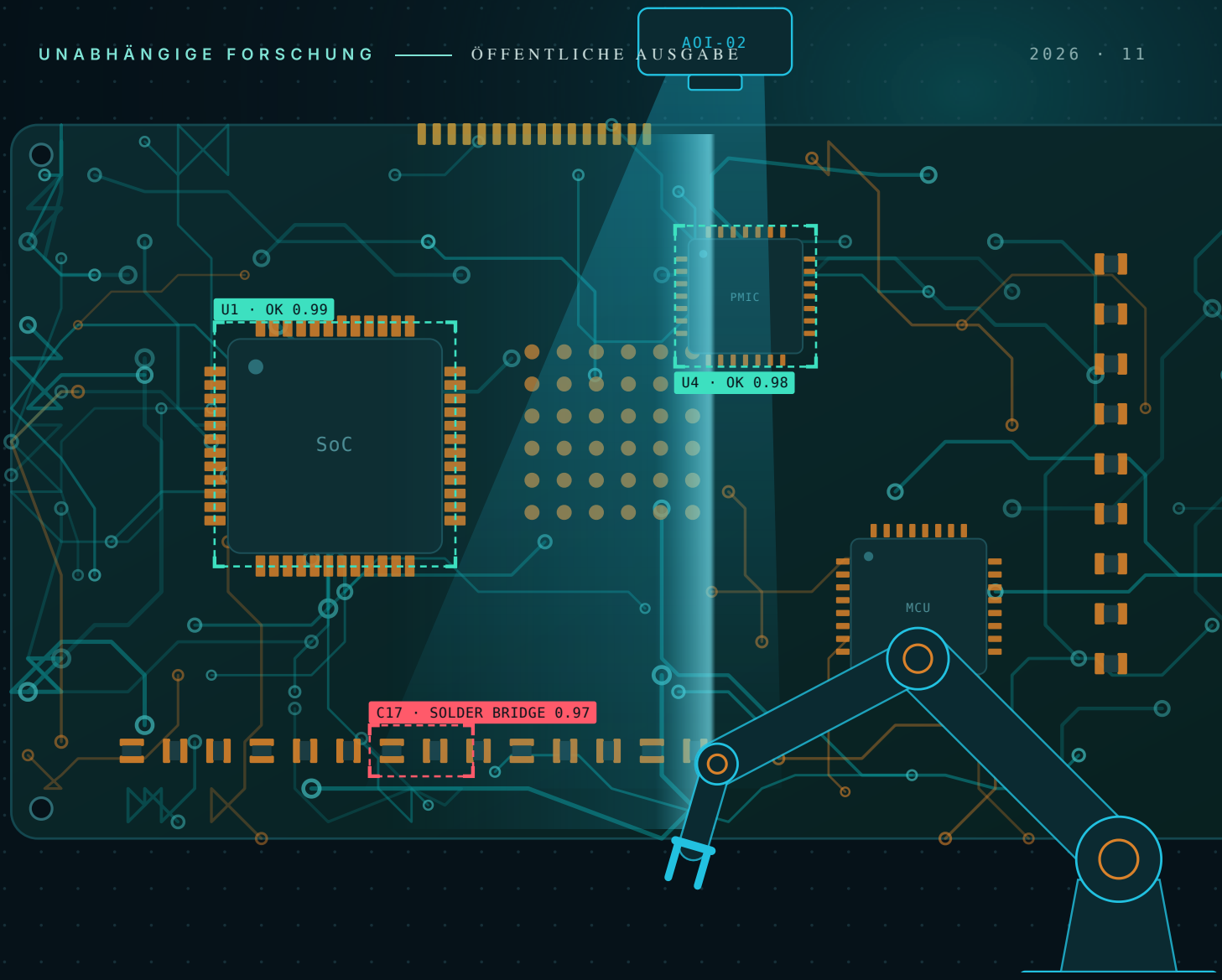


ELEKTRONIKFERTIGUNG · AI-INSPEKTION × ROBOTIK

AI-Sichtprüfung und Roboterarme in Elektronikfabriken

Bewertung 2026

Gängige Konfigurationen · Regionale Machbarkeit · Bereitstellung vor Ort · Lehrer-Schüler-Modelle
Return on Investment · Wettbewerb · Was nur ein Pilotprojekt beantworten kann

KERNERGEBNIS

AI-Nachprüfung lohnt sich, in engem Rahmen. Verkörperte Roboterarme können warten.

Tan Shuai
tanshuai.comi@tanshuai.com

Datenstand: 11. November 2026
Festlandchina · Taiwan · Südostasien · Vereinigte Staaten

INHALT

Inhalt



◆ **ZUSAMMENFASSUNG** 5
Lohnend, aber eng gefasst, jetzt und mit Partnern

01 **Die Linie heute** 9
Was Elektronikfabriken heute betreiben und was technisch machbar ist

02 **Vor Ort** 17
AI im Werk betreiben: Reinräume, Luftspalte, Hardware und Quantisierung

03 **Lehrer und Schüler** 27
Lehrer und Schüler: Frontier-Modelle können Inspektionsdaten vorab beschriften; bei Roboterarmen sind die wahren Lehrer weiterhin Menschen

04 **Kosten und Amortisation** 36
Kosten und Amortisation: AI-Nachprüfung amortisiert sich in Monaten, Maschinenbeschickung in Jahren, Ausschleusarme nie

05 **Markt und Wettbewerb** 44
Markt, Wettbewerb und wie Käufer zahlen

06 **Verkörperte AI** 53
Verkörperte AI: Zuerst ändert sie, wie Roboter unterrichtet werden, später wie sie sich bewegen

07 **Offene Fragen** 59
Was die öffentliche Dokumentation nicht beantworten kann und wie ein Pilot es klärt

◆ **Über den Autor** 62
Tan Shuai

◆ **Über den Autor** 63
Tan Shuai

Vertrauensgrade im Text

verified	zwei oder mehr unabhängige Quellen oder zwei unabhängige Primärdokumente stimmen überein
single source	nur eine Quelle, einschließlich Primärdokumente
vendor claim	vom Verkäufer des Produkts oder der Dienstleistung veröffentlicht
estimate	die eigene Rechnung oder Ableitung des Autors aus den zitierten Daten

So lesen Sie diesen Bericht

Schlüsselzahlen tragen einen Vertrauensgrad. **verified** bedeutet, dass zwei oder mehr unabhängige Quellen übereinstimmen oder zwei unabhängige Primärdokumente übereinstimmen; **single source** bedeutet nur eine Quelle (einschließlich Primärdokumente); **vendor claim** bedeutet, vom Verkäufer des Produkts oder der Dienstleistung veröffentlicht, selbst wenn vielfach wiederholt; **estimate** bedeutet meine eigene Rechnung oder Ableitung aus den zitierten Daten, mit dokumentierten Eingaben. »Primär« beschreibt die Art der Quelle (Geschäftsbericht, Prospekt, Verordnung, Norm, amtliche Statistik); es ändert nicht die Vertrauensklasse. Wo Quellen einander widersprechen, werden die Zahlen nebeneinander angegeben und nicht gemittelt. Währungsumrechnungen verwenden den Zentralkurs der People's Bank of China vom 9. November 2026, $1 = RMB 6.7188 (Xinhua), und für Taiwan die Fixierung von SMBC am selben Tag, $1 = NT$32.12 (SMBC). Lohn- und Subventionsdaten für das Festland verwenden das Perlflussdelta, wo die Elektronikfertigung am stärksten konzentriert ist (amtliche Lohndaten aus Shenzhen), mit dem Jangtse-Delta zum Vergleich. Behauptungen aus Brancheninterviews werden als zu prüfende Behauptungen behandelt, nie als Tatsachen. Fragen, die die öffentliche Dokumentation nicht beantworten kann, sind in Kapitel 7 gesammelt.

ABBILDUNGEN

Abbildungen

01	Wo KI heute in der Fabrik passt	7
02	Eine SMT-Serienlinie: Die Menschen sitzen hinter der AOI	10
03	Etwa 97–99% der AOI-Alarme sind Fehlalarme	11
04	Eine Allzweckkamera ist keine AOI	13
05	Eine luftspaltgetrennte Bereitstellung	17
06	Reinraum- und ESD-Zonen bestimmen, wo Rechenleistung stehen darf	18
07	Geschwindigkeitsgrenzen nach Einsatzzweck	20
08	On-Prem-Rechenleistung: Speicher, Preis und was in Festlandchina legal kaufbar ist	22
09	Eine große Workstation ergibt noch keinen gemeinsamen Assistenten	23
10	Quantisierung: 8-Bit ist sicher, 4-Bit hängt vom Rezept ab, 3-Bit und darunter fällt aus	24
11	Lehrerkandidaten: Die generischen Scores haben sich in einer Generation verdoppelt	28
12	„Gibt es einen Defekt?“ bleibt für allgemeine Modelle schwer	29
13	Zwei Lehrer-Schüler-Pipelines	30
14	Wie groß das Goldstandard-Set sein muss	31
15	Edge-Policy-Latenz	33
16	Roboter-Schüler: fast serienreif nur bei strukturierten Stationen	34
17	Was eine Inspektions- oder Handhabungskraft pro Jahr kostet	37
18	Die Preisleiter	38
19	Amortisation nach Region	40
20	Warum sich das Aussortieren schlechter Platinen nicht lohnt	41
21	US-Zölle auf Roboter, AOI und Edge-Computer	42
22	Marktgrößen unterscheiden sich um mehr als das Zehnfache	45
23	Bottom-up: 40.000 Linien → 16.000 bedienbar → ≈320 Mio. RMB einmalig	46
24	Das Nachrüstfenster beträgt etwa fünf Jahre	47
25	Der Nachprüf-Keil ist von drei Seiten umzingelt	47
26	Die Marge steckt in produktisierter Hardware	49
27	Wie Käufer zahlen	50
28	Days Sales Outstanding	51
29	Der Embodied-AI-Stack	53
30	Was In-Plant-Roboterbereitstellungen tatsächlich offenlegen	55
31	Netto ersetzte Arbeitskraft pro Roboter	55
32	Zeitstrahl zu Normen und Regulierung	57

ZUSAMMENFASSUNG

Lohnend, aber eng gefasst, jetzt und mit Partnern

KERNERGEBNIS

Das Urteil: Für Elektronikfabriken ist das, was 2026 ausgereift ist und wofür Fabriken zahlen werden, die AI-Nachprüfung und Nischeninspektion auf AOI-Bildern sowie die programmierte Roboter-Maschinenbeschickung, nicht »ein großes Modell, das Platinen betrachtet, plus verkörperte Roboter, die sie greifen«.

0.46–0.65 Jahre
Amortisation der AI-Nachprüfung in den USA, Taiwan und Shenzhen, bei 50.000 RMB pro Linie

≥95% · <20 PPM
Abdeckung und Durchschlüpfe der internen AI-Nachprüfung eines führenden EMS: die Messlatte für den Verkauf an Tier-1-EMS

100–700 Mio. RMB
Adressierbarer Drittparteien-Nachrüstmarkt in Festlandchina, einmalig, über einen Zeitraum von etwa fünf Jahren

0 Punkte
Veröffentlichte Benchmark-Werte für Industriedefekte der Flaggschiffe vom November 2026: Lehrer, keine Prüfer

1.1–1.5 Jahre
Amortisation einer Cobot-Maschinenbeschickungszelle über zwei Schichten in den USA, Taiwan und Shenzhen

0.003–0.14 FTE
Arbeit, die ein Ausschleus-Roboterarm pro Schicht an einer SMT-Linie einspart: amortisiert sich nie

Inspektion
ROI

Überwachtes Deep Learning, angewendet auf die Bilder, die ein vorhandenes AOI bereits erzeugt, senkt Fehlalarme oder Nachprüfungsstunden um 60–90% (vendor claims). DBG, ein großes EMS-Unternehmen, berichtet in seinem Geschäftsbericht, dass sein internes System mehr als 95% der gekennzeichneten Komponenten abdeckt, mit Durchschlüpfen unter 20 PPM und mehr als 85% weniger menschlicher Nachprüfung; das ist die Messlatte, die ein Drittparteien-Produkt überschreiten muss, um an Tier-1-EMS zu verkaufen. Bei 50.000 RMB pro Linie amortisiert sich die AI-Nachprüfung innerhalb eines Jahres in Festlandchina, Taiwan und den USA (Shenzhen 0.65 Jahre, Taiwan 0.53, USA 0.46), Cobot-Maschinenbeschickungszellen in 1.1–1.5 Jahren und Südostasien in 2–6 Jahren, während ein Roboterarm, der einer SMT-Linie nur zum Ausschleusen schlechter Platinen hinzugefügt wird, sich nirgendwo amortisiert, weil bei über 90% Erstausbeute das Ausschleusen nur 0.003–0.14 einer Person pro Schicht erfordert estimate .

Modelle
ROBOTER

Multimodale große Modelle können weiterhin nicht der primäre Prüfer sein. Zum 11. November 2026 besitzt keines der Flaggschiffe (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 und andere) einen veröffentlichten Wert auf einem Benchmark für Industriedefekte oder PCB, und in der ACM MM 2026 PCBA-Challenge erzielten die fünf besten Systeme, trainiert auf 6.600 echten Linienbildern, nur 77–82 bei »liegt ein Defekt vor«. Frontier-Modelle können vorab beschriften und triagieren. Bei Roboterarmen sind die wahren Lehrer weiterhin menschliche Demonstrationen, Reinforcement Learning am echten Roboter und Simulation: Schüler erreichen 90–99.4% an fixierten, strukturierten Stationen, aber nur 30–80% auf bewegten Förderbändern, ungesehenen Teilen und neu-

en Produkten; ein Frontier-Agent erledigte 104 von 108 quasistatischen Aufgaben am echten Roboter, mit jeweils 280 Sekunden.

Vor Ort
BEREITSTELLUNG

Luftspaltgetrennte Bereitstellung in sauberen, kontrollierten Fabriken ist vollständig praktikabel. Eine gewöhnliche SMT-Halle ist in der Regel ein ESD-Schutzbereich und kein klassifizierter Reinraum: eine lüfterlose Jetson- oder inländische NPU-Box steht an der Linie, und GPU-Server stehen im Serverraum. Nach den eigenen Exportklassifizierungen von NVIDIA können Jetson Thor und DGX Spark ohne US-Lizenz an eine gewöhnliche Festland-Fabrik verkauft werden; DGX Station nicht. Die Standardpräzision ist 8-Bit; quantisierungsbewusstes 4-Bit vom Anbieter ist unter Bedingungen nutzbar; 3-Bit und darunter ist ausgeschlossen.

