, Indonésie, Philippines	Incitations fiscales pour la construction d'usines ; 20 % d'amortissement supplémentaire de l'équipement aux Philippines	Presque rien pour les achats d'automatisation individuels
États-Unis	Amortissement de 100 % la première année ; ≈ 37,5–40 % de droits cumulés sur l'équipement fabriqué sur le continent	Amortissement ≈ 20 % plus court ; droits ≈ 10–40 % plus long

FIG 21 **Droits américains sur les robots, AOI et ordinateurs edge*Les bras fabriqués en Chine continentale paient ≈40 % ; expédiés d'Asie du Sud-Est ou de Taïwan, 10–15 %.***



Source : Chapitre 99 du tarif douanier harmonisé USITC ; Federal Register ; USTR (novembre 2026).

Par région : en Chine continentale, un boîtier de revue IA se rembourse en mois et une cellule cobot en 1,3–1,5 an, tandis qu'un nouvel AOI acheté seulement pour réduire la main-d'œuvre prend 3–4 ans et la détection-et-prélèvement intégrée 4,5–5,3, tous deux au-delà d'un filtre de deux ans. Taïwan fait légèrement mieux que Shenzhen (0,53 an pour la revue, 1,1 pour une cellule cobot), et le recrutement y est plus difficile. En Malaisie et en Thaïlande, un boîtier IA prend 1–1,5 an et une cellule cobot 2–3, environ 15–25 % plus vite avec les incitations fiscales locales pour les acheteurs rentables. En Indonésie, au Vietnam et aux Philippines, seul un boîtier IA très bon marché se rembourse en 2–3 ans ; la demande y vient davantage des exigences de qualité des clients et des pénuries de main-d'œuvre. Aux États-Unis, chaque option sauf un bras de prélèvement pur se rembourse en 0,2–3,4 ans, avec les droits de douane, la réévaluation Section 232 et la certification NRTL comme principales variables.

Marché et concurrence

05 Marché, concurrence et comment les acheteurs paient

Taille du marché : les chiffres publiés diffèrent par des multiples ; seul un décompte ascendant est utile

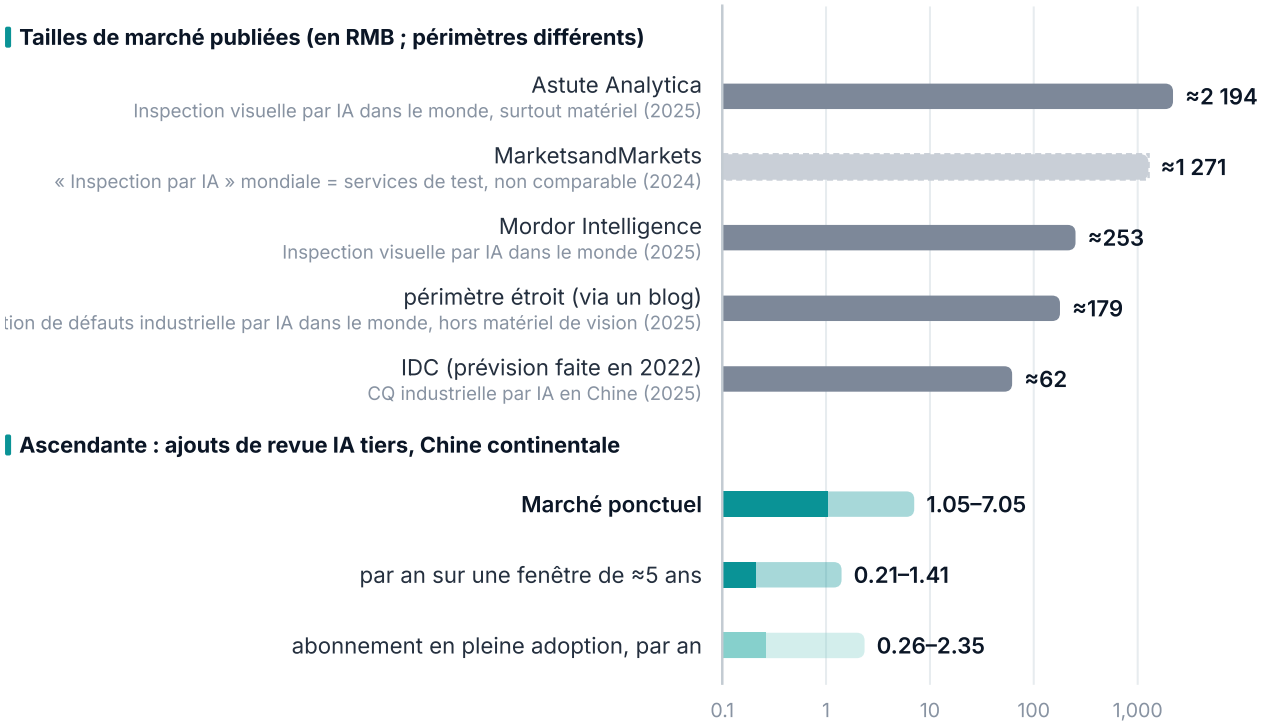
Chaque chiffre de marché doit être lu avec son périmètre. Le marché chinois de la vision industrielle était de 18,15 milliards de RMB en 2024 (Zhiyan) verified, l'électronique 3C étant le plus grand acheteur ; les estimations mondiales de la vision industrielle diffèrent d'environ trois fois et celles de l'AOI d'environ 6,5 fois (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Par agrégation ascendante des revenus de Koh Young, TRI, Jutze, Sinic-Tek et ViTrox, l'équipement mondial d'inspection SMT est de l'ordre de 1–2 milliards de dollars estimate.

Il n'existe pas de marché autonome pour les « logiciels d'inspection visuelle par IA ». Les chiffres publiés diffèrent de plus de dix fois et incluent tous du matériel ou des services :

Périmètre	Éditeur	Chiffre
« Détection de défauts industriels par IA » mondiale (étroit, hors matériel de vision)	Cabinet non nommé, via un blog iFactory	2,66 milliards de $ en 2025, TCAC de 8,6 % jusqu'en 2035, électronique 34 %
« Inspection visuelle par IA » mondiale	Mordor Intelligence	3,76 milliards de $ en 2025, 12,8 milliards de $ en 2031, TCAC de 22,67 %
« Inspection visuelle par IA » mondiale (surtout matériel)	Astute Analytica	≈ 32,66 milliards de $ en 2025
« Inspection par IA » mondiale (services de test et de certification ; non comparable)	MarketsandMarkets	18,92 milliards de $ en 2024
« Inspection qualité industrielle par IA » en Chine	IDC, prévision faite en 2022	958 millions de $ en 2025 (≈ 6,2 milliards de RMB), 91,5 % dans les communications, l'électronique et deux autres secteurs

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

FIG 22 **Les tailles de marché diffèrent de plus de dix fois** ***Seuls le périmètre étroit et un comptage ascendant sont utiles aux décisions.***



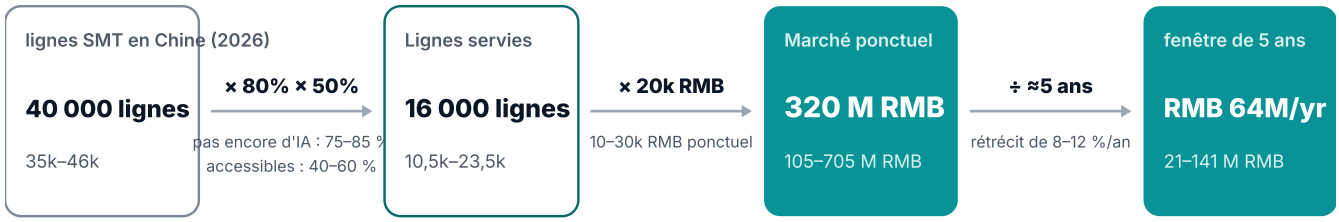
Source : Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory ; estimation ascendante de l'auteur.

Par agrégation ascendante, il n'existe qu'un seul décompte public des lignes SMT : New Venture Research a estimé environ 73 231 lignes dans le monde en 2020, 47 % en Chine (19 % en 2000), environ 70 % haut de gamme et 30 % bas de gamme (EMSNow, résumé NVR SMT2020) single source, 2020 ; NVR appartient désormais à IPC et les données plus récentes sont payantes (NVR). Une croissance de 3–5 % par an avec une part chinoise de 40–47 % donne environ 87 000–98 000 lignes dans le monde en 2026 et 35 000–46 000 en Chine continentale (base 40 000) estimate.

Entrée	Bas	Base	Haut	Base
Lignes SMT en Chine continentale (2026)	35 000	40 000	46 000	Base NVR 2020 × croissance × part
Part sans revue IA	75 %	80 %	85 %	Fenêtre de rétrofit, ci-dessous
Part qu'un tiers peut atteindre (pas interne, pas liée à l'IA OEM, peut payer)	40 %	50 %	60 %	Hypothèse : ≈ 30 % de lignes bas de gamme plus quelques lignes haut de gamme à marques mixtes, hors constructions internes de type DBG
Lignes adressables	**10 500**	**16 000**	**23 500**	Produit
Prix ponctuel par ligne	10 k RMB	20 k RMB	30 k RMB	Échelle et plafond de prix (chapitre 4)
Marché ponctuel adressable (base installée)	**105 M RMB**	**320 M RMB**	**705 M RMB**	Produit
Par an sur une fenêtre d'environ 5 ans	21 M RMB	64 M RMB	141 M RMB	La fenêtre recule de 8–12 % par an

Entrée	Bas	Base	Haut	Base
Abonnement (2,5 k / 5 k / 10 k RMB par ligne et par an, adoption complète)	26 M RMB par an	80 M RMB par an	235 M RMB par an	Point d'ancrage d'abonnement annuel de style Baidu

FIG 23 **Ascendant : 40 000 lignes → 16 000 servies → ≈320 M RMB ponctuel** ***La niche de revue IA tierce en Chine continentale est réelle, mais petite.***



Un nouvel entrant gagnant 2–5 % des lignes servies en 3 ans : ≈320–800 lignes, ≈2–5 M RMB par an
En abonnement (5k RMB par ligne et par an) : 1,6–4 M RMB récurrents, seulement 0,3–2,3 % de la prévision IDC du CQ industriel IA en Chine

Source : NVR (base 2020) ; estimations de l'auteur.

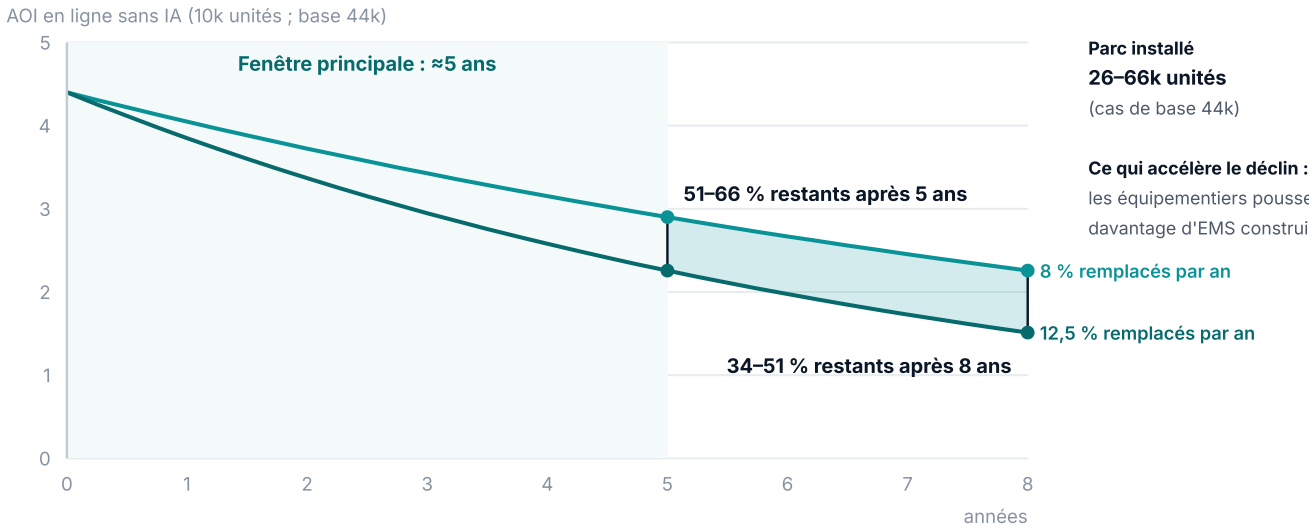
Tout cela est des estimations ; ni la part ni le prix n'ont été mesurés publiquement. La niche ne vaut que 0,3–2,3 % par an de la prévision d'IDC pour toute l'inspection qualité industrielle par IA de la Chine en 2025 (environ 6,2 milliards de RMB). Un nouvel entrant gagnant 2–5 % des lignes adressables de base en trois ans aurait environ 320–800 lignes : 6,4–16 millions de RMB au total à 20 000 RMB par ligne, environ 2–5 millions de RMB par an, ou 1,6–4 millions de RMB par an de revenus récurrents à 5 000 RMB par ligne estimate. Cela suffit pour une unité d'affaires au sein d'une entreprise, pas pour une entreprise autonome de taille venture. Chaque 1 000 lignes installées à Taïwan, en Asie du Sud-Est ou aux États-Unis vaut environ 3–15 millions de RMB ponctuels aux prix continentaux ; les prix sont plus élevés à Taïwan et aux États-Unis, mais les acheteurs y utilisent plus souvent des AOI 3D OEM avec options IA estimate.

La fenêtre de rétrofit est d'environ cinq ans

Aucun fabricant d'AOI ne publie la durée de vie des machines ; les EMS amortissent l'équipement sur 3–10 ans (rapport annuel 2025 de DBG), et les AOI anciens perdent mises à jour logicielles et pièces détachées après la fin de vie (I-Connect007) single source. J'utilise 8–12 ans (base 10).

À 1,0–1,65 AOI en ligne par ligne, le continent a environ 35 000–78 000 AOI en ligne en service (base environ 55 000), cohérent avec les expéditions connues (Circuits Assembly, rapport annuel 2025 de Sinic-Tek). Aucune source ne mesure combien font tourner une revue IA ; en supposant que les machines nationales l'offrent depuis environ 2023 et que 40–60 % des livraisons récentes l'utilisent, **environ 26 000–66 000 AOI en ligne en Chine continentale (base environ 44 000) ne font pas tourner l'IA** estimate. Si chaque nouvelle machine sort désormais avec l'IA, ce parc recule de 8–12,5 % par an : 51–66 % reste après cinq ans et 34–51 % après huit, plus vite si les fabricants d'AOI poussent les mises à niveau IA vers les machines plus récentes. **La majeure partie de l'opportunité se situe dans les cinq prochaines années environ**, avec en son cœur les EMS de taille moyenne exploitant d'anciens AOI 2D nationaux, des marques mixtes et aucune voie de mise à niveau IA.

FIG 24 **La fenêtre de retrofit est d'environ cinq ans** ***Les AOI en ligne sans IA diminuent de 8–12,5 % par an au fur et à mesure de leur remplacement ou mise à niveau.***

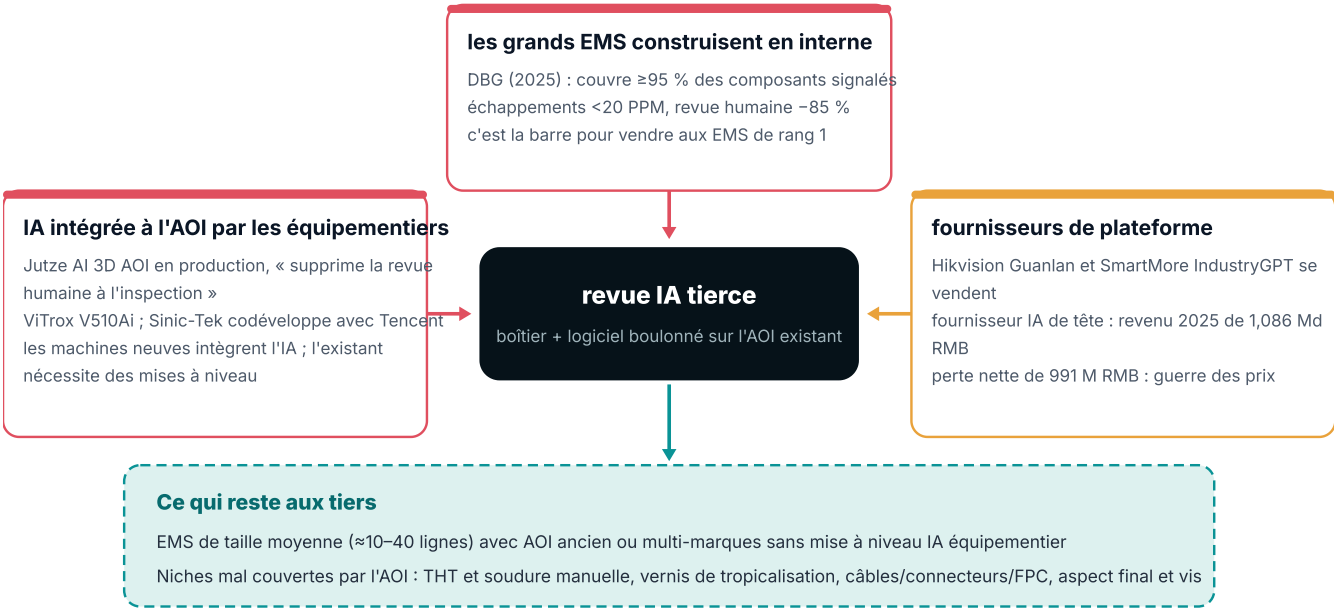


Source : Estimations de l'auteur ; parc et parts non vérifiés indépendamment.

Les concurrents sont déjà forts : le créneau de la revue est encerclé

Le créneau de la revue IA se rétrécit de trois directions. Les leaders de l'AOI et de la vision intègrent l'IA dans de nouvelles plateformes : Jutze a « pleinement adopté les algorithmes d'IA » (relations investisseurs Jutze, juin 2026), le chiffre d'affaires IA industriel d'OPT a crû de 96,77 % en 2025 (rapport annuel 2025 d'OPT) single source, primary, et ViTrox revendique jusqu'à 95 % de précision pour la revue autonome (ViTrox) vendor claim. Les grands fabricants construisent les leurs (iThome). Et des entreprises d'IA financées par le capital-risque vendent à grande échelle : les clients de capteurs edge de SmartMore sont passés de 65 à 251 en deux ans (36Kr) single source, Hikvision vend son modèle Guanlan pour l'inspection (Xinhua), et l'intégration « œil, cerveau et main » est déjà proposée par SmartMore, Hikrobot et Mech-Mind (36Kr, Securities Star).

FIG 25 **Le créneau de revue est encerclé sur trois côtés** ***La place des tiers est dans les EMS de taille moyenne et dans les niches d'inspection mal couvertes par l'AOI.***



Source : Rapports annuels Jutze, Sinic-Tek et DBG ; supports publics Hikvision et SmartMore ; compilé par l'auteur.