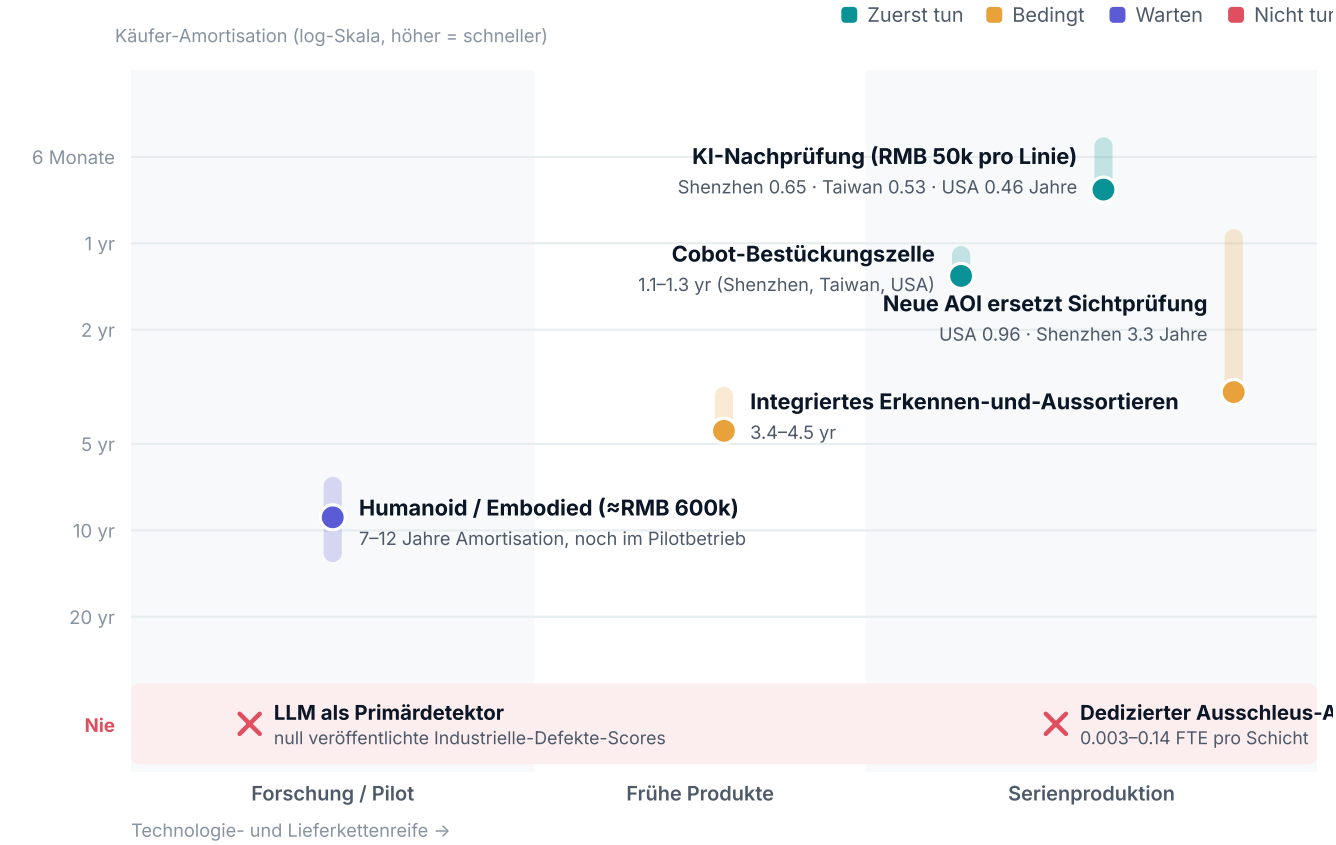
Markt
WETTBEWERB

Der Markt ist real, aber begrenzt. Festlandchina verfügt über etwa 35,000–46,000 SMT-Linien und etwa 26,000–66,000 Inline-AOIs, die nicht mit AI laufen; der adressierbare Markt für Drittparteien-Nachrüstungen beträgt etwa 100–700 Millionen RMB einmalig, über ein Fenster von etwa fünf Jahren, und AOI-Hersteller haben begonnen, die Nachprüfung in neue Maschinen einzubauen estimate. Der Wettbewerb ist bereits stark: Der Review-Keil wird von AOI-Herstellern, internen EMS-Teams und Plattformanbietern umzingelt, und die Marge liegt bei produktisierter Hardware.

Frage	Urteil	Schlüsselzahlen (Quellen und Vertrauensgrad in den Kapiteln)
① **Was betreiben gängige Elektronikfabriken heute?**	Eine SMT-Volumenlinie ist »3D SPI + AOIs vor und nach dem Reflow + stichprobenartige Röntgenprüfung + ICT/FCT + Reparatur und Sichtprüfung«, mit menschlicher Nachprüfung hinter jedem AOI. AI tritt auf drei Wegen in die Nachprüfung ein: von AOI-Herstellern eingebaut, von führenden EMS intern gebaut oder von Dritten hinzugefügt. Montage und Handhabung nutzen Arme, SCARAs und AMRs mit gelernter Wahrnehmung und programmierter Bewegung; verkörperte AI befindet sich noch in der Pilotphase	Takt ≈13–22 s pro Platine; ein ODM veranschlagte 30.6M RMB für AOIs für 18 Linien, ≈3–5 AOIs pro Linie; ≈97–99% der AOI-Alarme sind Fehlalarme
② **Wie weit reicht die Technologie, kann Geld null Menschen kaufen, und was kann jede Region nutzen?**	Inspektion ist industrietauglich als »dedizierter Detektor + AI-Nachprüfung + einige Menschen«; selbst unbegrenztes Geld erreicht keine null Menschen. Das Festland nutzt nur inländische Modelle oder lokale offene Gewichte; private Fabriken in Taiwan können fast alles nutzen, aber Fabs und kritische Infrastruktur schließen Festland-Marken in der Praxis aus; Südostasien hat die wenigsten Beschränkungen; die USA haben den vollständigen Stack plus Zölle und Zertifizierung	AI-Nachprüfung senkt Fehlalarme oder Nachprüfungsstunden um 60–90% vendor claims ; eine 4K-Kamera über eine ganze Platine ≈78 µm pro Pixel gegenüber 8–20 µm bei einem AOI; auf dem Festland gefertigte Arme zahlen ≈40% kombinierten US-Zoll, Instrumente vom AOI-Typ ≈37.5%
③ **Kann sie vor Ort in Reinräumen und luftspaltgetrennten Fabriken laufen, auf welcher Hardware, schnell und genau genug?**	Ja; die Beschränkungen sind prozessual, nicht rechnerisch. Eine lüfterlose Box ≤100 W an der Linie, GPU-Server im Serverraum; Thor für Stationen, ein Mac Studio oder ein inländisches Gerät für die Werks-Q&A kein DGX Station auf dem Festland. Die Geschwindigkeit passt, wenn die Modellgröße zur Aufgabe passt; 8-Bit beibehalten, quantisierungsbewusstes 4-Bit vom Anbieter mit Bedingungen nutzen, nie 3-Bit	≈0.5 s pro Ausschnitt für ein 8B-Vision-Modell auf Thor estimate ; Thor-Dev-Kit 59,104 RMB in China; DGX Station $123–140k, 139 Tokens/s insgesamt für 32 Nutzer auf einem 743B-Modell

Frage	Urteil	Schlüsselzahlen (Quellen und Vertrauensgrad in den Kapiteln)
④ **Kann das neueste Flaggschiff über Dutzende Runden als Lehrer wirken und in lokale Schüler destilliert werden, auch für flexible Fertigung und Roboterarme?**	Inspektion: Vorab-Beschriftung und Triage; Lehrer-Labels nur für Klassen mit »≥299 Defekten, null Übersehenen« auf einem Golden Set automatisch übernehmen. Roboter: Die Lehrer sind weiterhin menschliche Demonstrationen, RL am echten Roboter und Simulation, und Schüler nähern sich der Produktionsreife nur an strukturierten Stationen	Kein November-Flaggschiff hat einen Industriedefekt-Wert; Schüler 97.8% bei trainierten Steckverbindern, 29–67% bei ungesehenen, 37–59% auf bewegten Förderbändern; Frontier-Modelle, die Gelenke direkt steuern, 0–5.5%
⑤ **Was kostet es, und wie hoch ist der ROI je Region?**	AI-Nachprüfung amortisiert sich innerhalb eines Jahres in Festlandchina, Taiwan und den USA; Cobot-Beschickung 1.1–1.5 Jahre; Südostasien 2–6 Jahre; ein Ausschleusarm an einer SMT-Linie amortisiert sich nie	50k-RMB-Nachprüfbox pro Linie: Shenzhen 0.65 Jahre, Taiwan 0.53, USA 0.46, Vietnam 2.2 estimate ; akzeptabler Drittparteien-Add-on-Preis ≈10–30k RMB pro Linie einmalig oder 5–10k RMB pro Jahr
⑥ **Sind Wettbewerber bereits stark genug, und ist es besser, jetzt zu starten oder zu warten?**	Die Nachfrage ist nicht gesättigt, aber das Angebot ist überfüllt; OEM-AI begrenzt das Nachrüstfenster auf etwa fünf Jahre. Jetzt mit einem engen Produkt über Channel-Partner eintreten, ohne Plattformanbieter frontal anzugreifen; auf Produktionsdaten von 2027–2028 warten, bevor man auf verkörperte AI setzt	≈26–66k Inline-AOIs ohne AI auf dem Festland, jährlich um 8–12% schrumpfend estimate ; der führende AI-Vision-Anbieter erzielte 2025 einen Umsatz von 1.086 Mrd. RMB und einen Nettoverlust von 991 Mio. RMB; der 3C-Umsatz des nächsten börsennotierten Vergleichsunternehmens fiel im Januar–Mai 2026 um 79.9% im Jahresvergleich

FIG 01 **Wo KI heute in der Fabrik passt*Die KI-Nachprüfung von AOI-Fehlalarmen und die Maschinenbestückung per Cobot sind ausgereift und amortisieren sich schnell; ein dedizierter Ausschleus-Arm und ein LLM als Primärinspektor lohnen sich nicht.***



Quelle: Amortisation aus dem Modell des Autors (Kapitel 4); die Reifeachse ist eine Einschätzung des Autors.

Was man bauen sollte und was nicht

	Empfehlung
Bauen	AI-Nachprüfung und Defektklassifikation auf AOI-Bildern, für mittelgroße EMS mit älteren oder gemischten AOI-Marken und ohne AI-Upgrade vom AOI-Hersteller; Inspektion, die das SMT-AOI schlecht abdeckt (Durchsteck- und Handlötung, Conformal Coating, Kabel, Steckverbinder und FPCs, Enderscheinung und Schrauben); Cobot-Be- und Entladezellen, wo jedes Teil gehandhabt wird (Prüfvorrichtungen, Trays, Verpackung), mit Inspektion als Add-on; ein Golden Set und Schattenmodus auf einer echten Linie als Grundlage jeder Behauptung
Nicht bauen	Eine allgemeine »AI-plus-Roboterarm«-Plattform; ein multimodales LLM als primärer Prüfer; ein Roboterarm an einer SMT-Linie nur zum Ausschleusen von NG-Platinen (eine NG-Ausschleusung oder ein Puffer erledigt das); Humanoide und Robots-as-a-Service; Frontier-Modelle innerhalb eines Regelkreises oder einer Sicherheitsfunktion; Berge von GPU-Speicher an Stationen, um »vollwertige« Modelle laufen zu lassen; Direktverkäufe an Fabs, taiwanische kritische Infrastruktur oder US-Rüstungselektronikwerke
Mit wem man zusammenarbeiten sollte	AOI-Hersteller und Distributoren (Channel- und Bildschnittstellen; Distributoren brachten 46.68% des Umsatzes von Sinic-Tek 2025 ein); Cobot-Hersteller (zertifizierte Arme und Sicherheitsfunktionen); EMS-Unternehmen mit 10–40 Linien (rollierende Expansion und Anzahlungen); lokale Integratoren in Thailand und anderswo (zur Erfüllung der BOI-Local-Linkage-Regeln)

Die Linie heute

01 Was Elektronikfabriken heute betreiben und was technisch machbar ist

Die Standard-Volumenlinie: mehrere AOIs, stichprobenartige Röntgenprüfung, zentrale Nachprüfung, 13–22 Sekunden pro Platine

Primärunterlagen börsennotierter ODMs zeigen, wie eine Volumenlinie der Jahre 2025–2026 tatsächlich aussieht. Der Hongkong-Prospekt von Longcheer (Daten bis 30. September 2025) führt seine SMT-Qualitätsprüfungen als »Röntgeninspektion, Lotpasteninspektion (SPI), automatisierte optische Inspektion (AOI), PCBA-Funktionstest und Sichtprüfung« auf; sein Prozessablauf umfasst Pre-Reflow-AOI, Post-Reflow-AOI, stichprobenartige Röntgenprüfung, NG-Platinen zu einer SMT-Reparaturstation und dann eine abschließende Sichtkontrolle. Seine Werke in Huizhou, Nanchang und Vietnam verfügen über 34, 24 und 8 SMT-Linien sowie 67, 60 und 13 Endmontagelinien (Longcheer H-share prospectus) single source, primary filing. Huaqins Laptop-Werk von 2023 plante 18 SMT-Linien mit einem AOI-Budget von 30.6 Millionen RMB, etwa 1.7 Millionen RMB pro Linie, was bei inländischen 3D-AOI-Preisen drei bis fünf AOIs pro Linie für die Pre- und Post-Reflow-Inspektion doppelseitiger Platinen bedeutet (Huaqin A-share prospectus) primary filing + estimate. Eine Volumenlinie hat daher in der Regel zwei bis vier Nachprüfungspunkte, nicht einen.

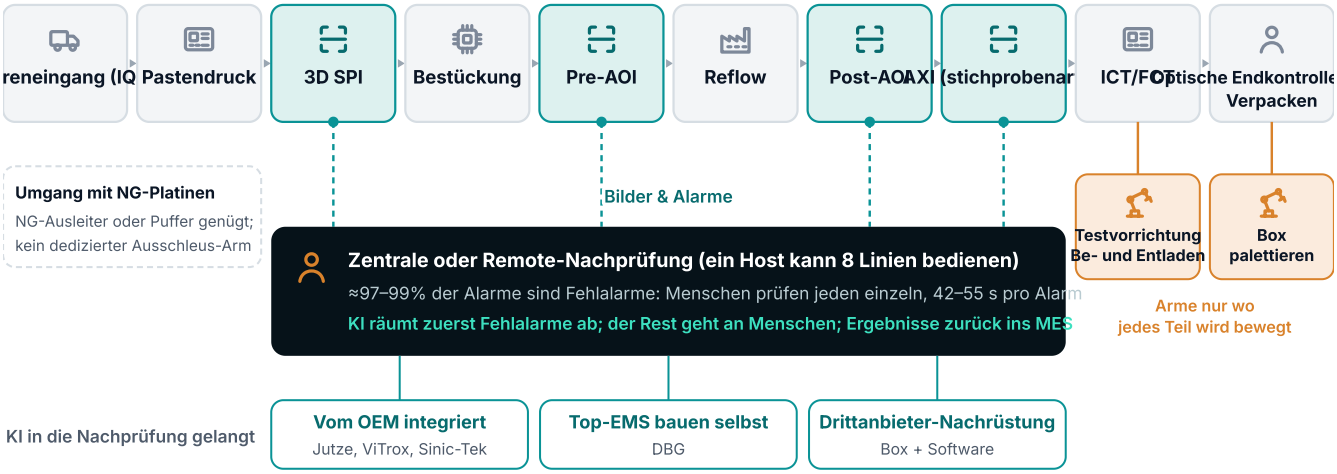
Die Maschinen sind bereits hochautomatisiert: Huaqin meldet 96% Automatisierung auf seinen SMT-Linien (Huaqin H-share prospectus) single source. Die verbleibenden Menschen sitzen in Nachprüfung, Reparatur, Sichtprüfung, Test und Endmontage, und kein Unternehmen veröffentlicht, wie viele Prüfer es pro Linie hat (Kapitel 7).

Drei unabhängige Schätzungen stimmen bei einem Takt von **13–22 Sekunden pro Platine** überein (eine Umweltunterlage von 2026 für fünf Hebei-Linien (Hebei SMT project EIA), Huaqins Laptop-Linien und ein iFactory-Fall) oder 7–11 Sekunden pro Seite bei doppelseitigen Platinen estimate. Das Schwierige an einer High-Mix-Linie ist nicht das Volumen, sondern die Einführung neuer Produkte und der Produktwechsel, und jeder Produktwechsel ist ein weiterer Kaltstart für ein AI-Modell.

Auf der Softwareseite sind die eigenen Nachprüfungsstationen der AOI-Hersteller bereits zentralisiert: Die Reparaturstations-Software von Sinic-Tek lässt »einen Host die Ausrüstung von acht SMT-Linien steuern« (Sinic-Tek prospectus) single source, primary, 2022, und nachdem Jabil 2026 die ViTrox-V-ONE-Fernsteuerung von AOIs eingeführt hat, konnte ein Ingenieur 5–8 AOIs fernbedienen gegenüber 1–2 lokal (AtlasPCB) vendor claim. Ein Drittparteien-AI-Nachprüfprodukt dockt genau an diese Zentralstation an, über den Bild- und Ergebnis-Export des AOI (IPC-CFX, IPC-HERMES-9852 oder das eigene Format eines Anbieters). Die AOIs selbst bekommen schnell AI: Yamaha gibt an, seine AI-Beurteilung dauere etwa eine Sekunde pro Aufruf, helfe Prüfern nach 30 NG-Bildern und werde nach 1,000 »sehr hoch« an Qualität (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Eine SMT-Serienlinie: Die Menschen sitzen hinter der AOI** ***Die Arbeit konzentriert sich nach der automatischen optischen Inspektion auf die Nachprüfplätze. Genau dort setzt KI an.***

Eine SMT-Serienlinie Takt ≈13–22 s pro Platine · 3–5 AOIs pro Linie



Quelle: Öffentliche Berichte und Herstellerunterlagen (Huaqin, DBG, Sinic-Tek u. a.); vom Autor zusammengestellt.