L'espace laissé à un nouvel entrant est étroit : les EMS de taille moyenne avec d'anciens AOI 2D/3D et sans mise à niveau IA du fabricant ; les lignes mélangeant plusieurs marques d'AOI (un fournisseur de rétrofit prend déjà en charge Omron, VI, Koh Young et Mirtec (Sohu sur Hengtianyi)) ; les niches que les AOI SMT couvrent mal, comme les joints traversants et soudés à la main, le vernis de tropicalisation, les câbles, connecteurs et FPC, et l'aspect final et les vis ; et les petites usines qui ne peuvent se payer un AOI à 230–280 k RMB.

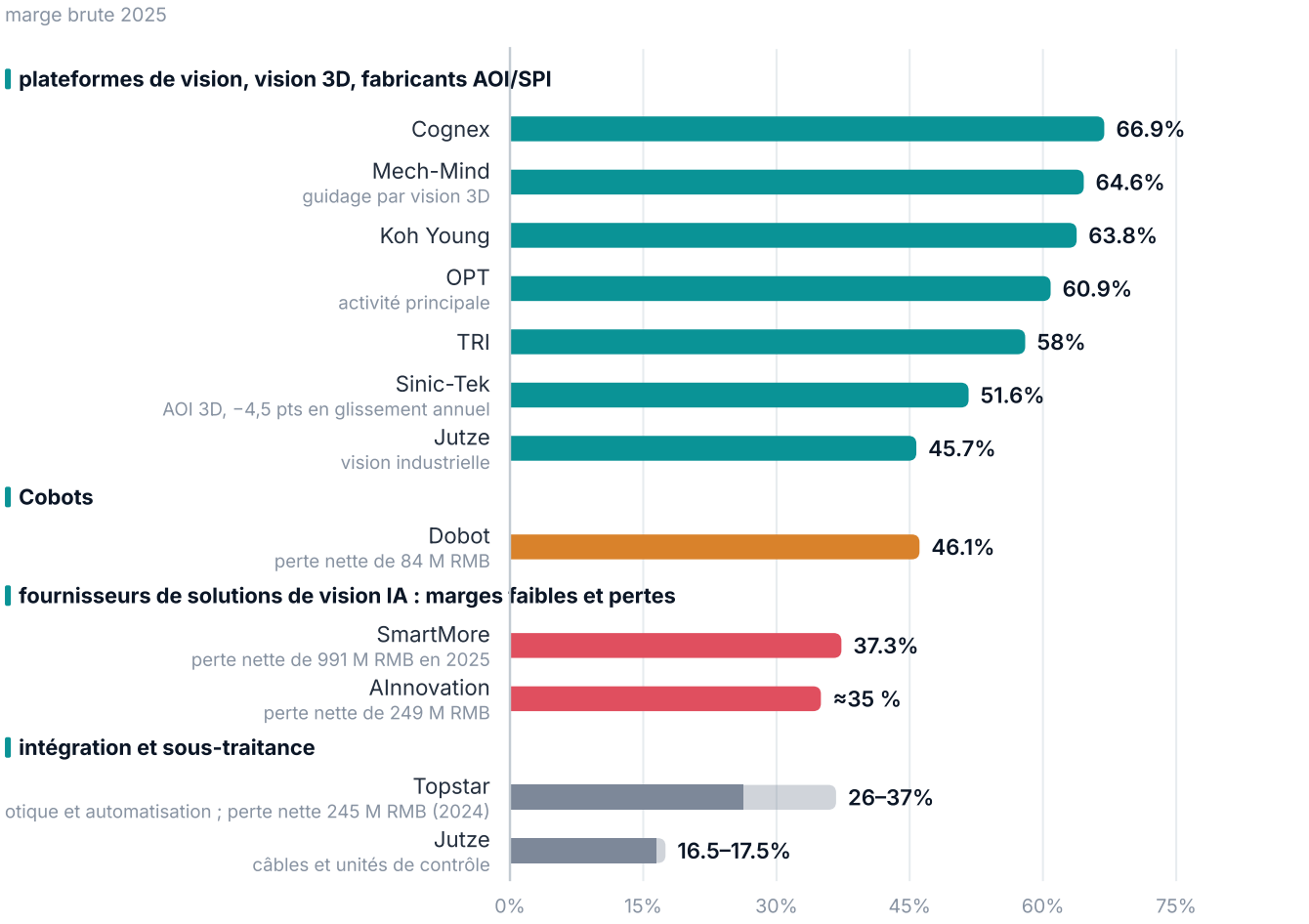
Le comparable coté le plus proche donne un avertissement. Weiyi Zhizao, qui vend des équipements et des robots d'inspection IA, a vu son chiffre d'affaires 3C chuter de 79,9 % sur un an en janvier–mai 2026, a reconnu « une intensification de la concurrence et une pression sur les prix », a reporté son introduction à Hong Kong le 25 octobre 2026, et a divulgué 15 projets terminés avant achèvement, dont 12 projets de robots incarnés (prospectus Weiyi Zhizao, avis de report) single source, primary . AInnovation a terminé 2025 avec 29,8 % de ses créances provisionnées (rapport annuel 2025 d'AInnovation). Le créneau de l'inspection 3C est encombré, et le pouvoir de fixation des prix par ligne est faible.

La marge se situe dans le matériel industrialisé

Position dans la chaîne	Entreprises représentatives et marge brute 2025	Profit et encaissements
Plateformes de vision	Cognex 66,9 %	Rentable ; cédant ≈ 22 M $ de revenus à faible marge
Équipement AOI/SPI	Koh Young 63,8 % ; TRI 58 % ; activité cœur d'OPT 60,85 % ; AOI 3D de Sinic-Tek 51,62 % (−4,53 points sur un an) ; vision industrielle de Jutze 45,74 %	Rentable ; jours de créances de 139 (Jutze) à 247–274 (OPT)
Guidage vision 3D	Mech-Mind 64,6 %	—
Fournisseurs de solutions de vision IA	SmartMore 37,3 % ; AInnovation ≈ 35 %	Déficitaires : perte nette de SmartMore de 991 M RMB en 2025, AInnovation de 249 M RMB
Cobots	Dobot 46,1 %	Perte nette de Dobot de 84 M RMB ; revenus d'UR de 326 M $ (2022) à ≈ 232 M $ (2025, prévision)
Intégration de systèmes	Robotique et automatisation Topstar 26,2–36,7 %	Perte nette de 245 M RMB (2024) ; jours de créances 180–240
Matériel sous contrat	Câbles et unités de contrôle Jutze 16,5–17,5 %	—

Sources : (Cognex, StockAnalysis sur Koh Young, Fugle sur TRI, rapport annuel 2025 d'OPT, rapport annuel 2025 de Sinic-Tek, rapport annuel 2025 de Jutze, Jiemian sur Mech-Mind, 36Kr sur SmartMore, Yicai sur SmartMore, TMTPost, DoNews sur AInnovation, The Paper sur Dobot, The Robot Report sur UR, Jiemian sur Topstar).

FIG 26 **La marge se situe dans le matériel industrialisé** ***Les plateformes de vision et équipementiers réalisent 46–67 % de marge brute ; les fournisseurs de solutions IA 35–37 % et perdent de l'argent.***



Source : Rapports annuels ou prospectus 2025.

Un nouvel entrant ne garde sa marge qu'en transformant l'inspection en produit, équipement ou logiciel ; l'ajustement projet par projet pousse les marges vers environ 35 % avec des pertes persistantes. Les mécaniques jouent aussi contre les petites entreprises : le revenu n'est reconnu qu'après acceptation sur site, les acomptes sont faibles (Jutze, OPT), les pilotes sont généralement gratuits (A3/Mindtrace), et les propriétaires de marques spécifient souvent l'équipement d'inspection sur les lignes de leurs sous-traitants.

Comment les acheteurs paient : 30 % à la signature, 60 % après acceptation, 10 % retenus un an

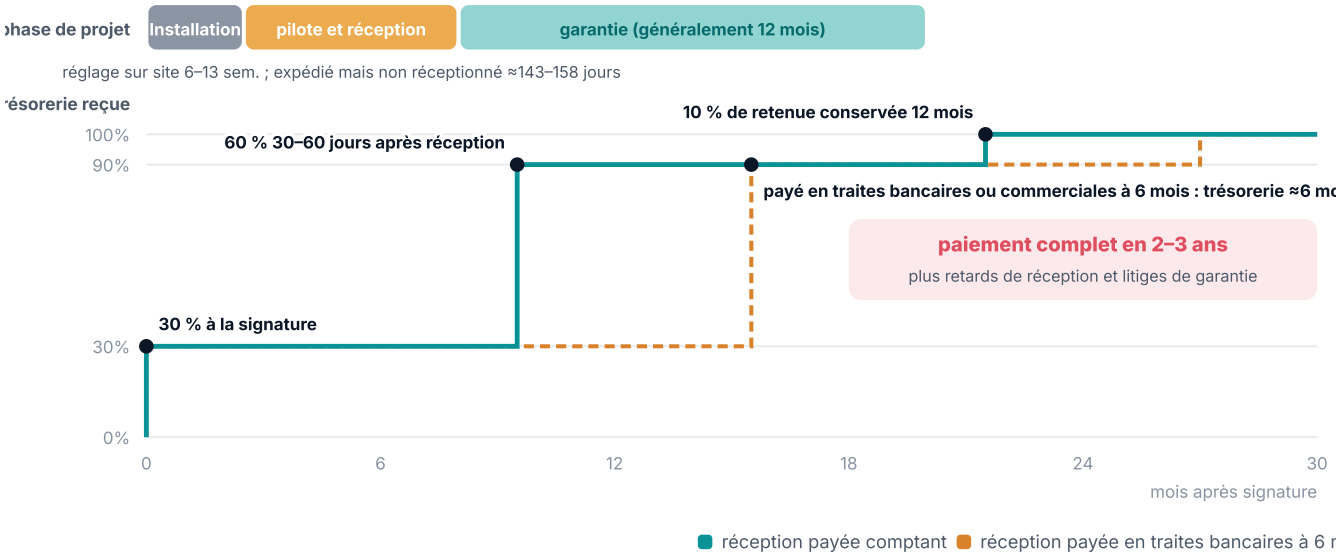
Les conditions de paiement que les usines électroniques continentales utilisent pour l'équipement d'inspection sont détaillées dans 13 contrats d'acheteurs nommés divulgués dans la réponse d'enquête de cotation de Sinic-Tek (réponse d'enquête Sinic-Tek 8-1-1) single source: one primary filing, several independent buyers, 2020–2022 :

Acheteur (année)	Signature	Livraison	Acceptation	Retenue ou solde
Une unité Hikvision (2021)	30 % sous 15 jours	—	65 % sous 30 jours après acceptation finale	5 % de retenue de garantie payés 12 mois après acceptation ; partie payée en billets à ordre bancaires à 6 mois et plus
Nanchang Huaqin (2020–2021)	30 % sous 10 jours ouvrés	—	60 % à 60 jours après acceptation	10 % après une garantie d'un an

Acheteur (année)	Signature	Livraison	Acceptation	Retenue ou solde
Guangdong Yinuo (2021, AOI 3D)	30 % sous 10 jours ouvrés	—	60 % sous 30 jours	10 % après trois mois de fonctionnement sans panne
Xiamen Hongxin, Jingmen Hongyi (2022)	30 % sous 10 jours ouvrés	—	60 % sous 30 jours ouvrés après acceptation finale, billets de banque à 6 mois	10 % après une garantie d'un an
AKM Panyu, Suzhou (2021–2022)	30 % sous 7–10 jours ouvrés	—	60 % un mois après acceptation et facturation, billets de banque à 180 jours	10 % après 360 jours
BYD Precision, Shenzhen, Huizhou (2020)	—	30 % à l'arrivée	70 % après acceptation et facturation	Tout en billets commerciaux à 6 mois
Shenzhen H&T (2020)	—	—	90 % après acceptation et facturation	10 % après six mois d'utilisation sans panne

D'autres documents primaires montrent la même structure : les grands projets de Mech-Mind paient 30 % à la signature, 30 % avant livraison, 30 % après acceptation et 10 % de retenue de garantie sur 12–15 mois (prospectus Mech-Mind) ; GKG dit que la retenue dans le secteur de l'équipement est généralement de 5–10 % du contrat, payée un an après acceptation (réponse d'enquête GKG) single sources, primary. La norme est donc d'environ 30 % à la signature (20–30 %, zéro dans certaines usines comme BYD), 60–65 % après acceptation (payés 30–60 jours plus tard, souvent en billets de banque ou commerciaux à six mois, ce qui retarde l'encaisse utilisable d'un semestre de plus), et 5–10 % de retenue sur 3–12 mois (12 le plus courant) verified across several contracts.

FIG 27 **Comment les acheteurs paient *30 % à la signature, 60 % après réception, 10 % retenus un an : le paiement complet prend 2–3 ans.***



Source : Réponse d'enquête Sinic-Tek, Mech-Mind, GKG, JT et autres dépôts ; schéma sur conditions usuelles, calendrier estimé.

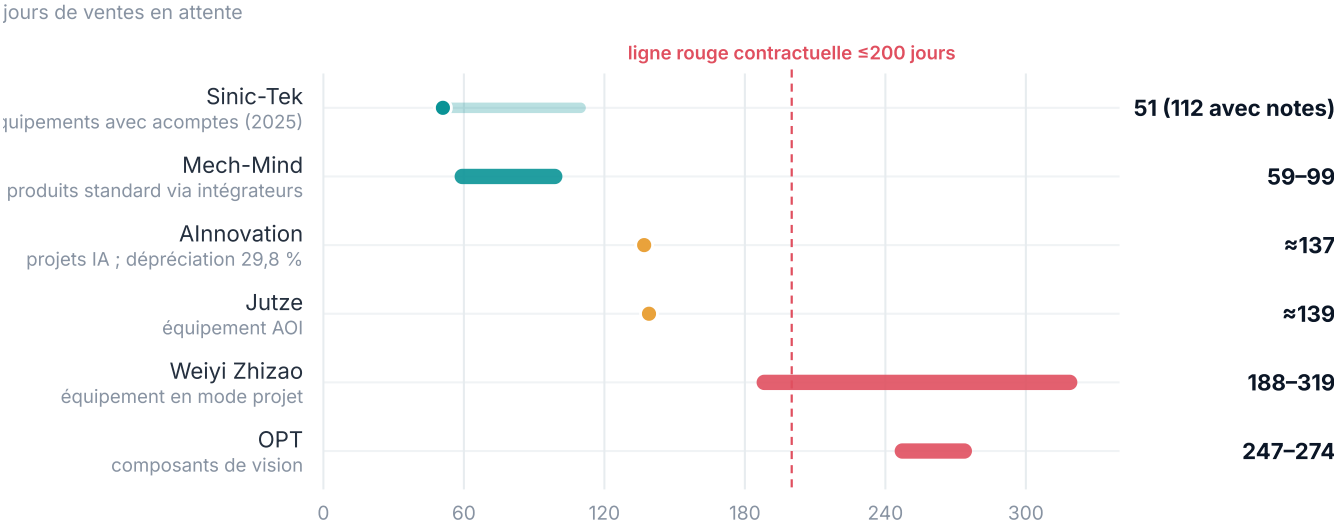
L'acceptation et les essais sont lents et payés par le vendeur. Les marchandises expédiées mais non acceptées de Sinic-Tek ont mis environ 143–158 jours à tourner, parce que les grands clients EMS et marques exigent des tests d'intégration de ligne entière et un essai prolongé (réponse d'enquête Sinic-Tek 8-1-2) ; les projets IA de Weiyi prennent 4–8 mois de l'évaluation au déploiement (prospectus Weiyi Zhizao) single sources, primary. Il n'y a pas de données de conversion pour les pilotes de revue IA ; Sinic-Tek n'a converti qu'environ 18–19 % de ses machines de

démonstration directement en ventes. Je prévois une conversion d'un sur trois à un sur cinq pour les pilotes gratuits et d'environ 80 % pour les pilotes payants après présélection estimate.

Entreprise	Modèle d'affaires	Jours de créances clients
Sinic-Tek (2025)	Équipement avec acomptes	≈ 51 (≈ 112 billets compris)
Mech-Mind	Produits standard via intégrateurs	59–99
Jutze	Équipement AOI	≈ 139
AInnovation (2025)	Projets IA	≈ 137 en fin d'année, avec 29,8 % de dépréciation
Weiyi Zhizao	Équipement par projet	188–319
OPT	Composants de vision	247–274

Sources : les rapports annuels et prospectus ci-dessus ; créances (actifs de contrat compris) ÷ revenu × 365 primary + estimate.

FIG 28 **Jours de ventes en attente*Produits standard 51–99 jours ; équipement en mode projet 188–319 jours.***



Source : Dépôts d'entreprise ; créances (y c. actifs contractuels) ÷ revenu × 365.

Les licences de revue IA vendues via des fabricants d'AOI ou des intégrateurs devraient encaisser comme les 60–100 jours de Mech-Mind ; les ventes de projet directes comme les 190–320 jours de Weiyi. Rassemblés en chronologie de trésorerie : pilote gratuit un à trois mois, qualification chez un grand ODM jusqu'à environ douze mois, construction, livraison et acceptation trois à sept mois ; 30 % arrivent à la commande, 60 % environ quatre à neuf mois après et les 10 % restants environ 16 à 21 mois après. **Du premier pilote chez un grand EMS au paiement intégral, il peut s'écouler deux à trois ans** ; les EMS de taille moyenne sans appels d'offres formels peuvent raccourcir la qualification à des semaines estimate. Les critères d'acceptation des contrats divulgués ne couvrent que l'intégration de ligne entière, le débit (un projet taïwanais exigeait 100 unités par heure) et des mois de fonctionnement sans panne ; aucun ne précise les taux d'échappement ou de fausses alertes (réponse d'enquête Anda) single source, primary. Les critères d'acceptation écrits pour la revue IA doivent être convenus avant le pilote.

Quatre schémas d'échec

Les échecs documentés tombent dans quatre schémas : expédier avant que le produit soit fiable (Rethink (The Robot Report)) ; une couche logicielle sur le matériel d'un leader que l'écosystème du leader a ensuite remplacée (READY Robotics (The Robot Report)) ; manquer d'argent en bas de cycle (Dextrous Robotics (Robots & Startups)) ;

et des entreprises de modèles finissant en acqui-hires (Covariant (GeekWire)). Les leçons : industrialiser de manière ciblée plutôt qu'ajuster projet par projet ; détenir le fonds de roulement pour des revenus basés sur l'acceptation et de longs encaissements ; fournir un service local pour les usines qui tournent 24 heures ; et ne pas passer à l'échelle avant que le produit soit prouvé.

Calendrier : l'inspection ne gagne rien à attendre ; le matériel des bras devient moins cher

Les arguments pour attendre sont faibles. Le matériel n'a pas baissé en 2026 : les fabricants de robots ont augmenté leurs prix de 5–15 %, certains produits de 25–50 % (Sina Finance) single source, et les kits NVIDIA ont augmenté une troisième fois (36Kr). Les VLA de niveau production n'arriveront pas dans la fenêtre (IFR, Instrumental). Pendant ce temps, chaque remplacement d'AOI supprime une cible de rétrofit et chaque déploiement d'IA OEM verrouille les données et le processus d'une ligne ; la capacité des modèles se banalise, et ce qui est rare, c'est l'accès aux usines et aux données d'usine étiquetées. Le support local a aussi des fenêtres : les règles « ville pionnière de l'IA » 2026 de Shenzhen, avec jusqu'à 2 millions de RMB pour les projets de démonstration d'application IA et des bons de modèles couvrant 30 % des achats, courent jusqu'au **30 juin 2027** (Yicai, mai 2026) terms single source.