Stufe	Gängige Produkte und Technologie	Repräsentative Anbieter	Wie weit AI gekommen ist
Nach dem Drucken	3D SPI (strukturiertes Licht, Moiré, teils Lasertriangulation)	Koh Young, Sinic-Tek, TRI	OEM-AI-Lotpastenbeurteilung in Produktion; Jutze sagt, einige Modelle seien »vollständig AI«
Nach Bestückung und Reflow	2D/3D-AOI mit mehrwinkligen, mehrfarbigen LEDs und strukturiertem Licht, 8–20 µm pro Pixel, 8–12-MP-Kameras	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek und andere	OEM-AI-Nachprüfung (Jutze in Produktion, ViTrox V510Ai, Sinic-Tek mit Tencent); führende EMS bauen eigene (DBG); Drittparteien-Add-ons
Unterseitig terminierte und verdeckte Lötstellen	Stichprobenartige oder Inline-Röntgenprüfung (AXI)	—	Hikvision behauptet, sein Guanlan-Modell beurteile Röntgen-Lunker in weniger als 1% der Fälle falsch vendor claim
Nachprüfung und Reparatur	Zentrale Nachprüfungsstation (ein Host, acht Linien), Fernprüfung, Reparaturstation	Software der AOI-Hersteller, ViTrox V-ONE	AI schließt Fehlalarme, die Menschen erledigen den Rest
Sichtprüfung, IQC, FQC	Überwiegend manuell; einige AI-Erscheinungsbild-Maschinen mit Roboterbeladung	SmartMore und andere	Erscheinungsbild-Maschinen in Produktion; keine öffentlichen Personalzahlen für IQC/FQC
Handhabung und Maschinenbeschickung	Sechsachsarme und Cobots, SCARA, AMRs, 2D/3D-Vision-Führung	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Gelernte Wahrnehmung plus programmierte Bewegung in Produktion; VLAs und Humanoide noch in der Pilotphase
Daten und Systeme	MES, IPC-CFX, IPC-HERMES-9852	—	Ergebnisse werden ins MES zurückgeschrieben; Unterstützung in älterer Ausrüstung ungleichmäßig

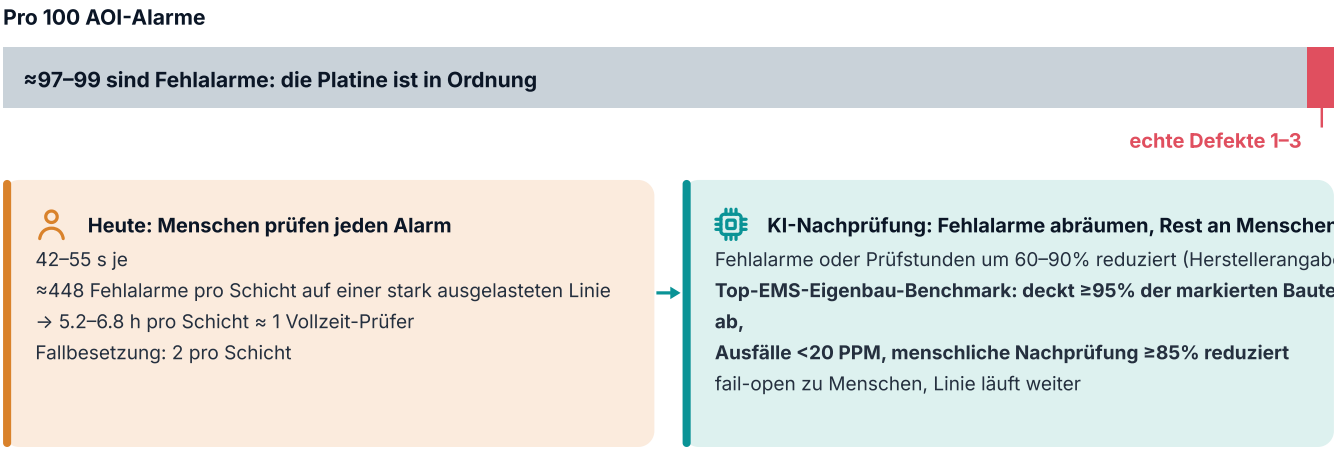
Quellen zur Tabelle: die oben genannten Unterlagen und (Koh Young, Jutze 2025 annual report, Sinic-Tek 2025 annual report, ViTrox, Sohu on Hengtianyi, Xinhua on Hikvision).

Die Menschen sitzen hinter dem AOI, weil die meisten Alarme falsch sind

Die Prüfkette einer gängigen SMT/PCBA-Linie läuft: 3D SPI nach dem Drucken für fehlende, unzureichende oder überschüssige Paste, Brücken und Versatz; AOI nach der Bestückung (optional) und nach dem Reflow für fehlende, verschobene, gedrehte, verkehrte oder tombstoned Teile und Lotvolumen; Röntgenprüfung bei BGAs oder anderen unterseitig terminierten Teilen; dann In-Circuit- oder Flying-Probe-Test, Funktionstest und manuelle Sichtprüfung. Jedes AOI hat einen Prüfplatz und eine Reparaturstation dahinter (Sinic-Tek prospectus, JLCPCB) verified. Das Post-Reflow-AOI ist das häufigste und wertvollste, sieht aber nur Form und Fläche; Lötstellen, Lunker und Benetzung unter einem BGA erfordern Röntgen- oder elektrische Prüfung (JLCPCB).

Die Arbeitskraft sitzt in der Nachprüfung, weil die meisten Alarme falsch sind. Ein britisches EMS verzeichnete **54,019 Fehlalarme gegenüber 544 echten Defekten über 917 Teilenummern, etwa 99% der Alarme** (arXiv 2505.03576) single source, preprint; im Werk Rastatt von Siemens waren **97.5% der vom AOI abgelehnten Platinen gut** (Siemens case study) vendor claim. Eine ausgelastete Linie verzeichnet etwa 448 Fehlalarme pro Schicht zu je 42–55 Sekunden, 5.2–6.8 Stunden pro Schicht, etwa ein Vollzeit-Prüfer (iFactory) vendor claim; hours are an estimate.

FIG 03 **Etwa 97–99% der AOI-Alarme sind Fehlalarme** ***Heute bestätigt ein Mensch jeden Alarm; die KI-Nachprüfung räumt zuerst die Fehlalarme ab und gibt den Rest an Menschen weiter.***



Quelle: Siemens-Fallstudie, iFactory (Herstellerangabe), DBG-Geschäftsbericht 2025; Prüfstunden vom Autor geschätzt.

Aus »Nachprüfungsstunden pro Schicht = Platinen × Fehlalarme pro Platine × Sekunden pro Aufruf« kann ein Prüfer etwa 0.7–1 Volumenlinie mit einem regelbasierten AOI abdecken. Sobald AI 60–90% der Fehlalarme entfernt, würde die reine Beurteilungszeit 2–14 Linien pro Person erlauben, aber Laufwege, NG-Pufferhandhabung und Reparaturübergaben begrenzen die lokale Abdeckung auf zwei bis vier benachbarte Linien; nur zentrale oder Fernprüfung nähert sich der theoretischen Obergrenze estimate; seconds per call come from one source. Da gängige AOIs 60–80 cm² pro Sekunde scannen (Yunheng 2026 AOI guide), dauert eine 20×15-cm-Seite 4–5 Sekunden; die AI muss nur die wenigen gekennzeichneten Bereiche innerhalb eines Takts verarbeiten, nicht die Platine neu scannen.

Was in der Produktion bereits funktioniert: überwachte Klassifikation auf den eigenen Bildern des AOI

Das bewährte Muster ist überwachte Deep-Learning-Klassifikation auf Bildern des AOI, das eine Fabrik bereits besitzt. Fujitsus Add-on auf ADLINKs AOI in Taiwan senkte Fehlalarme um **mehr als 90%** (Fujitsu case); OKI nutzt seit Juli 2026 AI, um die manuelle Bestätigung nach dem AOI im eigenen EMS-Geschäft um etwa **80%** zu senken (Circuits Assembly); Siemens meldet 60% weniger manuelle Analyse, €160,287 im ersten Jahr gespart und eine Amortisation von acht Monaten (Siemens case study) all vendor claims. Zwei Dinge sind diesen Fällen gemeinsam: Fast niemand veröffentlicht eine Durchschlupfrate, und alle lassen Menschen bei den verbleibenden Alarmen. Der eine

Fall, der eine Durchschlupfzahl angibt (0.8% auf 0.3%, gegenüber vertraglichen 0.5%), ist intern widersprüchlich (iFactory). AI senkt Nachprüfungsstunden, nicht den Nachprüfungsjob selbst.

Die AI-Nachprüfung ist von Drittparteien-Add-ons zu OEM-Funktionen und internen Systemen gewandert, was ändert, woran ein Neuling gemessen wird. Jutzes Geschäftsbericht 2025 führt sein »3D-AOI-Upgrade« als in Serienproduktion und sagt, es »nutzt AI-Algorithmen, um die manuelle Nachprüfung am Inspektionsende zu entfernen« (Jutze 2025 annual report); Sinic-Tek hat seit 2019 ein »multimodales AI-gestütztes manuelles Nachprüfsystem« entwickelt und entwickelt gemeinsam mit Tencent ein Top-and-Bottom-3D-AOI mit dem Ziel von über 99.5% Defektabdeckung und 98% Erkennungsgenauigkeit (Sinic-Tek 2025 annual report) vendor claims in primary filings. Große EMS-Unternehmen bauen eigene. DBGs Geschäftsbericht 2025 führt »AI-Beurteilung für die Sichtprüfung« als abgeschlossenes Projekt auf, das »mehr als 95% der vom AOI gekennzeichneten Komponenten abdeckt«, mit einer »Durchschlupfrate unter 20 PPM« und »mehr als 85% weniger manueller Nachprüfung«, vollständig rückverfolgbar (DBG 2025 annual report) single source, primary. Das ist die einzige Elektronikwerk-Zahl, die ich gefunden habe, die Abdeckung, Durchschlüpfe und Reduktion zusammen angibt. **Ein Drittparteien-Produkt, das an Tier-1-EMS verkauft wird, muss DBGs 95%-Abdeckung, 20 PPM Durchschlüpfe und 85% Reduktion erreichen**; unterhalb von Tier 1 ist der Benchmark die 50–85%-Reduktion, die Anbieter bewerben.

Anomalieerkennung, die nur auf guten Teilen trainiert ist, ist schnell (EfficientAD-S braucht 2.2 ms pro Bild (EfficientAD, WACV 2024)), aber schwach bei »logischen« Anomalien wie fehlenden oder falschen Teilen, und die besten Methoden erreichen im Schnitt unter 60% AU-PRO auf dem härteren MVTec AD 2 (MVTec AD 2) single source. Fehlende und falsche Teile werden besser gegen CAD und BOM geprüft.

Die Obergrenze um jeden Preis: Schichten von Maschinen, sehr wenige Menschen, aber nicht null

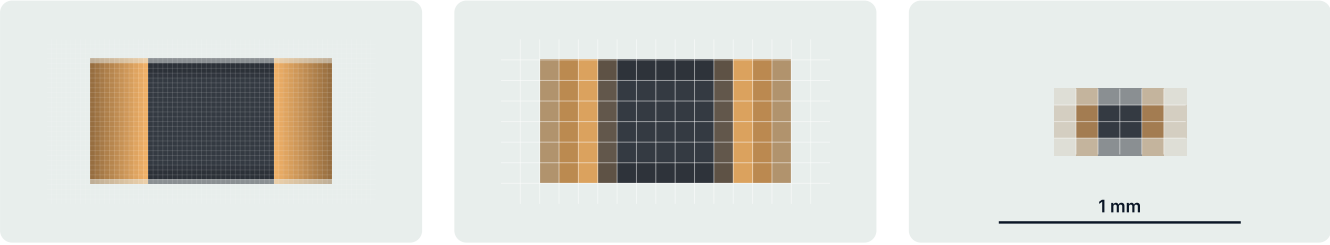
Ich habe kein Elektronikwerk gefunden, das die manuelle Nachprüfung vollständig entfernt und null Durchschlüpfe erreicht hat. Die Praxis für Produkte der IPC-A-610-Klasse 3 (hohe Zuverlässigkeit) lässt weiterhin geschulte Prüfer über Grenzfälle entscheiden, die automatische Schwellen nicht klassifizieren können, und ein PCB-Bestücker nennt »automatisierte Inspektion kann die menschliche Freigabe ersetzen« ein »weit verbreitetes Missverständnis« (PCBCart) single source. Auch die menschliche Ausgangsbasis muss korrigiert werden: Unter Berufung auf eine Sandia-Studie von 2012 erkennt ein einzelner Sichtprüfer bestenfalls etwa **80%** der Defekte, zwei in Reihe etwa 96% (Instrumental) single source, cited by a vendor. Die vertretbare Behauptung ist »AI plus eine kleine menschliche Nachprüfschleife schlägt rein manuelle Inspektion«, nicht »AI ersetzt die Inspektion«.

Ohne Budgetgrenze ist die Obergrenze 3D SPI, 3D AOI, Röntgenprüfung, AI-Nachprüfung und AI-Erscheinungsbild-Stationen übereinander gestapelt, wobei Menschen nur noch für Grenzfälle, neue Produkte und Produktwechsel, neue Defekttypen und die Kundenfreigabe übrig bleiben. Flaggschiff-Werke melden 85–92% weniger manuelle Inspektion (I-Connect007, 2021, WiAdvance) vendor claims; meine Schätzung der Obergrenze ohne Budget ist Nachprüfarbeit bei 8–15% des heutigen Niveaus estimate. Die Endmontage bleibt arbeitsintensiv: Die Automatisierung der iPhone-16-Tasten wurde wegen hoher Defektraten aufgegeben (9to5Mac) single source.

Eine allgemeine hochauflösende Kamera ist kein AOI

Produktions-AOIs bilden Kachel für Kachel mit 8–20 µm pro Pixel ab, mit 8–12-MP-Kameras, programmierbarer mehrwinkliger mehrfarbiger Beleuchtung und strukturiertem 3D-Licht (Koh Young, Saki) verified. Die Branche hat schon lange festgestellt, dass ein 01005-Teil bei 25 µm pro Pixel nur etwa 8×16 Pixel abdeckt, »möglicherweise nicht genug« für eine zuverlässige Beurteilung (Circuits Assembly, 2007). Nach derselben Rechnung estimate erreicht eine 4K-Kamera über eine 300-mm-Platine etwa **78 µm pro Pixel**: Ein imperiales 0402-Teil ist etwa 13×6 Pixel, genug für Anwesenheit und groben Versatz, aber nicht für die Lötstellenform, und ein 01005 ist etwa 5×3 Pixel, praktisch unsichtbar. Um 15 µm pro Pixel über eine ganze Platine zu erreichen, wären etwa 330 Megapixel nötig, weshalb AOIs eine Kamera über die Platine bewegen. Wenn jemand »0402« sagt, prüfen Sie, welches gemeint ist: metrisch 0402 ist imperial 01005, weit jenseits jeder Weitwinkelkamera.

FIG 04 **Eine Allzweckkamera ist keine AOI***Eine 4K-Kamera über einem 300-mm-Panel liefert etwa 78 µm pro Pixel: Ein 0402-Bauteil schrumpft auf ≈13×6 Pixel, ein 01005 verschwindet.*



AOI: 8–20 µm pro Pixel
0402 ≈ 100×50 px

4K-Kamera über ein 300-mm-Panel: ≈78 µm/px
0402 schrumpft auf ≈13×6 px: genug für Anwesenheit und grobe Verschiebung

Dieselbe Kamera auf einem 01005
≈5×3 px: praktisch unsichtbar

Quelle: Circuits Assembly; Berechnung des Autors für ein 300-mm-Panel und einen 4K-Sensor. Die Pixelraster sind illustrativ.

Auf Linien, die bereits AOIs haben, ist der beste Ort für AI oben auf den Bildern des AOI. Nur Linien ohne AOI oder End-of-Line-Stationen rechtfertigen eine geschlossene Kamerastation mit kontrollierter Beleuchtung; eine offene hochauflösende Kamera über der Linie ist Umgebungslicht und Schatten am stärksten ausgesetzt (CSDN).

Montage und Handhabung: gelernte Wahrnehmung plus programmierte Bewegung ist in Produktion

Der IFR-Bericht 2026 sagt, »Machine Vision für Inspektion und Roboterführung ist ausgereift und industrietauglich«, während »allgemeine Physical AI und Lernen aus Simulation und Erfahrung noch nicht industrietauglich sind« (IFR World Robotics 2026 executive summary) single source, primary. Elektronik ist der größte Roboterkäufer (142,235 weltweit installierte Einheiten 2025, 59% aller Installationen in China (IFR, IFR China release)), doch China hat nur 166 Roboter pro 10,000 Fertigungsbeschäftigte gegenüber 307 in den USA (IFR, April 2026). Das Muster in der Produktion ist gelernte Wahrnehmung plus deterministische Bewegung; End-to-End gelernte Bewegung ist nur an strukturierten Stationen nahe der Produktionsreife (Kapitel 3 und 6).

Was jede Region nutzen kann

Festlandchina setzt auf inländische Modelle und lokale Gewichte; in Taiwan entscheidet die Identität des Kunden; in den USA entscheiden Zölle und Zertifizierung.

Element	Festlandchina	Taiwan (private Werke)	Südostasien	Vereinigte Staaten
Claude (jeder Kanal)	● Nicht verfügbar	● Bedingt: Der Anbieter darf nicht mehrheitlich im Eigentum oder unter der Kontrolle von Personen oder Einrichtungen in Festlandchina, Hongkong oder Macau stehen	● Gleich	● Gleich
GPT- und Gemini-Cloud	● Nicht verfügbar	● Verfügbar	● Verfügbar (Listen nicht Land für Land geprüft)	● Verfügbar

Element	Festlandchina	Taiwan (private Werke)	Südostasien	Vereinigte Staaten
Chinesische Cloud-Modelle (Qwen, DeepSeek, Doubao, GLM, Kimi)	● Verfügbar (registrierte Modelle; das Anbieten eines Dienstes erfordert Registrierung)	● Legal, aber Festland-Hosting besteht selten die Sicherheitsprüfung des Kunden; im öffentlichen Sektor, in Staatsbetrieben und kritischer Infrastruktur verboten	● Legal (internationale Knoten)	● Legal, strenge Prüfung
Lokale offene Gewichte chinesischen Ursprungs (Qwen, GLM, DeepSeek)	● Erste Wahl; die private Bereitstellung im Werk erfordert keine Modell- oder Algorithmusregistrierung (Lesart einer Anwaltskanzlei)	● Verfügbar (öffentlicher Sektor eingeschränkt)	● Verfügbar	● Bedingt: Ursprungsprüfung
Lokale offene Gewichte nichtchinesischen Ursprungs (z. B. Gemma 4)	● Verfügbar	● Verfügbar	● Verfügbar	● Verfügbar
Klassische nichtgenerative Vision-Modelle	● Verfügbar (außerhalb der Regeln für generative AI)	● Verfügbar	● Verfügbar	● Verfügbar
Jetson Thor/T5000/T4000, DGX Spark	● Verfügbar: NVIDIA klassifiziert sie als 5A992.c, keine US-Lizenz für ein gewöhnliches ziviles Werk erforderlich; Spark hat eine China-SKU; Kanalpreise etwa 40–60% über der US-Liste	● Verfügbar	● Verfügbar	● Verfügbar
RTX 5090, RTX PRO 6000, DGX Station	● Nicht verfügbar: Die ersten beiden sind 4A090.a, DGX Station 5A992.z; legale Alternativen sind 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	● Verfügbar	● Verfügbar, aber Käufer mit Hauptsitz in China oder letztlich in chinesischem Besitz gelten als chinesisch; Malaysia verlangt Genehmigungen für Export und Umladung	● Verfügbar
Mac Studio M5 Ultra 512 GB	● Verfügbar (Apple China Store)	● Verfügbar	● Vermutlich verfügbar	● Verfügbar
Huawei Ascend und andere chinesische Beschleuniger	● Verfügbar	● Hohes Risiko	● Hohes Risiko	● In der Praxis nicht nutzbar
Auf dem Festland gefertigte Festarme, AOIs, Industriekameras	● Verfügbar	● Importierbar; seit dem 1. Dezember 2025 können die Behörden direkt gegen Festland-Marken vorgehen, und Behörden, Staatsbetriebe, bestimmte Stiftungen und Anbieter kritischer Infrastruktur können bei der Nutzung eingeschränkt oder ausgeschlossen werden	● Verfügbar	● Legal, aber kombinierter Zoll beträgt etwa 40% auf Arme, 37.5% auf Instrumente und Kameras vom AOI-Typ, 25% auf Edge-Computer
Im Ausland gefertigte AMRs und mobile Manipulatoren	● Verfügbar	● Verfügbar	● Verfügbar	● Neue Modelle benötigen bedingte Genehmigung des US-Kriegsministeriums

Element	Festlandchina	Taiwan (private Werke)	Südostasien	Vereinigte Staaten
Hikvision- und Dahua-Kameras	● Verfügbar	● Private Nutzung erlaubt, öffentlicher Sektor verboten	● Verfügbar	● Neue FCC-Zulassungen blockiert; für Bundesauftragnehmer verboten
Robotersicherheit und -zertifizierung	● GB 11291.1-2011 und GB 11291.2-2013 verbindlich, seit April 2026 in Überarbeitung; CR-Zertifizierung freiwillig; GB/T 39404-2026 behandelt die Cybersicherheit von Steuerungen	● Standardmäßig Umzäunung; zäunungslose Zusammenarbeit muss CNS 14490 oder ISO 10218 mit schriftlicher Bewertung erfüllen	● Je Kunde und Exportziel, vor dem Versand schriftlich bestätigt	● Elektrische Ausrüstung muss »zugelassen« sein (in der Praxis ein NRTL-Zeichen oder eine Feldbewertung); UL 1740 verlangt seit dem 11. August 2025, dass Roboterzellen ISO 10218-2 erfüllen
Fabs und Werke für kontrollierte Produkte	—	● Die Einkaufsverträge von TSMC verlangen SEMI E187; Festland-Ausrüstung wird aus 2-nm-Fabs entfernt	—	● ITAR-Fachdaten umfassen »Fotografien«; Verweigerungspolitik für China; vom Festland kontrollierte Anbieter können Rüstungselektronikwerke in der Praxis nicht bedienen

Quellen zur Tabelle: (Anthropic supported regions, OpenAI supported countries, Gemini availability, CAC interim measures on generative AI, NVIDIA export classification data, eCFR 15 CFR 774, NVIDIA China DGX Spark, Apple China, Covington on new BIS rules, Greenberg Traurig, June 2026, Fulcrum, Tom's Hardware on Ascend guidance, Taiwan BOFT import list, Taiwan Cyber Security Management Act history, Taiwan review rules for products endangering national cybersecurity, Taipei Times, USITC HTS chapter 99, Latham & Watkins, IPVM, China national standards platform, GB 11291.2, Intertek UL 1740 update, eCFR 22 CFR 126.1).

Festlandchina: ein vollständiger Stack, aber weder geschlossene Frontier-Modelle noch die stärksten Einzelmaschinen können hineinkommen. Registrierte inländische Cloud-Modelle, lokale offene Gewichte (Qwen unter Apache-2.0, GLM und DeepSeek unter MIT), klassische Vision-Modelle sowie inländische AOIs, Kameras und Cobots sind alle nutzbar. NVIDIAs eigene Exportklassifizierungen führen das Jetson-AGX-Thor-Kit, die T5000/T4000-Module und DGX Spark als 5A992.c, sodass ein gewöhnliches EMS- oder PCBA-Werk keine US-Lizenz benötigt, während RTX 5090 und RTX PRO 6000 (4A090.a) und DGX Station (5A992.z, TPP 58,484) auf dem Festland nicht gekauft werden können (NVIDIA export classifications, eCFR) primary; the legal reading is mine and not legal advice. RS China listet das Thor-Kit mit 59,104 RMB inklusive Steuer, etwa das 1.4-Fache des US-Listenpreises (RS China).

Vier Anwaltskanzleien stimmen überein, dass eine private Bereitstellung im Werk für den Eigenbedarf des Werks keine Modell- oder Algorithmusregistrierung erfordert, während ein gehosteter Mehr-Mandanten-Dienst in einer Grauzone liegt (Covington, FPF, JunHe) verified: consistent readings, no official ruling. GB 11291.1-2011 und GB 11291.2-2013 bleiben die verbindliche Robotersicherheits-Basislinie und sind seit dem 28. April 2026 in Überarbeitung; GB/T 39404-2026 zur Cybersicherheit von Steuerungen trat am 1. November 2026 in Kraft (revision plan 20262542, GB/T 39404-2026) single source, primary. Bis die neue GB 11291 kommt (2028–2029 estimate), ist nach ISO 10218:2025 zu konstruieren und sind GB-11291-Konformitätsunterlagen zu führen.

Taiwan: private Werke können fast alles nutzen; die Identität des Kunden entscheidet. Frontier-APIs, offene Gewichte und Festland-Ausrüstung sind alle legal (data.gov.tw). Aber seit das geänderte Cyber Security Management Act am **1. Dezember 2025** in Kraft getreten ist, können die Behörden direkt gegen Produkte vorgehen, die »Festland-Marken« sind, und sie bei Regierungsstellen, Staatsbetrieben und Anbietern kritischer Infrastruktur einschränken (Laws & Regulations Database, review rules) verified; die Liste der Anbieter kritischer Infrastruktur ist vertraulich; und TSMC hat SEMI E187 in seine Einkaufsverträge aufgenommen (CNA) und entfernt Berichten zufolge Festland-Ausrüstung aus seinen 2-nm-Fabs (TrendForce) single source. Private EMS-, Montage- und kleine PCBA-Werke sind offen; Fabs, Parkbetriebe und regierungsnahe Kunden sind in der Praxis geschlossen. Ein vom Festland kontrollierter Anbieter sollte auf E187 hin konstruieren, eine Software-Stückliste ausliefern, Offline-Updates anbieten und sich niemals vom Festland aus remote verbinden.

Südostasien: die wenigsten Beschränkungen; die Probleme sind langsame Amortisation und chinesische Muttergesellschaften. Alles ist nutzbar, aber die US-Leitlinie vom 31. Mai 2026 verlangt eine Lizenz für den Export fortschrittlicher Rechenartikel an Einrichtungen mit Hauptsitz in China oder letztlich in chinesischem Besitz, selbst wenn der Empfänger in einem Drittland sitzt (Greenberg Traurig) single source, und Malaysia hat eigene Genehmigungen für fortschrittliche Chips (Fulcrum). Die Nachfrage wandert hierher: TrendForce erwartete, dass Thailand, Vietnam und Malaysia 2026 fast 50% der Server-ODM-SMT-Kapazität erreichen (TrendForce) single source, a 2023 forecast.

Vereinigte Staaten: der vollständige Stack, zum Preis von Zöllen, Zertifizierung und kontrollierten Daten. Nach Kapitel 99 des US-Zolltarifs zahlt ein auf dem Festland gefertigter Roboterarm etwa 40% kombinierten Zoll, Instrumente und Kameras vom AOI-Typ etwa 37.5% und Edge-Computer 25%; in Malaysia oder Indonesien gefertigt etwa 10–12.5%, in Vietnam, Thailand oder den Philippinen 12.5–15%, in Taiwan 10%, mit Computern bei 0% außerhalb des Festlands (USITC HTS chapter 99, Great Lakes Customs Law) primary schedule + estimate. Die Mitteilung des USTR vom 10. November verlängerte 178 alte Section-301-Ausnahmen bis zum 9. Januar 2027, von denen keine Roboter, Inspektionsinstrumente, Kameras oder Computer abdeckt (Federal Register 2026-22925) verified.

OSHA verlangt, dass elektrische Arbeitsplatzausrüstung »zugelassen« ist (in der Praxis ein NRTL-Zeichen oder eine Feldbewertung von etwa $3,000 bis $20,000+ pro Zelle), und seit dem 11. August 2025 bringt UL 1740 ganze Roboterzellen unter ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). Rüstungselektronik ist für vom Festland kontrollierte Anbieter in der Praxis geschlossen: ITAR behandelt Produktionsfotografien als Fachdaten und wendet eine Verweigerungspolitik auf China an (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Eine für die USA bestimmte Zelle sollte einen Festarm, unbeschränkte Kameras und lokale oder US-Cloud-Modelle verwenden, mit Endmontage in Südostasien, wo dies den Ursprung tatsächlich ändert.

Bei jedem Budget unterscheidet sich die Obergrenze je Region, erreicht aber nirgendwo null Menschen estimate. Das Festland kann inländische 3D SPI, 3D AOI, Röntgenprüfung, AI-Nachprüfung und ein internes AI-Team stapeln, um die Nachprüfarbeit auf 8–15% zu senken, aber ohne geschlossene Frontier-Modelle oder DGX Station muss werksweite LLM-Q&A auf lokalen offenen Gewichten und inländischer Rechenleistung laufen (ein vollwertiges Modell für viele Nutzer benötigt mindestens zwei Ascend-Server mit acht Karten). Taiwan und Südostasien können technisch das Beste von allem nutzen, einschließlich Frontier-APIs und DGX Station, begrenzt durch Kundenidentität, Festland-Marken und Lizenzierung mit chinesischer Muttergesellschaft. Die USA haben keine technischen Grenzen, zum Preis von Zöllen, NRTL-Zertifizierung und Personaltrennung für kontrollierte Daten. Überall bleiben Menschen für Grenzfälle, neue Produkte und Produktwechsel, neue Defekttypen und die Kundenfreigabe.

Ein über Regionen hinweg verkauftes Produkt benötigt daher zwei Modell-Stacks: einen Festland-Stack auf registrierten inländischen Modellen oder lokalen offenen Gewichten und einen Übersee-Stack auf Frontier-APIs plus einem lokalen Modell nichtchinesischen Ursprungs. Die Richtlinie von Anthropic, wirksam ab dem 12. November 2026, urteilt nach Eigentum und Kontrolle, sodass ein vom Festland kontrollierter Anbieter nicht auf Claude aufbauen kann, auch nicht über eine Übersee-Tochter, eine Cloud-Plattform oder einen Wiederverkäufer (Anthropic) single source, primary.