Ce qui devient encore moins cher rapidement, ce sont les humanoïdes et les mains dextres (les mains sont passées d'environ 50 000 RMB à moins de 10 000 en deux ans (Sina Finance)), et les cobots d'entrée de gamme à mesure que le mix produit descend ; cela plaide pour différer la partie incarnée, pas la partie inspection. **En résumé : inspection maintenant, en produit ciblé ; cellules robotisées maintenant, via des partenaires et l'intégration ; IA incarnée plus tard, avec seulement un poste de R&D aujourd'hui.**

IA incarnée

06 IA incarnée : d'abord elle change la manière d'enseigner aux robots, plus tard leur manière de bouger

Ce qui change, c'est la perception et la programmation ; le mouvement reste déterministe

Chaque leader de bras robotisés commercialise désormais l'« IA physique », mais ce qui a été livré se répartit en trois couches : la perception apprise sans enseignement (AI Masterless Vision de FANUC, livrée en janvier 2027 (FANUC) ; la sélection neuronale de préhension de KUKA (KUKA)) ; les grands modèles qui écrivent des programmes et planifient des tâches (iiQWorks.Copilot de KUKA, qui avertit que le code généré par IA doit être vérifié dans un jumeau numérique (KUKA) ; Yaskawa avec Gemini Robotics-ER, où « l'IA générative décide quoi faire » et le robot décide « comment bouger » (Yaskawa)) ; et les plateformes de calcul et de données (l'AI Accelerator d'UR avec un Jetson Orin (Universal Robots)) vendor claims. Aucun leader ne publie de taux de réussite en production pour ses fonctionnalités IA, et les leaders japonais s'associent à Google, SoftBank, Microsoft et d'autres plutôt que de construire des VLA frontières (index d'actualités FANUC) : la couche de politique apprise viendra des entreprises de modèles, les bras et contrôleurs certifiés des fabricants de bras.

FIG 29 **La pile d'IA incarnée** ***Les grands modèles planifient hors ligne, les petites politiques s'exécutent dans la cellule, le mouvement et la sécurité restent déterministes.***

Couche	Ce qu'il fait, et jusqu'où va la preuve	Échelle de temps · qui la maîtrise
Planification L4	**modèles de pointe (hors ligne)** écrire ou modifier les programmes robot hors ligne, changements de série, expliquer les pannes réécrit les programmes ABB RAPID à 77–99,7 %, déployé après simulation boucle de contrôle à seulement 0,2–0,4 Hz : reste hors temps réel	**secondes à minutes** géants des modèles
Compétences / poli L3	**ACT, politiques de diffusion, petits VLA** postes structurés avec montage 90–99,4 % puces edge : Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**blocs d'action 100–300 ms** fabricants de puces · fabricant
contrôle de mouver L2	**Contrôleur robot** trajectoires, asservissement, contrôle de force, détection de collision code déterministe ; les politiques apprises ne fixent que les objectifs	**millisecondes (le temps réel exige >83 Hz)** fabricants de robots
Sécurité L1	**PLC de sécurité, scanners, limites de force et de vitesse** certifié selon ISO 10218-1/-2:2025 et les règles locales Pas de ML dans les fonctions de sécurité : évaluation par organisme notifié UE à partir du 20 jan 2027	**temps réel strict** robot certifié + intégrateur

Les petites entreprises gagnent dans l'intégration et le matériel de détection : intégrer la perception apprise et de petites politiques dans des cell

Source : Synthèse de l'auteur (chapitres 3 et 6) ; ISO 10218:2025 ; règlement machines UE.

À quelle distance du travail non structuré, flexible et en petites séries : postes structurés dans 1–3 ans, assemblage flexible 5+

De vrais chiffres d'usine commencent à apparaître, mais ils sont rapportés par les vendeurs, étroits et issus de postes structurés. Les huit humanoïdes d'AgiBot sur la ligne de test final des tablettes de Longcheer ont enregistré 64 828 tâches à 99,99 % sur plus de 64 heures (AgiBot, Yicai Global) vendor and customer, mais c'est un travail d'alimentation qu'un bras conventionnel avec vision peut faire. UBTech dit que son Walker S2 travaille à 30–50 % de l'efficacité humaine (Irish Times/FT) single source. Les 11 mois du Figure 02 chez BMW Spartanburg ont produit des chiffres de volume (1 250+ heures, 90 000+ pièces), tandis que ses chiffres de cycle, de succès et d'intervention sont des cibles, pas des résultats (Figure, BMW Group) verified, numerators only. Nucleus dit qu'environ 40 % de sa vidéo d'usine est téléopérée (Interesting Engineering) vendor claim, et une enquête de Reuters a cité environ 20 % de la production d'un ouvrier pour certains humanoïdes chinois (Notebookcheck) single source. Aucun déploiement ne publie ensemble taux de réussite, temps de cycle et taux d'intervention, et aucun n'a été audité, pourtant ces trois-là sont exactement ce qu'un calcul de retour sur investissement exige.

Déploiement	Taux de réussite	Takt	Interventions ou téléopération	Disponibilité ou intervalle de panne	Volume
Figure 02 @ BMW Spartanburg	Cible seulement (> 99 % par équipe)	Cible seulement (84 s)	Cible seulement (0 par équipe)	Non divulgué (point de défaillance haut de l'avant-bras)	1 250 h
Agility Digit @ GXO	Non divulgué	Non divulgué	Non divulgué	Non divulgué	100 000 bacs
Apptronik Apollo @ Mercedes-Benz	Non divulgué	Non divulgué	Non divulgué	Non divulgué	Non divulgué
Agile ONE @ Audi	Non divulgué	Non divulgué	Non divulgué	Non divulgué	Non divulgué
Nucleus (usine non nommée)	Non divulgué	Non divulgué	≈ 40 % téléopéré (méthode non divulguée)	Non divulgué	≈ 2 h de vidéo
AgiBot G2 @ Longcheer	99,99 % de 64 828 tâches	18–20 s par étape	Non divulgué	Une panne de communication, ≈ 2 min pour récupérer	64+ h
Cellule hybride Neuromeka	322/324	12,8 % plus lente qu'une personne	Non divulgué	Non divulgué	324 cycles
Physical Intelligence π0.6 sur site client	—	—	96,4 % autonome sur une équipe	Non divulgué	Une équipe

Sources : ce chapitre et le chapitre 3 mostly vendor claims.

FIG 30 **Ce que divulguent réellement les déploiements robot en usine** ***La plupart ne divulguent que le volume ; aucun ne divulgue ensemble taux de succès, takt et taux d'intervention.***

	taux de succès	Takt	Interventions / téléopération	Disponibilité / pannes	Volume
Figure 02 @ BMW	objectif seul >99 %/équipe	objectif seul 84 s	objectif seul 0/équipe	n.d.	1 250 h
Agility Digit @ GXO	n.d.	n.d.	n.d.	n.d.	100 000 bacs
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 (usine non nommée)	n.d.	n.d.	≈40 % téléopéré*	n.d.	≈2 h de vidéo
AgiBot G2 @ Longcheer	99.99%	18–20 s par étape	n.d.	une panne, 2 min	64+ h
cellule hybride Neuromeka	322/324	12,8 % plus lent	n.d.	n.d.	324 cycles
π0.6 @ site client	—	—	96,4 % autonome	n.d.	1 équipe

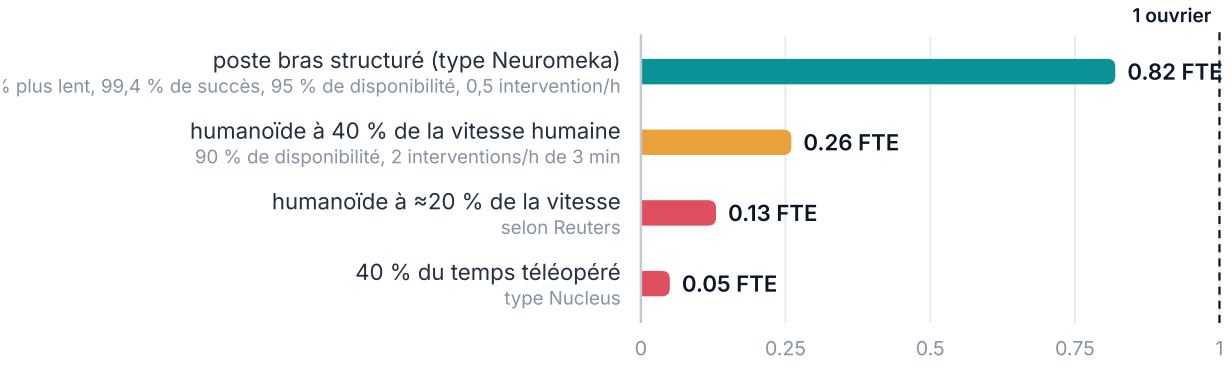
* méthode non divulguée

Divulgué (surtout allégations fournisseur) objectifs seuls n.d

Source : Supports publics de BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence et autres.