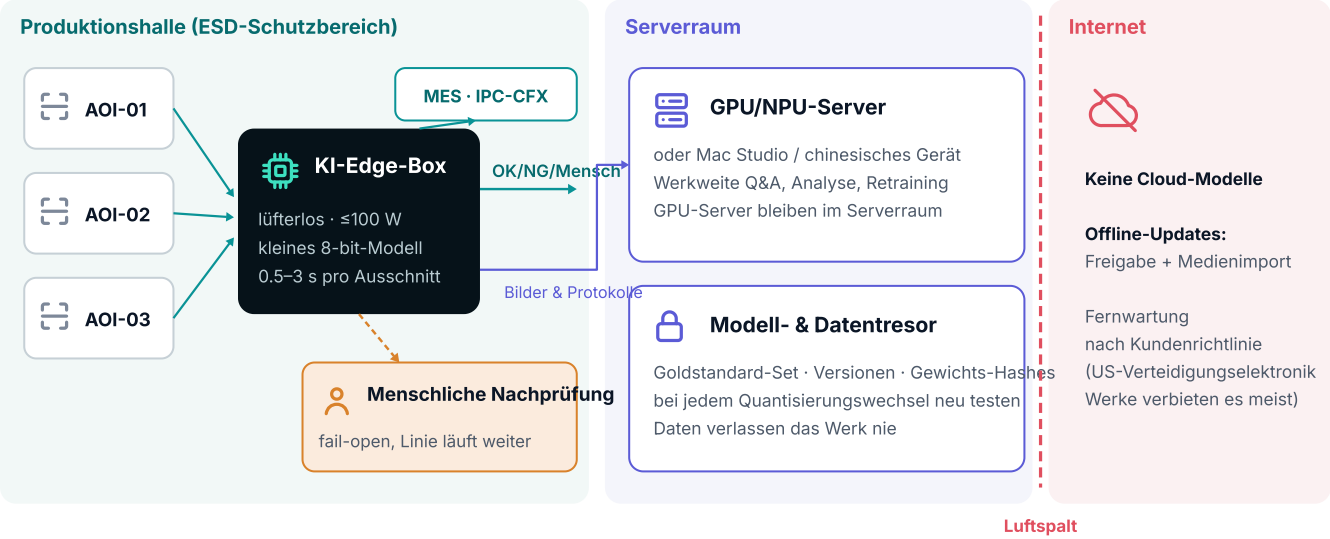
Vor Ort

02 AI im Werk betreiben: Reinräume, Luftspalte, Hardware und Quantisierung

In einer luftspaltgetrennten Fabrik ist die Beschränkung der Prozess, nicht die Rechenleistung

Vollständig offline betriebene Sichtprüfung ist Routine; die Beschränkungen sind prozessual. Chinas MIIT-Leitfaden von 2024 zur Sicherheit industrieller Steuerungssysteme verlangt Zonierung, Firewalls oder Datendioden, geprüfte zeitlich begrenzte Fernwartung, gescannte Medien und getestete Patches (guide text) single source ; Halbleiterkunden ergänzen SEMI E187 und E188, die TSMC 2023 in seine Einkaufsverträge aufgenommen hat (SEMI, TXOne) verified . Ergebnisse gehen über IPC-CFX an das MES zurück, und die Platinenidentität läuft über IPC-HERMES-9852 (Hermes standard). Die häufige Falle ist die Lizenzierung, die ein Netzwerk voraussetzt: Der Offline-Modus von Roboflow puffert Gewichte auf einer 30-Tage-Lease (Roboflow). Ein Offline-Produkt benötigt Offline-Aktivierung, signierte Updates auf gescannten Medien, eine Software-Stückliste, USB-Kontrolle, lokale Audit-Protokolle und einen CFX-Adapter.

FIG 05 **Eine luftspaltgetrennte Bereitstellung** ***Eine kleine Box an der Linie entscheidet; der Serverraum beantwortet Fragen; bei Ausfällen übernehmen Menschen, und Daten verlassen das Werk nie.***



Quelle: Zusammenfassung des Autors (Kapitel 2).

Geschlossene Frontier-Modelle gelangen in der Praxis nicht in den Serverraum einer Fabrik. Gemini wird luftspaltgetrennt nur über Google Distributed Cloud für Regierungen und Großunternehmen angeboten (Google Cloud); Microsofts Foundry Local öffnet große Modelle nur für »qualifizierte Kunden« und nennt kein GPT-Modell (Microsoft); Claude wird nur in Cloud-Regierungsregionen angeboten, ohne Bereitstellung vor Ort (Claude for government); keines der drei wird in Festlandchina angeboten. Fabrikausrüstung muss um offene Gewichte herum konstruiert werden; Frontier-Modelle können höchstens Bilder in der Zentrale offline beschriften oder erklären, wenn der Kunde anonymisierte Bilder hinauslässt.

Serverraum-Hardware sollte als Büroqualität behandelt werden: Das Mac Studio ist für 10–35 °C und DGX Spark für 5–30 °C ausgelegt, mit Berichten über thermische Abschaltungen bei anhaltender Inferenz (Apple, NVIDIA forum) verified. Die Linie bekommt lüfterlose Industrieboxen, etwa einen Jetson-T4000-Industrie-PC mit Nennbereich −25 bis 60 °C (AAEON) vendor claim. Da eine Stunde Ausfallzeit ein großes Werk $36,000 bis $2.3 Millionen kostet (Siemens True Cost of Downtime 2024) vendor survey, muss jeder AI-Schritt an der Linie **fail open**: Bei jedem Fehler fällt er auf den bestehenden AOI-plus-Menschen-Prozess zurück.

Reinraum- und ESD-Zonen: Eine gewöhnliche SMT-Halle ist ein ESD-Schutzbereich, kein klassifizierter Reinraum

Was begrenzt, wo Rechenleistung stehen kann, sind Statik und Luftstrom, nicht die Reinheitsklasse. Die meisten SMT-Montagebereiche sind »formal nicht partikelkontrolliert«, und Weltklasse-Werke »können ISO 8 erreichen oder sich ihr nähern« (Circuitnet expert panel) single source. Medizinische PCBA und Kameramodule liegen typischerweise in ISO 7, Bare Die und optische Ausrichtung in lokalen ISO 5–6 (PCBSync) single source; Reinraumbauer nennen ISO 7–8 für SMT allgemein (Angstrom Technology) vendor claim.

Die universelle Anforderung ist ESD-Kontrolle: ANSI/ESD S20.20-2021 verlangt isolierte Leiter unter 35 V und hält starkfeldige Isolatoren mindestens 30 cm von empfindlichen Bauteilen entfernt (ANSI, ESDA, Desco) verified. Fabs stellen Rechenleistung in die Sub-Fab oder das Rechenzentrum, in flüssiggekühlte Racks, die SEMI S2/S8/F47 erfüllen (Prodrive) vendor claim.

Ob Ausrüstung in einen Reinbereich darf, wird auf Partikelemission nach ISO 14644-14 geprüft; Roboter halten solche Zertifikate (Fraunhofer IPA), aber ich habe keinen Industrie-PC, keine Edge-AI-Box und keinen GPU-Server mit einem solchen gefunden. CE, FCC und IEC 60068 zeigen, dass eine Box ihre Umgebung toleriert, nicht dass sie einen ESD-Schutzbereich betreten darf; das prüft der ESD-Koordinator des Kunden. GPU-Server kommen nie an die Linie: Eine DGX Station zog im Test 1,292 W an der GPU (StorageReview).

FIG 06 **Reinraum- und ESD-Zonen bestimmen, wo Rechenleistung stehen darf** ***Eine normale SMT-Halle ist ein ESD-Schutzbereich, kein Reinraum. GPU- und NPU-Server gehören in den Serverraum.***



Quelle: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, SEMI-Normen und Herstellerunterlagen; vom Autor zusammengestellt.

Zone	Typische Umgebung	Rechenleistung an der Linie	GPU/NPU-Server
Gewöhnliche SMT/PCBA-Halle (Konsum, Industrie)	ESD-Schutzbereich nach S20.20 oder IEC 61340-5-1, meist kein klassifizierter Reinraum (≈ISO-8-Äquivalent); 40–60% RF	Lüfterlose Box in geerdetem Metallgehäuse (Jetson T4000/Thor, chinesische NPU-Box) neben dem AOI oder Prüfplatz; Kunststoffmonitore und -gehäuse außerhalb des Bereichs oder ≥30 cm von Platinen	Serverraum oder klimatisierter IT-Schrank
Medizin, Kameramodule, EV-Elektronik (ISO 7–8 Reinraum)	Reinraum plus ESD-Schutz	Lüfterlos, IP65/66 abgedichtet, glatt und leicht abwischbar; ISO-14644-14-Daten oder ein IPA-Zertifikat anfordern	Außerhalb des Reinraums, vernetzt hinein
Bare Die, Drahtbonden, optische Stationen (lokal ISO 5–6)	Mini-Environment	Nur Kameras und Licht innen; Rechenbox außen	Serverraum
Fab-Reinraum (ISO 5 und sauberer)	Fab-spezifische ESD- und Elektrostatik-Anziehungspläne, Ausrüstung nach SEMI E78; Cybersicherheit nach SEMI E187/E188	Nur werkzeugintegrierte, fab-qualifizierte Steuerungen; keine Add-on-PCs mit Lüftern	Flüssiggekühlte Sub-Fab-Racks oder das Fab-Rechenzentrum

Die Tabelle ist meine Synthese der obigen Quellen estimate. Für ein Produkt bedeutet das eine lüfterlose liniennahe Box unter etwa 100 W, ein Metallgehäuse mit Schutzleiterverbindung, 12–24 V Gleichstromeingang, Betrieb im weiten Temperaturbereich bei realem Schrank-Luftstrom und einen ISO-14644-14-Partikelemissionsbericht für Kunden bei ISO 7 und sauberer. Jede Konstruktion, bei der »ein GPU-Server vier Linien bedient«, stellt den Server in den Serverraum und lässt nur eine kleine Box oder eine Kamera an der Linie.

Geschwindigkeitsgrenzen je Einsatz: Genauigkeit ist schwieriger als Geschwindigkeit

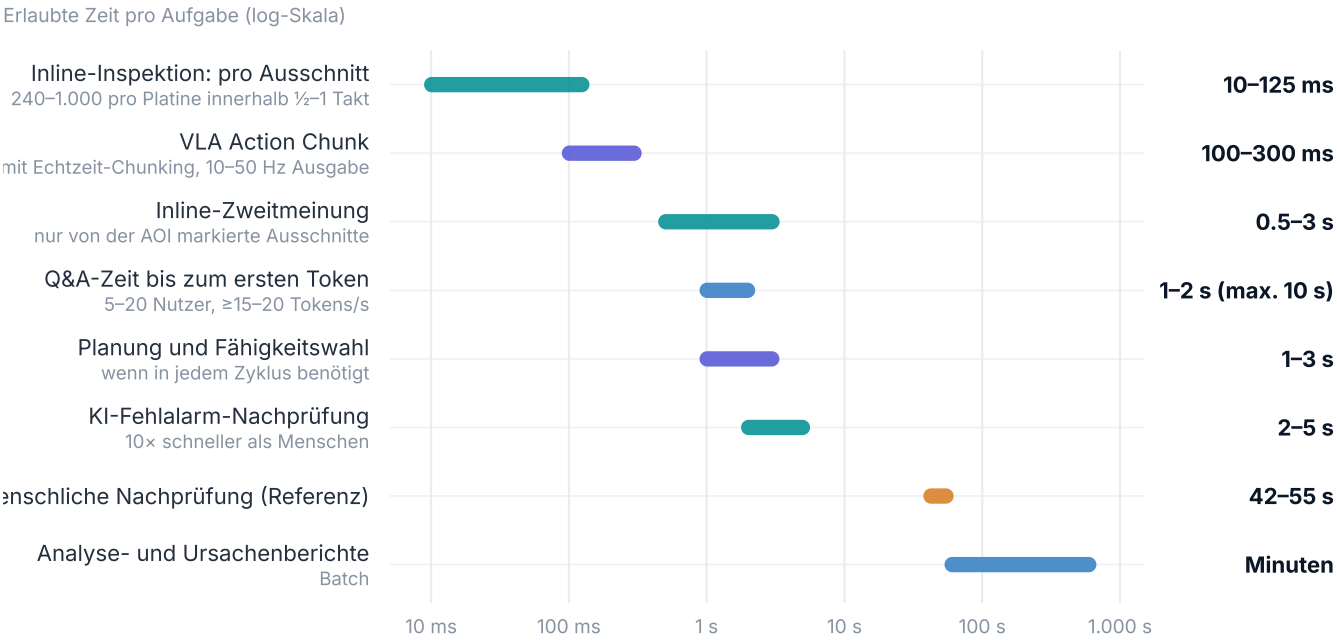
Fabriken und Anbieter veröffentlichen fast nie AI-Latenz- oder Verfügbarkeitsziele. Die folgenden Schwellen sind aus Takt, menschlicher Nachprüfzeit, Robotersteuerungsraten und Interaktionsforschung abgeleitet estimate:

Einsatz	Geschwindigkeitsgrenze	Genauigkeitsgrenze	Verfügbarkeit
Inline-Inspektion jedes Teils	Innerhalb ½–1 Takt: 240–1,000 Ausschnitte in einem 10–30-s-Takt, 8–100 Ausschnitte pro Sekunde	Durchschlüpfe nicht schlechter als das heutige »AOI plus Menschen« (Vertragsbeispiel <0.5%)	Fail open, nie die Linie stoppen; Ziel ≥99.9% (keine öffentliche SLA)
Inline-Zweitmeinung nur auf AOI-gekennzeichneten Ausschnitten	Kennzeichnungen (meist ≤5–20 pro Platine) innerhalb ½ Takt: 0.5–3 s pro Ausschnitt	Gleich; auf den eigenen Bildern der Linie trainiert	Fällt auf menschliche Nachprüfung zurück
AI-Nachprüfung von Fehlalarmen	2–5 s pro Aufruf, mehr als zehnmal schneller als die 42–55 s einer Person	Durchschlüpfe nicht schlechter als ein menschlicher Prüfer	Degradation fügt Menschen hinzu, stoppt nie die Linie
Robotersteuerung (VLA)	100–300 ms pro Aktions-Chunk mit Echtzeit-Chunking, 10–50 Hz Ausgabe	≥99% Aufgabenerfolg, um programmierte Bewegung zu schlagen	Sicherheitssystem unabhängig vom Modell
Roboterplanung und Fertigkeitswahl	1–3 s bei Bedarf in jedem Zyklus; ≤1 min nur bei Ausnahmen oder Produktwechseln	Ausgabe muss verifizierbarer Code oder eine Fertigkeits-ID sein	Fällt auf das feste Programm zurück

Einsatz	Geschwindigkeitsgrenze	Genauigkeitsgrenze	Verfügbarkeit
Q&A (SOPs, MES-Abfragen)	Erstes Token idealerweise ≤1–2 s, höchstens 10 s; ≥15–20 Tokens/s pro Nutzer; 5–20 gleichzeitige Nutzer	Genaue Zitate; treu über langen Kontext	Büro-IT-Qualität (≈99–99.5%)
Analyse und Ursachenberichte	Minuten sind ok (Batch)	Höchste; treu über langen Kontext	Büro-IT-Qualität

Physical Intelligence maß 97 ms Inferenz für eine VLA der 3B-Klasse (etwa 108–139 ms mit Netzwerk und Vorverarbeitung), stellte fest, dass Echtzeit-Chunking den Durchsatz bis zu 200 ms zusätzlicher Verzögerung unverändert lässt, und erledigte feine Aufgaben wie das Anzünden eines Streichholzes oder das Einstecken eines Netzwerkkabels bei über 300 ms (Physical Intelligence) vendor research ; Interaktionsforschung setzt die Grenze für das Halten eines Gedankenfadens bei einer Sekunde und für Aufmerksamkeit bei zehn (NN/g). Genauigkeit ist der schwierige Teil: Das offene Qwen3-VL-8B erreicht im Schnitt nur 56.8 Zero-Shot über fünf industrielle Anomaliedatensätze (IndusAgent) single source . Keine Hardware rettet ein Vision-Language-Modell, das nicht auf den eigenen Bildern der Linie abgestimmt wurde.

FIG 07 **Geschwindigkeitsgrenzen nach Einsatzzweck** ***Die Inline-Inspektion misst in Millisekunden, Frage-Antwort in Sekunden, Analyse in Minuten. Menschen brauchen 42–55 s pro Alarm.***



Quelle: Zusammenfassung des Autors (Kapitel 1–2).

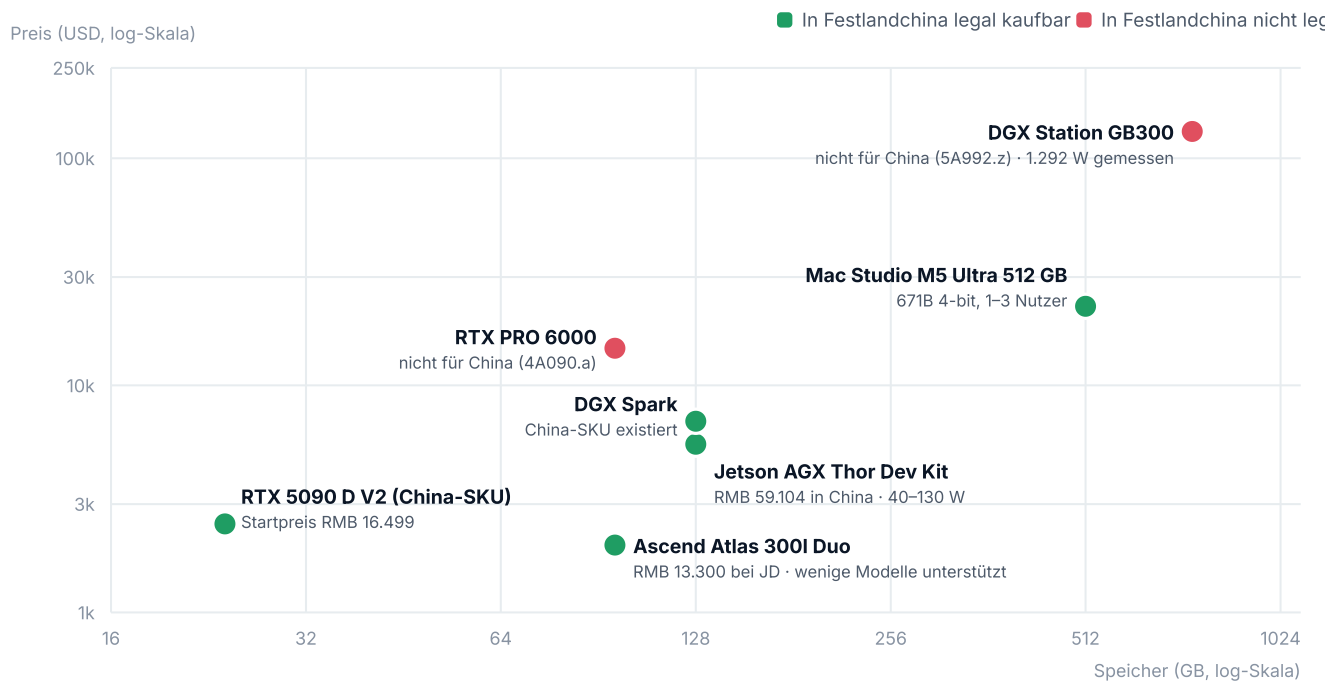
Hardware: Thor für die Station, ein Mac Studio oder inländisches Gerät für die Werks-Q&A, und kein DGX Station auf dem Festland

Gerät	Speicher und Bandbreite	Preis (November 2026)	Wofür es gut ist	Festlandchina
Jetson AGX Thor Dev Kit / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 W	$3,499 zum Start → $5,499 (Juli 2026); T5000 $4,999, T4000 $2,999; RS China Dev Kit 59,104 RMB inkl. Steuer	2–8B-Vision-Modelle zur Nachprüfung von Ausschnitten; 3B-Klassen-VLAs in Echtzeit	● Ja (5A992.c, keine Lizenz für zivile Werke)

Gerät	Speicher und Bandbreite	Preis (November 2026)	Wofür es gut ist	Festlandchina
DGX Spark	128 GB, 273 GB/s	$3,999 → $4,699 → $6,950; das same-chip ASUS GX10 China-Modell 32,999 RMB auf JD (Okt 2025)	Multi-User-8–35B-MoE	● Ja (5A992.c, China-Modell, autorisierte Kanäle)
RTX PRO 6000	96 GB GDDR7	≈$8,565 → $13,250 (Juni 2026) → ≈$16,005 durchschnittliches eBay-Angebot im November	Multi-User-27–35B-FP8; 120B-Klassen-MoE in 4-Bit	● Nein (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB Unified, 1.2 TB/s; 480 W max. Dauerlast	Ab $22,299 in den USA; ≈173–186k RMB geschätzt in China	DeepSeek 671B in 4-Bit für 1–3 Nutzer	● Ja
DGX Station GB300	748 GB (252 GB HBM3e bei 7.1 TB/s + 496 GB LPDDR5X bei 396 GB/s); 1,292 W an der GPU gemessen	$123–140k Straßenpreis im November (≈830–940k RMB vor Zoll); $95–117k im August	Hohe Gleichzeitigkeit auf 120–235B-MoE, das in HBM passt; nur wenige Nutzer auf 433–743B-Modellen	● Nein (5A992.z, TPP 58,484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	13,300 RMB von einem JD-Händler	Batch-Nachprüfung mit geringem Volumen; wenige unterstützte Modelle	● Ja (inländisch)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Nicht ermittelt; volles DeepSeek benötigt zwei bei W8A8, vier bei BF16	Werksweite Multi-User-Q&A und Analyse	● Ja (inländisch)
RTX 5090 D V2 (China-Modell)	24 GB	16,499 RMB zum Start (Aug 2025), seitdem mit Speicherpreisen gestiegen	8B-Vision-Modelle, kleine VLAs	● Ja

Quellen: (NVIDIA China Thor, NVIDIA Thor launch, CNX on price rises, NVIDIA export classification data, RS China, Implicator, IThome on ASUS GX10, Tom's Hardware RTX PRO 6000, GPU Poet, Apple specifications, TechRadar, IThome on M5 Ultra China pricing, Wccftech, StorageReview, CyberPowerPC, monitors.com, Tom's Hardware DGX Station, vLLM-Ascend support matrix, China Telecom Cloud, Huawei Atlas 800I A2, TweakTown). Festlandpreise für das 512-GB-Mac Studio sind aus US-Konfigurationsaufschlägen extrapoliert und wenig vertrauenswürdig **estimate**.

FIG 08 **On-Prem-Rechenleistung: Speicher, Preis und was in Festlandchina legal kaufbar ist*Exportklassifizierungen bestimmen die Auswahlliste, bevor Benchmarks es tun.***



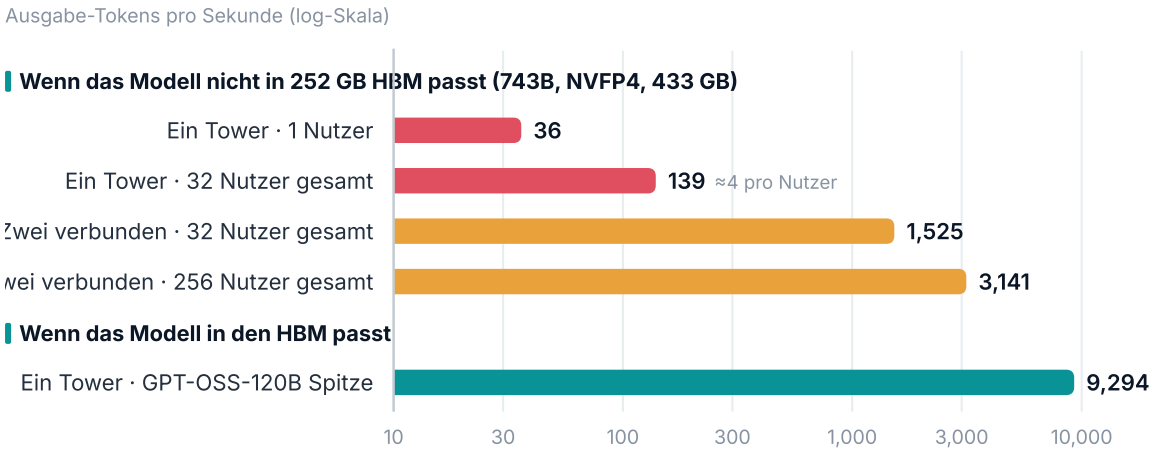
Quelle: NVIDIA-Exportklassifizierungen, Apple, RS China, JD.com und Testpreise (November 2026).

Jetson Thor macht die Bereitstellung auf Stationsebene möglich. Die TensorRT-Edge-LLM-Tests von NVIDIA zeigen Qwen3-VL-8B (NVFP4) mit Prefill von 2,048 Tokens in 104.5 ms und Decoding von 44.4 Tokens/s auf Thor (TensorRT Edge-LLM) vendor claim; für einen AOI-Ausschnitt und eine 20-Token-Antwort sind das etwa 0.5 s mit dem 8B-Modell und 0.2 s mit dem 2B, oder 8.5–21 Ausschnitte pro Sekunde im Batch estimate. Die Ausgabelänge, nicht die Parameterzahl, bestimmt die Latenz, daher ist eingeschränktes Decoding auf ein Klassen-Token oder kurzes JSON Teil des Designs. GR00T N1.7 (3B) läuft mit 25.1 Hz auf Thor (Jetson AI Lab) vendor claim. Ein Orin NX braucht etwa 1.8 s pro Ausschnitt, und ein Atlas 300I Duo hat ein p95 über 23 s, nur für Batch-Arbeit geeignet (CSDN) single source. Auf dem Festland sind NVIDIA-Produkte mit 40–60% über dem US-Listenpreis zu kalkulieren estimate.

Ein 512-GB-Mac Studio macht ein vollwertiges Modell auf einer Maschine möglich, für wenige Nutzer. Das M5 Ultra (August 2026) bietet bis zu 512 GB bei 1.2 TB/s ab etwa $22,299 (Apple, TechRadar); DeepSeek 671B in 4-Bit (≈380–404 GB) decodiert 17–20 Tokens/s auf einem M3 Ultra (Pinggy) verified und geschätzt 25–29 auf dem M5 Ultra. Lange Prompts sind die Schwachstelle: Ein 10,000-Token-Retrieval-Prompt dauert grob 17–70 s estimate, jenseits der 10-Sekunden-Interaktionsgrenze.

DGX Station ist die stärkste Einzelmaschine, aber auf dem Festland nicht verfügbar und fünf- bis sechsmal so teuer wie ein Mac. Das ausgelieferte GB300 hat **748 GB**, von denen nur 252 GB schnelles HBM sind (Wccftech, Tom's Hardware, Aug 2026) verified. Unabhängige Tests zeigen etwa 9,300 Ausgabe-Tokens/s auf GPT-OSS-120B, das in HBM passt, aber ein 743B-Modell, das in LPDDR5X überläuft, gab 32 Nutzern insgesamt nur 139 Tokens/s, etwa 4 je Nutzer; zwei verbundene Einheiten erreichten 1,525 (StorageReview, Aug 2026, StorageReview, Oct 2026) single source. Die Straßenpreise stiegen bis November auf $123–140k (CyberPowerPC, monitors.com). Auf dem Festland sind die realistischen Optionen für denselben Job das Mac Studio, CPU+GPU-Server im KTransformers-Stil (KTransformers) vendor test oder inländische Ascend-Server.

FIG 09 **Eine große Workstation ergibt noch keinen gemeinsamen Assistenten** ***Wenn das Modell nicht in den High-Bandwidth-Speicher passt, erhalten 32 gleichzeitige Nutzer jeweils etwa 4 Token pro Sekunde.***



Quelle: StorageReview (Aug. und Okt. 2026); ServeTheHome. Unabhängige Tests aus Einzelquellen.

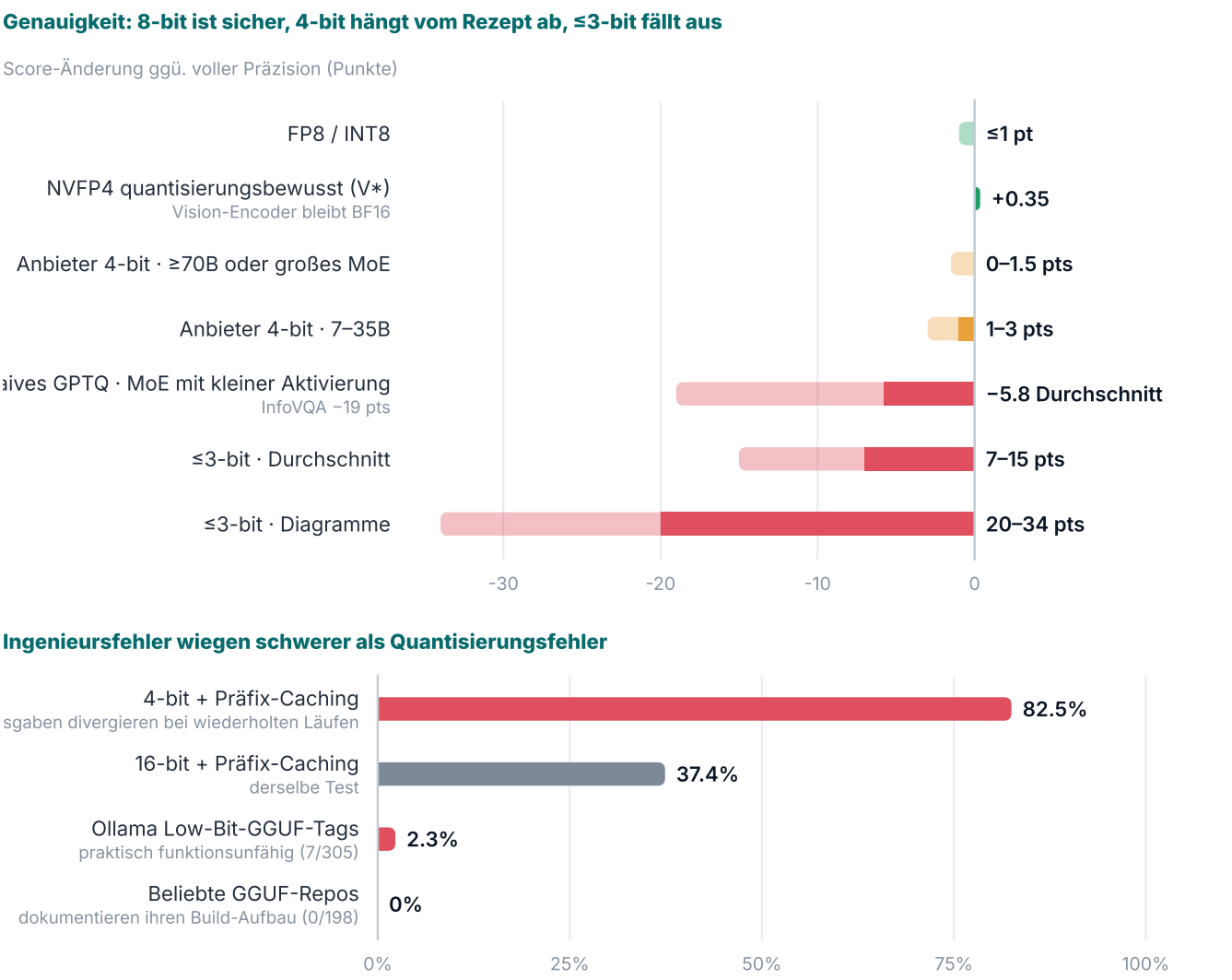
Speicherknappheit bewegte dieselben Teile innerhalb eines Jahres um 20–100% (DGX Spark stieg von $3,999 auf $6,950 (TechRadar)), daher sollten Ausrüstungsangebote indiziert statt fixiert werden.

Inländische Beschleuniger führen vollwertige Modelle aus; Softwarewartung ist der Engpass. Ascend 910B hat 64 GB HBM, aber vLLM-Ascend markiert DeepSeek V3/R1 als »nicht mehr gepflegt« und unterstützt nur wenige Modelle auf kleineren Karten (vLLM-Ascend) single source, primary, und jedes neue Modell braucht ein bis zwei Monate zur Anpassung (Caijing/Anue); liefern Sie eine eingefrorene, validierte Modellversion aus. Das Ausliefern von vollem DeepSeek 671B an 20–100 Nutzer benötigt mindestens zwei Atlas-800I-A2-Server mit acht Karten (China Telecom Cloud) verified, doch öffentliche Ausschreibungen für LLM-Appliances in China beliefen sich im Januar–September 2025 auf nur 27.5 Millionen RMB (C114) single source: Fabriken zahlen für Integration und Service, nicht für eine Kiste Rechenleistung.

Quantisierung: 8-Bit ist sicher, 4-Bit hängt vom Rezept ab, 3-Bit und darunter ist ausgeschlossen

8-Bit (FP8 oder INT8) ist für Vision-Language-Modelle im Wesentlichen verlustfrei (Red Hat, Hugging Face) verified. Bei 4-Bit hängt der Verlust von Größe und Rezept ab: Qwens 72B-AWQ-Build verliert unter einem Punkt auf MMMU, während der 7B 2.8 verliert (Qwen AWQ) verified, und detailintensive Aufgaben fallen weiter (naives 4-Bit-GPTQ kostet Qwen3-VL-30B-A3B 19 Punkte auf InfoVQA (VEQ)). Bei 3-Bit und darunter liegt der durchschnittliche Verlust bei 7–15 Punkten und Diagramme verlieren 20–34 (VLMQ) verified; jenseits von 64k Tokens verliert 4-Bit bis zu 59%, schlimmer bei nichtenglischer Eingabe (arXiv 2505.20276). Halten Sie den Vision-Encoder bei FP16/BF16.