La téléopération compte plus pour l'économie que le taux de réussite. Personnes nettes remplacées par robot et par équipe ≈ (cycle humain ÷ cycle robot) × disponibilité × succès, moins le temps d'intervention et la téléopération estimate, my formula. Un poste structuré de type Neuromeka (12,8 % plus lent, 99,4 % de succès, 95 % de disponibilité, 0,5 intervention de 2 minutes par heure) remplace environ 0,82 personne ; un humanoïde à 40 % de la vitesse humaine, 90 % de disponibilité et deux interventions de 3 minutes par heure environ 0,26 ; aux 20 % de Reuters, environ 0,13 ; et si, comme Nucleus, un opérateur le téléopère 40 % du temps, environ 0,05, presque rien. À environ 600 k RMB par humanoïde et 85 k RMB par personne et par an sur le continent, les trois derniers se remboursent sur des décennies ; seul le poste à bras structuré a un retour réaliste (chapitre 4) estimate.

FIG 31 **Main-d'œuvre nette remplacée par robot** ***Une cellule à bras structuré remplace ≈0,82 travailleur ; un humanoïde téléopéré 40 % du temps, ≈0,05.***



Source : Estimations basées sur les formules de l'auteur.

Hors des postes structurés, les modèles généraux obtiennent encore 40–90 % sur de nouvelles tâches : Gemini Robotics 2 atteint 74,2 % sur le pick-and-place en laboratoire, 78,9 % sur le kitting d'outils et 44 % sur le nouage d'un sac (eesel/DeepMind) ; sur RoboChallenge Table30, 30 tâches sur robot réel, π0.5 obtient 43,7 % et le meilleur modèle ouvert 50,3 % (RoboChallenge, QbitAI).

Les prévisions extérieures diffèrent, mais aucune ne soutient un usage généralisé sous deux ans. L'IFR dit que l'IA physique générale n'est « pas encore de niveau industriel » et que la plupart des humanoïdes sont « encore des

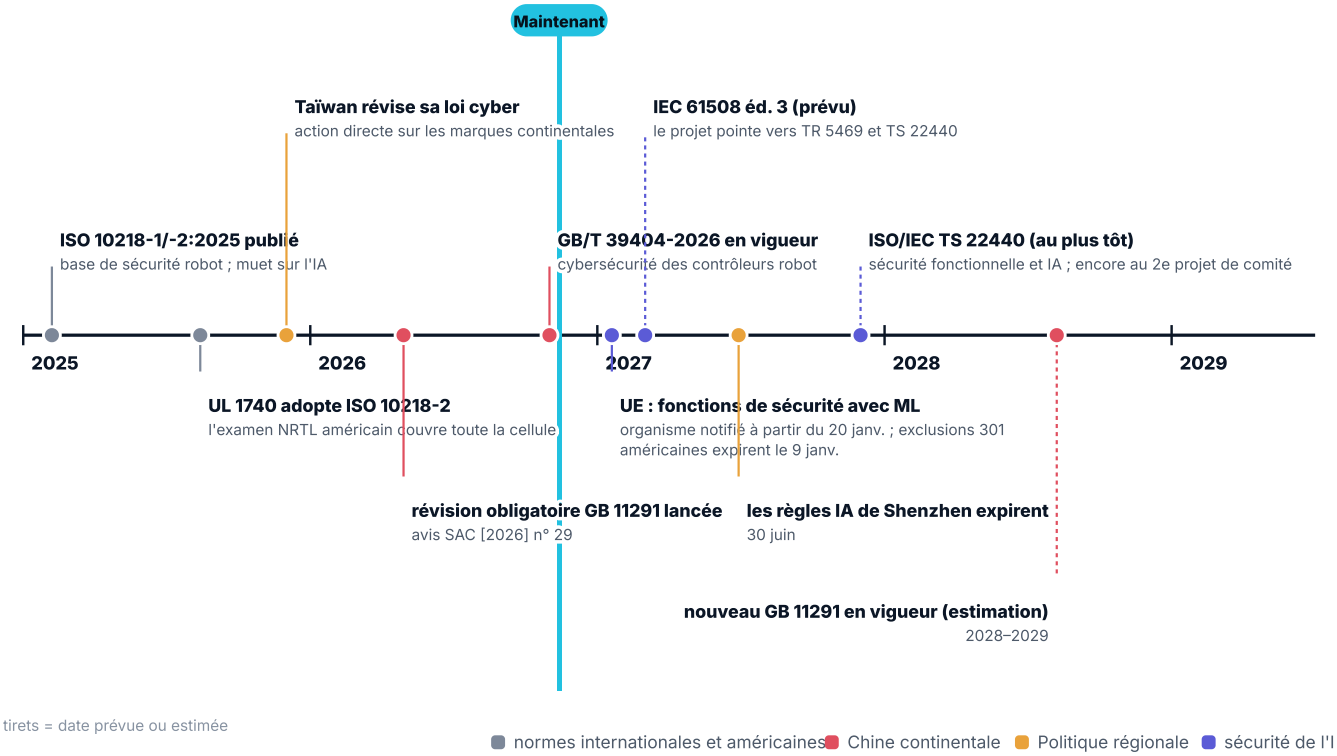
prototypes ou des pilotes » (IFR) ; Interact Analysis estime que seulement environ 10 % des humanoïdes produits en 2025 sont entrés en usage réel, le point de bascule commercial étant « après 2032 » (automation.com) ; UBS voit un « moment EV » des humanoïdes improbable avant 2030 (Yicai Global) ; des initiés chinois estiment encore deux à trois ans avant l'usage à l'échelle de l'usine (Gasgoo) single sources . Mon jugement estimate : dans les usines électroniques et automobiles bien dotées avec le vendeur sur site, les **tâches de poste structuré** (alimentation, tri OK/NG, insertion fixée) atteindront ≥ 99 % à une vitesse proche de l'humain dans environ 1–3 ans ; le **kitting de pièces mixtes** avec des VLA généraux dans environ 3–5 ; les **faisceaux à routage libre et l'assemblage flexible multi-étapes** dans plus de 5. Le retour sur investissement des humanoïdes ne fonctionne pas aujourd'hui : un Walker S2 à environ 600 k RMB (10jqka) single source , à 30–50 % de l'efficacité humaine sur deux équipes, équivaut à environ 0,6–1,0 ouvrier, soit un retour simple de 7–12 ans aux coûts de main-d'œuvre continentaux avant intégration, maintenance et temps d'arrêt estimate .

Là où ça bloque

Maillon	Preuve en 2026
Données	Le secteur estime qu'au moins 10 millions d'heures de données réelles de haute qualité sont nécessaires et que seulement environ 500 000 avaient été collectées début 2026 (Gasgoo) ; Reuters rapporte environ 500 000 heures contre environ 100 millions nécessaires (Notebookcheck) ; un téléopérateur novice a besoin d'environ 300 tentatives par mouvement utilisable ; les usines ne partageront pas les données de processus ; les politiques entraînées en simulation chutent fortement sur les vrais robots
Fiabilité	Postes structurés 98–99,99 % (vendor claims, échec non défini) ; modèles généraux 40–90 % sur de nouvelles tâches ; connecteurs inédits 29–67 %
Vitesse	Humanoïdes à 30–50 % de la vitesse humaine (UBTech), certains à environ 20 % (Reuters) ; agents frontières environ 280 s par tâche quasi statique sur robot réel ; boucles de contrôle LLM à 0,2–0,4 Hz contre environ 83 Hz nécessaires (Anthropic)
Coût	Walker S2 en moyenne environ 600 k RMB ; retour de 7–12 ans estimate
Certification de sécurité	ISO 10218-1/-2:2025 rend la sécurité fonctionnelle plus explicite et absorbe les exigences collaboratives, mais ne dit rien sur l'IA (The Robot Report) ; ISO/IEC TS 22440 (sécurité fonctionnelle et IA) était encore un second projet de comité en novembre 2026 (ISS, ISO Open Data) ; le guide IEC 61508-3 déconseille l'IA au-dessus de SIL 1 (Analog Devices) ; sous le règlement européen sur les machines, les fonctions de sécurité avec apprentissage automatique exigent une évaluation par organisme notifié à partir du 20 janvier 2027 (EUR-Lex)
Mains et outils	Les mains dextres gèrent environ 30 N (≈ 3 kg) par doigt contre environ 10 kg pour une main humaine (Gasgoo) ; les changements d'outils restent manuels
Intégration	Lectures de capteurs erronées, incompatibilités de protocoles et liaisons MES sur site ; taux d'intervention presque jamais divulgués

L'implication de sécurité est directe : les normes n'interdisent ni ne reconnaissent les politiques apprises. ISO 10218-1/-2:2025 ne dit rien sur l'IA ; ISO/IEC TS 22440 (sécurité fonctionnelle et IA) en est encore au projet de comité et peu probable avant fin 2027 ou 2028 estimate ; ISO/IEC TR 5469:2024 n'est qu'un guide, recommandant des replis non IA, des contraintes de supervision et un vote redondant ; l'IEC 61508 édition 3 est attendue début 2027 (ISO Open Data, BSI, 61508 Association) verified . Aucune norme publiée n'offre encore de voie acceptée pour certifier la perception par apprentissage automatique comme fonction de sécurité. Les règles régionales bougent aussi : la Chine révise la GB 11291 (en vigueur vers 2028–2029) (plan de révision 20262543) ; UL 1740 a intégré ISO 10218-2 à l'examen NRTL américain depuis août 2025 ; l'UE exige l'évaluation par organisme notifié des fonctions de sécurité avec apprentissage automatique à partir du 20 janvier 2027 ; et aucun humanoïde n'est certifié pour l'usage en usine sans clôture tant qu'ISO 25785-1 est encore en projet (i-SCOOP). L'architecture défendable : l'apprentissage automatique ne prend que des décisions de processus (« quelle pièce », « cette carte est-elle bonne »), tandis qu'une couche de sécurité déterministe certifiée (généralement PL d / Catégorie 3) et un automate de sécurité décident de la vitesse, de la force et des arrêts. Traitez toute revendication d'« IA caméra pour la sécurité sans clôture » comme non prouvée jusqu'à ce qu'un certificat PL d ou SIL 2 nomme le composant d'apprentissage automatique.