FIG 10 **Quantisierung: 8-Bit ist sicher, 4-Bit hängt vom Rezept ab, 3-Bit und darunter fällt aus** ***Ingenieursfehler im Serving-Stack kosten mehr Genauigkeit als die Quantisierung selbst.***



Quelle: NVIDIA-Nemotron-Modellkarte, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 u. a.

Zwei Warnungen zählen mehr. Erstens: **Zum 11. November 2026 hat keine Studie und keine Modellkarte die Quantisierung auf PCB/SMT-Defektbildern gemessen**, also ist auf dem Golden Set erneut zu validieren. Zweitens sind Serving-Stack-Fehler oft größer als der Quantisierungsfehler: Dasselbe Modell erzielte unter Ollama etwa 30 Punkte weniger als unter der Referenzimplementierung (SiliconBench); 7 der 305 offiziellen Code-Modell-Tags von Ollama waren im Wesentlichen kaputt, obwohl die Dateien intakt waren (arXiv 2609.05881); keines von 198 beliebten GGUF-Repositories dokumentierte seinen Build (arXiv 2610.19498); und mit Prefix-Cache-Wiederverwendung divergierten 82.5% der Episoden bei 4-Bit gegenüber 37.4% bei 16-Bit (arXiv 2609.04748) single sources. Quantisierungsbewusstes 4-Bit vom Anbieter ist bei Detailaufgaben nahezu verlustfrei (NVIDIA Nemotron 3.5) vendor claim, aber 4-Bit-VLAs verlieren unter Störung 1.5–4.3 Punkte und laufen auf Thor langsamer (LoopQ) single source. Behandeln Sie ein quantisiertes Modell wie eine neue Messeinrichtung: Gewichts-Hashes festpinnen, den Build dokumentieren, Prefix-Caching zurücksetzen und nach jeder Änderung von Präzision, Engine oder Template erneut validieren.

Präzision	≥70B dicht oder ≥100B MoE	7–35B	≤4B oder detailintensive Aufgaben	Urteil
FP8 / INT8	Sicher (≤1 Punkt)	Sicher	Sicher; der Standard für die Inspektion	Nutzen
4-Bit AWQ/GPTQ (Anbieterrezept, Vision-Encoder FP16)	0–1.5 Punkte	1–3 Punkte; naives GPTQ auf small-active MoE −5.8 im Schnitt, −19 auf InfoVQA	2–3 Punkte; Detailaufgaben zuerst validieren	Bedingt
NVFP4 (Anbieterrezept)	≈0.3–1 Punkt unter FP8 (Anbieterdaten)	Quantisierungsbewusste Builds nahezu verlustfrei auf V*, OCRBench	Keine öffentlichen Daten	Bedingt (quantisierungsbewusste Builds bevorzugen)
GGUF Q4_K_M	Ungetestet	Hohe Varianz bei Domänenaufgaben	Nicht für kleine Texte oder Defekte	Nicht für Detailbeurteilung
GGUF Q5_K_M und höher	Vermutlich sicher	Nahezu verlustfrei in zwei Tests	Die Untergrenze für Details auf llama.cpp	Nutzen
≤3-Bit (einschließlich 1T-Modelle, auf 1.8–2 Bits auf einer Maschine gequetscht)	Nicht für Entscheidungen	Nein	Nein	Nein

Welche Konfiguration für welchen Job

Konfiguration	Inline-Inspektion	AI-Nachprüfung von Fehlalarmen	Roboterplanung und -steuerung	Q&A und Analyse
Dedizierte Detektoren oder Anomaliemodelle (YOLO, RF-DETR, PatchCore-Klasse), INT8/FP16, jede NPU- oder GPU-Edge-Box	● Ja: der Hauptentscheider, 2–12 ms pro Ausschnitt	● Ja (Klassifikator, auf die Linie abgestimmt)	● Nur Wahrnehmung	—
2–4B-Vision-Modelle, NVFP4/AWQ/FP8 mit FP16-Vision-Encoder, auf Thor/T4000 oder DGX Spark	● Bedingt: AOI-gekennzeichnete Ausschnitte bei 0.2–0.3 s je Stück nachprüfen	● Bedingt, unter 1 s nach Abstimmung	● VLA-Backbone (2–3B-VLA bei ≈40–130 ms pro Chunk)	● Schwach
7–9B-Vision-Modelle (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 oder Anbieter-4-Bit, auf Thor, Spark oder einer 24–32-GB-Karte	● Bedingt: ≈0.5 s pro gekennzeichnetem Ausschnitt	● Ja: 0.5–1 s pro Aufruf, bestes Preis-Leistungs-Verhältnis	● Planung und Fertigkeitswahl (1–3 s), nicht im Regelkreis	● Grundlegende SOP-Q&A für einen oder wenige Nutzer
26–35B MoE (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● Nein (zu schwer)	● Ja, als Erklärer oder für schwere Fälle	● Planung	● Ja (8 Nutzer auf einem Spark, ≈0.1 s erstes Token)
106–241B MoE in 4-Bit auf einem Mac Studio, RTX PRO 6000 oder Spark	● Nein	● Nein	● Offline-Planung	● Ja, für einen oder wenige Nutzer

Konfiguration	Inline-Inspektion	AI-Nachprüfung von Fehlalarmen	Roboterplanung und -steuerung	Q&A und Analyse
DeepSeek 671B oder Kimi K2.x 1T in 4-Bit auf einem 512-GB-Mac, KTransformers-Server oder DGX Station	● Nein	● Nein	● Nur Offline-Programmgenerierung	● Bedingt: 1–3 Nutzer und kurze Prompts; lange Prompts dauern 16–170 s
Alles bei ≤3-Bit; GGUF Q4_K_M für Detailarbeit kleiner Modelle; Ollama für viele Nutzer	● Nein	● Nein	● Nein	● Nein

Quellen: die obigen TensorRT-Edge-LLM-, Jetson-AI-Lab-, Apple-ML-Research- und KTransformers-Daten sowie (DGX Spark tests, RTX PRO 6000 tests, Hardware Corner on Atlas, Ultralytics YOLO26, RF-DETR) **community and vendor tests, mostly single source** .

Zu den beiden üblichen Sorgen also: **Es muss nicht langsam sein**, wenn das Modell zum Job passt (kleine und mittlere Vision-Modelle auf Thor oder Spark erfüllen jedes Inline- und Nachprüfungsbudget; 671B–1T-Modelle bleiben von der Linie fern), und **Quantisierung muss es nicht falsch machen**, mit FP8/INT8 als Standard oder Anbieterrezept-4-Bit über 7B, auf dem Golden Set validiert. Die größeren Risiken sind die rund 57-Punkte-Zero-Shot-Industriegenauigkeit allgemeiner Modelle und Serving-Stack-Fehler. Geben Sie Inspektionsanfragen feste Batchgrößen und eine eigene Warteschlange (Temperatur 0 ist unter Batching nicht bit-reproduzierbar (vLLM documentation)), halten Sie einen N+1-Hot-Standby bereit und lassen Sie jeden Schritt an der Linie fail open.

Lehrer und Schüler

03 Lehrer und Schüler: Frontier-Modelle können Inspektionsdaten vorab beschriften; bei Roboterarmen sind die wahren Lehrer weiterhin Menschen

Lehrerkandidaten November 2026: Generische Bilderkennung verdoppelt, Industriedefekt-Werte fehlen

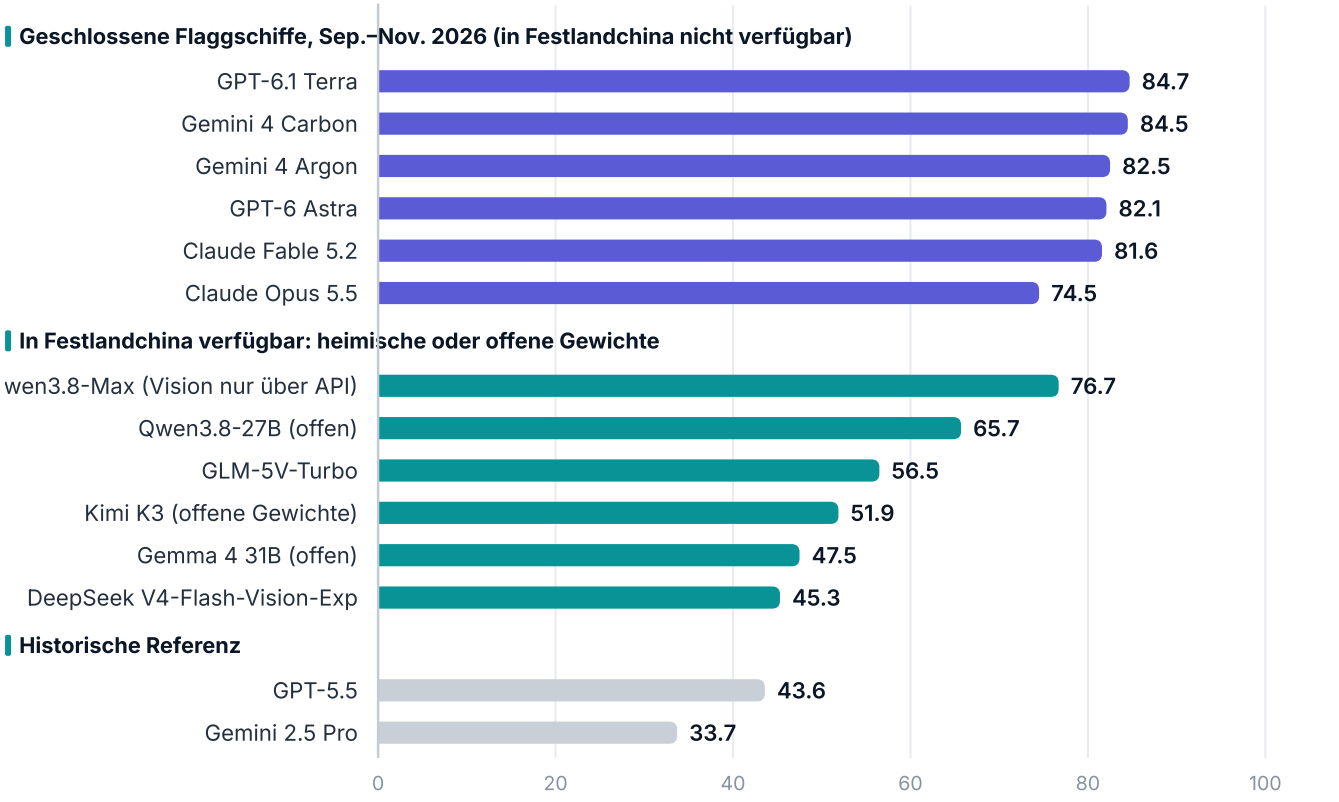
Jedes geschlossene Flaggschiff, das im November 2026 als Lehrer wirken könnte, nimmt Bilder an: Claude Fable 5.2 (allgemein verfügbar seit 9. November, $8/$40 pro Million Tokens) und Opus 5.5 ($4/$20) (Anthropic, Claude models overview) verified; GPT-6 Astra ($10/$50) und GPT-6.1 Terra ($1/$5) (OpenAI pricing); und Gemini 4 Argon und Carbon, stufenweise geöffnet (Google Carbon). In Festlandchina nutzbar sind Qwen3.8-Max (Vision via API) und offene Modelle wie Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp und Gemma 4 (Roboflow blog, Hugging Face).

Die entscheidende Lücke muss klar benannt werden: Keines dieser Modelle hat einen veröffentlichten Wert auf MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB oder einem PCB/SMT-Datensatz. MMAD hat weiterhin kein Leaderboard (MMAD GitHub), und die beiden neuesten Industrieanomalie-Papiere (August–September 2026) verwenden weiterhin Gemini 2.5 Pro und Claude Sonnet 4 als ihre neuesten allgemeinen Baselines (InspectorGPT, Anomaly-LR). Eine arXiv-Volltextsuche am 11. November ergab null Treffer für »Fable 5.2« oder »Gemini 4 Carbon« in diesem Kontext (arXiv API) single source, absence of evidence. Die nächstgelegenen Ersatzwerte:

- Das neueste Frontier-Modell mit einem MMAD-Wert ist GPT-5.5: 69.95% bei »Defekt oder nicht« und 76.20% im Schnitt über sieben Aufgaben auf einer 500-Bilder-Teilmenge, gegenüber 95.24% und 86.65% für menschliche Experten (AnomalyClaw) single source; GPT-4o erzielte 2024 68.63% und 74.92% (MMAD, ICLR 2025), etwa 1.3 Punkte Fortschritt in zwei Jahren.
- Auf LDU-Bench, echten SEM-Review-Bildern aus der IC-Lithografie und dem einzigen Elektronik-Benchmark mit 2026er-Modellen, ist die Gut-Schlecht-Triage ordentlich (Makro-F1 0.932 für GPT-5.4, 0.884 für GLM-5V-Turbo, 0.876 für Claude Opus 4.6, 0.504 für Gemma 4 31B), aber die Klassifikation von 11 Defektmorphologien erreicht nur 0.20–0.41 (LDU-Bench) single source.
- Bei Roboflows unabhängigen Vision-Evals (aktualisiert am 10. November 2026) liegt die Objekterkennungs-mAP@50 bei 84.5 für Gemini 4 Carbon, 84.7 für GPT-6.1 Terra und 81.6 für Fable 5.2, gegenüber 33.7, 43.6 und 38.6 für Gemini 2.5 Pro, GPT-5.5 und Claude Opus 4.8 (Roboflow Vision Evals) single source, independent evaluator, private datasets.

FIG 11 **Lehrerkandidaten: Die generischen Scores haben sich in einer Generation verdoppelt*Doch keines der 2026er-Flaggschiffe hat einen veröffentlichten Score in einem Industrielle-Defekte-Benchmark.***

Roboflow generische Detektion mAP@50



Industrielle-Defekte-Benchmarks (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): null veröffentlichte Scores für diese 2026er-F
Generische Detektionsscores haben sich in einer Generation verdoppelt, ersetzen aber keine Defektbeurteilung; auf dem eigenen Goldstandard-Set testen

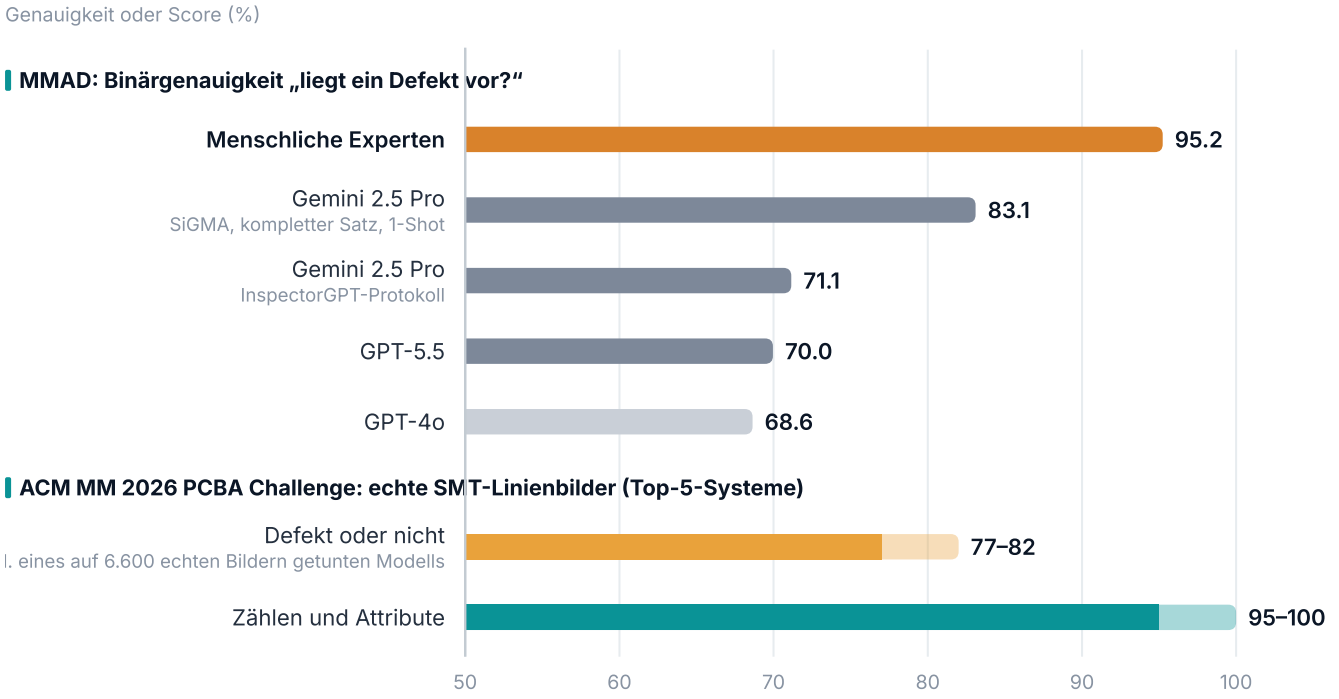
Quelle: Roboflow Vision Evals; Modell-Release-Notes; MMAD, MVTec AD 2 und weitere Benchmark-Seiten (Stand 11. Nov. 2026).

Lehrerkandidat (Nov 2026)	Preis pro M Tokens (in/out)	Roboflow-Erkennung mAP@50 / Zählung	Industriedefekt-Benchmarks	In Festlandchina nutzbar
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Keine	Nein
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Keine	Nein
GPT-6 Astra	$10 / $50	82.1 / 80.2	Keine	Nein
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Keine	Nein
Gemini 4 Carbon	$2 / $10 (Aktionspreis)	84.5 / 90.1	Keine	Nein
Qwen3.8-Max (Vision nur via API)	$2 / $6	76.7 / 81.1	Keine	Ja
Qwen3.8-27B (offen)	Lokal	65.7 / 64.9	Keine	Ja
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench Triage 0.884	Ja
Gemma 4 31B (offen)	Lokal	47.5 / 51.4	LDU-Bench Triage 0.504	Ja (lokale Gewichte)
Referenz: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD binär 69.95% / 68.63%	Nein

Lehrerkandidat (Nov 2026)	Preis pro M Tokens (in/out)	Roboflow-Erkennung mAP@50 / Zählung	Industriedefekt-Benchmarks	In Festlandchina nutzbar
Menschliche Experten	—	—	MMAD binär 95.24%	—

Neuere Papiere testen weiterhin die vorherige Generation, weisen aber in dieselbe Richtung. Auf dem vollen MMAD-Set erzielen menschliche Experten 95.2% bei Defekt-oder-nicht und Gemini 2.5 Pro 83.1% (SiGMA), doch dasselbe Modell erzielte in einem anderen Papier nur 71.14% (InspectorGPT): Das Testprotokoll bewegt den Wert ebenso stark wie eine Modellgeneration. In der CVPR 2026 VAND 4.0 Challenge erreichten handelsübliche Vision-Language-Modelle höchstens 61.88 AP gegenüber 90.84 für überwachte Methoden, und die Organisatoren schlossen, dass sie »feinabgestimmte Klassifikatoren derzeit nicht ersetzen können« (VAND 4.0) single source. In der ACM MM 2026 PCBA-Challenge erzielten die fünf besten 76.92–82.47 bei Defekt-oder-nicht, während Komponententyp und Pin-Zählung 95–99.5 erreichten (arXiv 2609.21276) single source: Erkennen ist viel einfacher als Beurteilen.

FIG 12 **„Gibt es einen Defekt?“ bleibt für allgemeine Modelle schwer** ***Das beste allgemeine Modell liegt auf MMAD mehr als 12 Punkte hinter menschlichen Experten; angepasste Systeme schneiden auf echten Linienbildern besser ab.***



Quelle: MMAD; SiGMA; InspectorGPT; ACM MM 2026 PCBA Challenge.

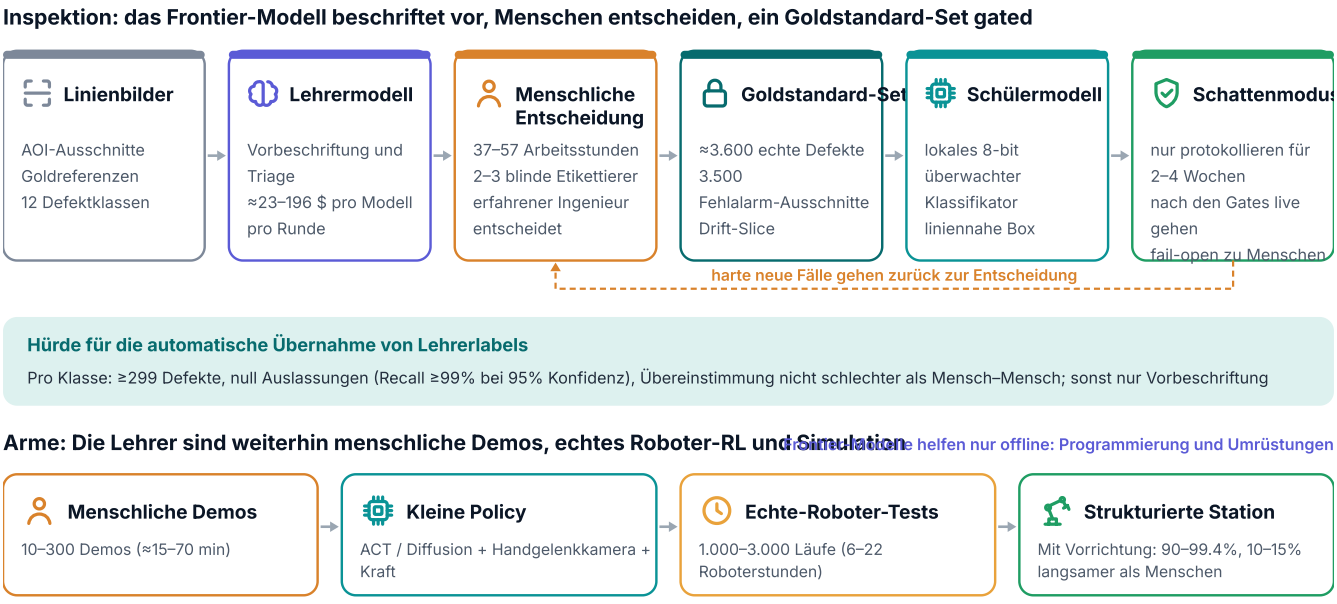
Daraus folgen drei Schlussfolgerungen. Sehen und Zählen verdoppelten sich in einer Generation, aber das Beurteilen von Anomalien bewegt sich langsam, hängt stark vom Protokoll ab und bringt systematische Verzerrungen mit; wie viel die neuen Modelle gewinnen, lässt sich nur auf den eigenen Bildern einer Fabrik messen. Innerhalb ihrer Trainingsverteilung schlagen feinabgestimmte 7B-Spezialisten weiterhin allgemeine Modelle (Anomaly-LR, UniP-CB). Und echte Liniendaten lassen die Werte einbrechen: Auf bestückten Elektronik-Panels fiel die bildbezogene AUROC von 97–98 auf MVTec AD auf 31.5–78.6 (arXiv 2606.01992) single source. Für Festlandunternehmen liegt der stärkste verfügbare Lehrer, Qwen3.8-Max, etwa 8 Punkte unter der globalen Spitze.

Dutzende oder Hunderte Runden laufen lassen: Geld ist nicht das Problem, die Label-Qualität ist es

Geld ist nicht das Problem. Zu Listenpreisen vom November 2026 kosten 1,000 Ein-Megapixel-Bilder etwa $3 (GPT-6.1 Terra) bis $32 (GPT-6 Astra) pro Runde, und hundert Runden über 100,000 Bilder etwa $32,000–135,000 vor Mengenrabatten estimate (Claude vision docs, OpenAI vision guide, Roboflow Vision Evals). Für einfache Labels kostet eine Person etwa so viel wie Terra; der Vorteil des Lehrers ist Geschwindigkeit und Skala, nicht Preis. Die NDAs der Kunden halten Produktbilder meist im Werk, daher wird das Golden Set vor Ort beschriftet.

Das eigentliche Hindernis ist Qualität, und mehr Runden beheben keine systematischen Fehler. Roboflow führt jede Aufgabe dreimal aus, und die Werte bewegen sich nur um ±0.1–1.4 Punkte (Roboflow), sodass ein Modell seine Fehler meist wiederholt und Mehrheitsvoting es nicht retten kann. Runden helfen nur, wenn jede Information hinzufügt: Ein oder zwei gute Referenzbilder bringen etwa 3–4 Punkte, mehr bringen nichts (AD-Copilot); wenn ein Modell über mehrere Züge gegen eine gute Referenz argumentiert, kommen 6.23 Punkte domänenübergreifende AUROC für GPT-5.5 und 2.95 auf MMAD hinzu, mit Höhepunkt nach etwa drei Aufrufen (AnomalyClaw) single sources. »Ein auf Lehrer-Labels trainierter Schüler nähert sich menschlichen Labels« wurde nur auf natürlichen Bildern gezeigt: Ein ResNet-101, trainiert auf Qwen3-VL-Labels von COCO und Objects365, erreichte 95.7–99.4% des Mensch-Label-Ergebnisses, und die Autoren sagen, Domänendaten seien ungetestet (arXiv 2602.20972) single source.

FIG 13 **Zwei Lehrer-Schüler-Pipelines** ***Bei der Inspektion beschriftet ein Frontier-Modell vor und Menschen entscheiden; bei Roboterarmen sind die eigentlichen Lehrer weiterhin menschliche Demonstrationen.***



Quelle: Zusammenfassung des Autors (Kapitel 3); Kosten und Stichprobengrößen sind Schätzungen.

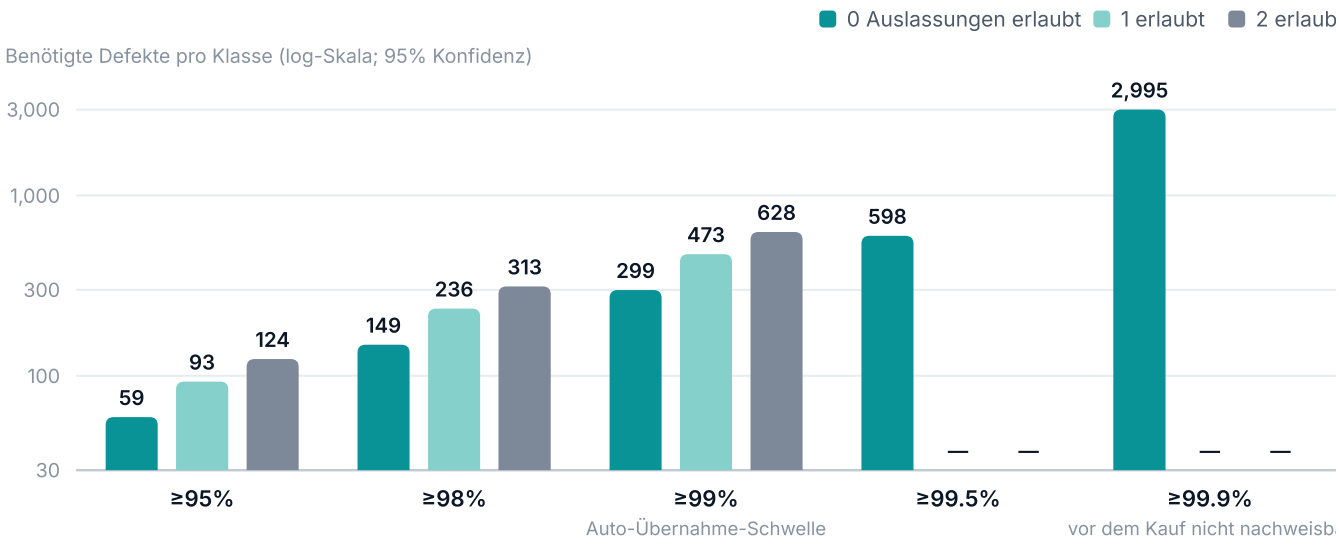
Ein praktikables Protokoll estimate:

1. **Sammeln Sie das Golden Set aus dem eigenen AOI-Archiv der Fabrik und zurückbehaltenen NG-Platinen**, geschichtet in 12 echte Defektklassen (fehlend, Versatz, Tombstone, Polarität, falsches Teil, Brücke/Kurzschluss, unzureichendes Lot, übermäßiges Lot, Lotkugel, abgehobene Leitung/offen, beschädigtes Teil, Fremdmaterial) und vier Fehlalarm-Ursachen (Schatten, Flussmittel, Siebdruck, Toleranz), plus einen »Drift-Slice« von mindestens 300 Defekten aus einer zweiten Linie oder anderer Beleuchtung, Platinenfarbe oder Teilenummern.

2. **Beschriften Sie blind mit zwei oder drei IPC-A-610-geschulten Prüfern**, wobei ein leitender Ingenieur entscheidet und Röntgenprüfung, elektrische Prüfung oder Querschliffe mehrdeutige Proben klären. Akzeptieren Sie eine Klassendefinition nur, wenn die Prüferübereinstimmung κ ≥0.75 erreicht (≥0.90 für kritische Klassen); die menschliche Übereinstimmung bei 11 Elektronikdefektmorphologien in LDU-Bench betrug nur 0.727, daher müssen einige Klassen zusammengelegt oder neu definiert werden. Eine AIAG-artige Attributvereinbarung (Effektivität ≥90%, Übersehen ≤2%, Fehlalarm ≤5%, κ >0.75; >0.90 für die Freigabe) ist ein vernünftiges Minimum (Fabrico, iFactory).
3. **Bemessen Sie es mit exakten binomialen (Clopper–Pearson) einseitigen 95%-Grenzen** (binomial confidence intervals, rule of three:

Zu beweisender Recall	0 Übersehene erlaubt	1 erlaubt	2 erlaubt
≥95%	59 Defekte	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2,995	—	—

FIG 14 **Wie groß das Goldstandard-Set sein muss*Der Nachweis von 99% Recall ohne Ausfälle erfordert 299 Defekte pro Klasse; 99,9% lassen sich vor dem Kauf nicht nachweisen.***



Quelle: Exakte binomiale Konfidenzintervalle (95%).

Sechzig Defekte pro Klasse reichen für einen Bestanden/Durchgefallen-Screen; 300 werden gebraucht, um über die automatische Übernahme zu entscheiden. Der volle Satz, etwa 3,600 Defekte plus 3,500 Fehlalarm-Ausschnitte, braucht etwa 37–57 Arbeitsstunden zum Beschriften und Beurteilen, und das Durchlaufen eines Lehrerkandidaten kostet $23–196; die wahren Kosten sind zwei bis sechs Wochen Sammlung, bestimmt durch die seltenste Klasse estimate. Senden Sie Einzelkomponenten-AOI-Ausschnitte, nicht ganze Platinen, mit dem Ziel von mindestens 42–48 Pixeln Durchmesser (arXiv 2610.09313).

1. **Übernehmen Sie Lehrer-Labels automatisch nur für eine Klasse mit null Übersehenen bei ≥299 Golden-Set-Defekten**, Modell-Wahrheit-Übereinstimmung nicht schlechter als Mensch-Mensch-Übereinstimmung und mindestens 50% der Fehlalarme bereinigt. Kritische Klassen wie Kurzschlüsse, fehlende Teile, Polarität und falsche Teile benötigen 99.5–99.9% (598–2,995 Defekte); andernfalls geht alles, was die AI nicht mit hoher

Konfidenz bereinigen kann, an einen Menschen. Eine Durchschlupfrate unter 0.1% der Defekte kann vor dem Kauf nicht bewiesen werden (2,995 Defekte mit null Übersehenen; 0.01% würde 29,956 erfordern), daher behält die Linie menschliche Stichproben der von AI bereinigten Aufrufe, nachgelagertes ICT/FCT als Rückhalt und SPC bei Durchschlüpfen. Verifizierte Labels werden dann in einen konventionellen Detektor oder