FIG 32 **Chronologie des normes et de la réglementation** ***Une voie de certification pour l'IA dans les fonctions de sécurité ne se stabilisera pas avant 2027–2029.***



Source : ISO, IEC, UL, EUR-Lex, SAC et bases de données réglementaires nationales.

Quelle couche est l'opportunité : les modèles appartiennent aux géants, les boîtiers edge aux fabricants de puces, les petites entreprises à l'intégration et au matériel de détection

La couche des modèles exige du capital et des données et appartient aux géants (AgiBot a entraîné GO-2 sur des « dizaines de milliers d'heures » (iFeng Tech) ; les modèles les plus forts sont fermés ou cloud-only). La couche des boîtiers edge appartient aux fabricants de puces : Jetson Thor est la norme de fait, et sur le continent les cartes RDK de D-Robotics (D-Robotics) et les coprocesseurs Rockchip (CSDN) rivalisent, surtout sans résultats VLA publiés.

Les petites entreprises ont deux ouvertures. L'une est le matériel de détection qui achète directement de la fiabilité : le toucher fait passer le succès d'insertion de ContactDP de 58,7 % à 97,8 %, et les capteurs de force/couple et les caméras au poignet font un travail similaire ; les kits de collecte de données se vendent aussi (10jqka), bien que mains, capteurs et bras baissent tous de prix. L'autre est la couche d'intégration entre les fabricants de bras et les entreprises de modèles, en possédant les données de tâche, l'enveloppe de sécurité et la liaison MES de chaque poste, ce que les fabricants de bras ne vendent pas comme produit aujourd'hui. Les concurrents se répartissent en quatre groupes : les leaders des bras (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot et autres), les entreprises vision-plus-bras (Mech-Mind, Hikrobot, SmartMore), les entreprises d'IA incarnée (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) et les entrants des puces automobiles.

Acceptation par l'acheteur : pas de chiffre sans dénominateur

Comme aucun vendeur ne publie de taux d'intervention, d'intervalles de panne ou de décomptes de succès à long terme, ceux-ci appartiennent aux conditions d'acceptation du contrat. Statistiquement, 3 000 cycles de production consécutifs sans échec prouvent un taux d'échec d'au plus 0,1 % à 95 % de confiance ; 10 échecs sur 3 000 ne prouvent que ≤ 0,57 % ; prouver ≤ 1 %, ≤ 0,5 %, ≤ 0,1 % et ≤ 0,01 % sans échec exige 299, 598, 2 995 et 29 956 cycles `estimate, exact binomial`. À cette aune, les 322/324 de Neuromeka ont un intervalle bilatéral à 95 % de 97,8–

99,9 % et ne prouvent pas ≥ 99 % ; les quelque 6 échecs d'AgiBot sur 64 828 tâches impliquent 99,98–100 %, mais seulement si « échec » est défini comme l'acheteur le définit estimate. Voici les exigences d'acceptation que j'imposerais à toute cellule robotisée ou incarnée ; les seuils s'appuient sur les propres cibles de Figure chez BMW et sur les cas Neuromeka et AgiBot, et relèvent du jugement, pas d'une norme estimate :

Élément	Exiger	Seuil de passage suggéré
Définitions	Définitions signées du cycle, du succès, de l'échec (succès partiel, reprise de préhension, chute, échappement qualité), de l'intervention (toucher, téléopération à distance, réinitialisation, arrêt d'urgence) et du calcul de l'autonomie	Les définitions précèdent toute métrique ; pas de pourcentage sans formule
Interventions et téléopération	Par heure-robot, par type et durée ; temps moyen entre interventions ; minutes de téléopération en part du temps de fonctionnement ; opérateurs par robot	Postes structurés ≤ 0,5 intervention par heure-robot (≥ 2 h d'écart) ; zéro téléopération pendant l'acceptation
Succès à long terme	≥ 3 000 cycles de production consécutifs sur ≥ 3 équipes, ≥ 2 références et une variation normale d'éclairage et d'approvisionnement ; échecs classifiés ; borne inférieure exacte	Borne inférieure ≥ 99,5 % pour les postes en ligne arrêtant la ligne, ≥ 99,0 % hors ligne ; échappements qualité rapportés séparément
Temps de cycle	P50/P90/P99 par pièce, récupération comprise, contre ≥ 30 cycles humains chronométrés au même poste	≤ 1,15 × P50 humain, ou dans le takt de la ligne
Disponibilité et pannes	Heures de fonctionnement, journal d'arrêts, pannes par composant, temps d'échange de pièce	≥ 95 % de disponibilité sur ≥ 2 semaines ; MTTR ≤ 20 min avec pièces détachées sur site
Récupération	Injection de pannes scriptée : chutes, occlusion, changement d'éclairage, arrêt de convoyeur, perte MES ou réseau, coupure de courant, arrêt d'urgence	≥ 90 % de récupération autonome, retour en production sous 2 min
Changement de série	Heures, démonstrations, compétence opérateur et coût de collecte de données par nouvelle référence ou montage	Postes structurés : ≤ 1 équipe et ≤ 100 démos par nouvelle référence
Sécurité	Bras et cellule selon ISO 10218-1/-2:2025, évaluation des risques de la cellule, règles locales (GB 11291 sur le continent ; CNS 14490 plus évaluation écrite à Taïwan ; ANSI/A3 R15.06-2025 et UL 1740 aux États-Unis)	Aucune fonction de sécurité mise en œuvre par apprentissage automatique ; une couche de sécurité certifiée (généralement PL d / Cat 3) limite la vitesse et la force ; pas de robots à pattes sans clôture jusqu'à publication et évaluation d'ISO 25785-1
Calcul et comportement hors ligne	Matériel d'inférence, latence P99, comportement sans réseau, versionnage des modèles, revalidation après chaque mise à jour	L'inférence critique pour le takt ne dépend pas du cloud ; ré-acceptation complète après tout changement de modèle, de quantification ou de moteur
Preuves	Vidéo non montée et journaux bruts pour toute la fenêtre d'acceptation ; attestés par l'équipe qualité du client ou un tiers	Pas de montages choisis ni de résumés fournis par le vendeur uniquement
Conditions commerciales	Jalons de paiement liés aux métriques d'acceptation ; pénalités de disponibilité lorsqu'elle est facturée par heure-robot ; propriété des données et règles d'accès à distance	Paiement final et retenue libérés seulement après l'acceptation des 3 000 cycles et ≥ 4 semaines de fonctionnement stable

Élément	Exiger	Seuil de passage suggéré
Économie	Personnes nettes remplacées = rapport de vitesse × disponibilité × succès − temps d'intervention − part de téléopération	Rejeter toute proposition dont le remplacement net ne peut soutenir le retour revendiqué

Sources : (Figure, Neuromeka, Yicai Global, i-SCOOP) ; les seuils sont mes suggestions.

Questions ouvertes

07 Ce que les sources publiques ne peuvent répondre, et comment un pilote le règle

Les questions auxquelles les sources publiques ne peuvent répondre se répartissent en cinq groupes : à quel prix la revue IA se vend-elle réellement et ce que les acheteurs paieront ; comment les fabricants d'AOI regroupent l'IA et mettent à niveau les anciennes machines ; quel âge ont les AOI installés et combien font déjà tourner l'IA ; combien de personnes et d'heures la revue prend-elle réellement par ligne ; et comment les pilotes se convertissent et comment les acheteurs paient. Par-dessus, comment les modèles les plus récents, les compilations quantifiées et les bras robotisés se comportent réellement dans les conditions d'une usine donnée. Chacune peut être transformée en un test avec un seuil clair. Les seuils ci-dessous sont des règles de décision que j'ai dérivées des preuves de ce rapport estimate ; recalculez-les dès l'arrivée des premières données d'usine.

Question ouverte	Comment la régler	À quoi ressemble un bon résultat
Prix et disposition à payer pour la revue IA	Un appel d'offres standard (par exemple quatre lignes avec d'anciens AOI 2D et de plus récents AOI 3D mixtes, deux équipes, environ 2 000 cartes par équipe, 12 classes de défauts, sur site uniquement, intégration MES ou IPC-CFX) envoyé à plusieurs fournisseurs tiers ; une enquête de sensibilité au prix en quatre questions auprès de 20–30 usines EMS	La plupart des devis entre 40–120 k RMB ponctuels ou 20–40 k par an pour quatre lignes ; au moins la moitié des usines acceptent 10–30 k RMB par ligne avec un retour sous 12 mois
Regroupement IA des fabricants d'AOI et mises à niveau des anciennes machines	Courts entretiens structurés avec environ une douzaine de fabricants d'AOI et de distributeurs	Si la revue IA est standard, optionnelle ou absente sur les modèles 2025–26 ; si les machines antérieures à 2020 peuvent être mises à niveau et à quel prix
Âge des AOI installés et part faisant tourner l'IA	Enquête sur site d'environ 30 usines EMS, stratifiée par taille et région, avec listes d'équipements examinées sous NDA	Un âge médian des AOI ≥ 5 ans et ≥ 60 % sans IA signifie que le parc de rétrofit est réel
Main-d'œuvre et heures de revue	Entretiens et 2–4 semaines de journaux d'alarmes AOI, de revue, de réparation et d'équipes dans deux usines ou plus, avec échantillonnage du travail	≥ 1 réviseur dédié par ligne et par équipe et ≥ 3 heures de revue par ligne et par équipe, avec au moins 0,5 FTE par ligne que l'usine est prête à redéployer
Performance de la revue IA elle-même	Jeu de référence plus mode fantôme sur une ligne réelle (chapitre 3) : accord sur les données en semaine 0, référence et étiquetage en semaines 1–2, mode fantôme en semaines 2–4, exécution assistée facultative en semaines 4–8	Fausses alertes ou heures de revue réduites de ≥ 50–60 % ; bornes d'échappement par classe pas pires que le niveau humain ; zéro manqué sur les classes critiques ; latence p95 ≤ ½ takt ; bascule automatique vers les personnes en mode ouvert
Qualité d'étiquette-enseignant des modèles frontières	Faire tourner deux ou trois candidats sur le même jeu de référence, trois formulations de prompt chacun, une bonne image de référence, recadrages à composant unique ≥ 42–48 px	Acceptation automatique seulement pour les classes sans manqué sur ≥ 299 défauts et avec accord de niveau humain ; sinon pré-étiquetage et tri uniquement

Question ouverte	Comment la régler	À quoi ressemble un bon résultat
Compilations quantifiées	Ne changer que la précision ; épingler les hachages de poids, le gabarit, le prétraitement, le décodage, le moteur et la taille de lot ; désactiver ou réinitialiser le cache de préfixe ; vérifier que la compilation de référence reproduit les résultats du vendeur ; noter les paires sur le jeu de référence et la tranche de dérive	Aucun nouveau manqué sur les classes critiques ; ≤ 1 nouveau manqué sur 1 000 défauts ; clôture en baisse de ≤ 2 points ; ≥ 99,5 % de décisions identiques sur trois passages ; sinon revenir au FP8/INT8
Matériel côté ligne, propreté et ESD	≥ 2 unités de chaque boîtier candidat testées en contrainte à deux fois le taux d'alarmes attendu ; trempage de 72 h au flux d'air réel de l'armoire ; contrôle ESD par le coordinateur du client ; test ISO 14644-14 pour les clients ISO 7 et plus propre	p95 ≤ ½ takt ; aucun plantage ni bridage sur 72 h ; validation ESD ; rapport de particules à la bonne classe
Conditions de paiement, règles de pilote et conversion	Clauses de paiement de contrats récents d'équipement d'inspection dans les usines enquêtées ; entretiens sur les règles de pilote, les critères d'acceptation et le temps de qualification	Environ 30/60/10 avec encaissement sous 60 jours après acceptation ; pilotes ≤ 8–12 semaines avec un responsable de validation nommé ; ≥ 1 essai sur 3 converti ; un accord-cadre couvrant ≥ 5–10 lignes la première année
Données, informatique et responsabilité	Entretiens et revue de contrat avec un conseil juridique	Un boîtier local peut se connecter et accéder aux données ; responsabilité limitée aux défauts logiciels
Conformité régionale avant de partir à l'étranger	Taïwan : statut écrit du client sous la loi sur la gestion de la cybersécurité, auto-évaluation E187 ; États-Unis : décision douanière sur la classification d'une cellule, devis NRTL, contrôle ITAR ; Asie du Sud-Est : lien local BOI et éligibilité à l'abattement malaisien, maison mère du client ; continent : périmètre CCC des PC et alimentations regroupés	Le client n'est pas un organisme contrôlé, et le retour reste ≤ 2 ans droits et certification compris
Acceptation d'une cellule robotisée ou incarnée	Le tableau d'acceptation en 12 éléments du chapitre 6 : définitions signées, ≥ 3 000 cycles de production consécutifs sur ≥ 3 équipes et ≥ 2 références, injection de pannes scriptée, ≥ 30 cycles humains chronométrés, attestés	Borne inférieure ≥ 99,5 % en ligne (≥ 99,0 % hors ligne) ; cycle ≤ 1,15 × humain ; ≤ 0,5 intervention par heure et zéro téléopération ; ≥ 95 % de disponibilité ; MTTR ≤ 20 min ; ≥ 90 % de récupération
Politiques ACT et VLA sur puces edge chinoises	Quatre semaines : portage et micro-benchmark (≥ 1 000 inférences, distribution de latence, puissance, trempage de 60 min) ; simulation en boucle fermée avec latence mesurée injectée ; A/B sur robot réel contre un GPU ; endurance de 8–24 h (≥ 3 000 cycles)	Blocs ACT p95 ≤ 100 ms, VLA ≤ 300 ms ; succès sur robot réel à 3 points près du GPU et cycle ≤ 1,2 × humain ; sinon garder la politique sur Jetson ou un GPU
Prix des cobots, AMR et de l'intégration	Devis modèle par modèle auprès de plusieurs distributeurs de cobots, et devis de cellule complète auprès de deux ou trois intégrateurs pour le chargement de montage de test (un poste, deux équipes)	Cellule complète ≤ 200 k RMB et retour ≤ 1,5 an sur deux équipes à Shenzhen
Décomptes de lignes SMT à Taïwan, en Asie du Sud-Est et aux États-Unis	Données IPC (anciennement New Venture Research) ou base installée des distributeurs d'équipement locaux	Assez de lignes adressables par région pour justifier l'ingénierie d'application locale

De toutes ces questions, quatre décident des chiffres les plus sensibles de ce rapport : le prix que les acheteurs paieront, la main-d'œuvre que la revue IA libère réellement, si la revue IA ne manque réellement pas plus que les personnes, et à quelle vitesse l'argent revient. Elles valent d'être réglées avant tout engagement significatif, par une usine envisageant un achat ou par un fournisseur envisageant un produit.

Une séquence de pilote qui fonctionne pour les deux côtés

Pour une usine, et pour un vendeur qui veut une référence crédible, l'ordre est le même :

1. **Semaines 0–2 : convenir des données et des définitions.** Signer l'accord sur les données, définir les classes de défauts et ce qui compte comme un échappement, et enregistrer une référence : fausses alertes par carte, secondes de revue par appel, échappements trouvés en aval.
2. **Semaines 1–4 : construire le jeu de référence** depuis les archives AOI et les cartes NG conservées, étiqueté en aveugle et arbitré (chapitre 3).
3. **Semaines 2–4 (ou plus) : mode fantôme.** Le modèle journalise ses décisions pendant que les personnes décident comme avant ; comparer par classe.
4. **Semaines 4–8 : exécution assistée sur des équipes aléatoires,** le modèle ne clôturant que les classes ayant passé le seuil et tout le reste allant aux personnes.
5. **Acceptation, par écrit, avant montée en charge :** les résultats par classe, la latence, le comportement en mode ouvert, et qui possède les modèles et les étiquettes. Monter en charge ligne par ligne, et revalider à chaque nouveau produit, version de modèle, quantification ou changement de moteur.

Limites de ce rapport

Il s'agit d'une recherche documentaire. Je n'ai mené aucun entretien primaire avec des usines pour cette édition ; une grande partie des données de ligne provient de vendeurs et n'a pas été auditée indépendamment. Plusieurs sites gouvernementaux, bases de données d'appels d'offres et plateformes de recrutement n'ont pas pu être consultés, de sorte que les taux de subvention de modernisation 2026 de Shenzhen, la guidance salariale du delta du Yangtsé, les prix de transaction de la revue IA, les réviseurs par ligne et les taux de conversion des pilotes sont des estimations ou des analogies ; chacun apparaît ci-dessus comme une question ouverte. Les taux de change sont le cours pivot de la Banque populaire de Chine du 9 novembre 2026 (1 $ = 6,7188 RMB) et, pour Taïwan, le fixing de SMBC du même jour (1 $ = 32,12 NT$) ; le dong vietnamien utilise des données de mai 2026. Tous les seuils sont des règles de décision dérivées des preuves citées, pas des normes.

À propos de l'auteur

Tan Shuai

Tan Shuai est un technologue et fondateur avec plus de vingt ans de travail dans les puces, le matériel et les logiciels système, et 20 brevets délivrés. Son parcours l'a mené à travers la Silicon Valley, New York, Chicago, Taïwan, la Grande Baie (Shenzhen et Hong Kong), le Vietnam et la Malaisie, avec une attention de longue date portée à la fabrication avancée et aux chaînes d'approvisionnement. Aujourd'hui, il travaille comme chercheur et développeur en IA, en se concentrant sur la manière dont l'IA est réellement utilisée en atelier.

Il accueille volontiers les échanges avec les personnes travaillant sur l'IA dans la fabrication, que ce soit pour monter une équipe, chercher un partenaire ou comparer des notes : i@tanshuai.com · tanshuai.com

Notes et avertissement

Ce rapport est une recherche documentaire indépendante fondée sur des documents publics disponibles jusqu'au 11 novembre 2026. Il est fourni à titre informatif uniquement et ne constitue pas un conseil en investissement, juridique, fiscal ou d'ingénierie. Les noms de produits, les prix, les sorties de modèles, les classifications d'exportation, les droits de douane et les réglementations changent rapidement ; vérifiez tout ce sur quoi vous vous appuyez auprès de la source primaire. Les chiffres marqués vendor claim ou estimate doivent être lus comme tels. Les noms d'entreprises et de produits appartiennent à leurs propriétaires et sont utilisés uniquement à des fins d'identification ; aucune entreprise n'a relu ni sponsorisé ce rapport.

© 2026 Tan Shuai. Vous pouvez partager ce rapport sans modification avec attribution.

FABRICATION ÉLECTRONIQUE · INSPECTION PAR IA × ROBOTIQUE

Inspection visuelle par IA et bras robotisés dans les usines électroniques : évaluation 2026

Chaque chiffre clé porte sa source et une balise de niveau de confiance : verified, single source, vendor claim ou estimate. Ce que les sources publiques ne permettent pas de trancher est rassemblé au chapitre 7, avec la manière dont un pilote règle chaque question.

Tan Shuai
tanshuai.comi@tanshuai.com

ÉDITION PUBLIQUE · DONNÉES AU 11 NOVEMBRE 2026

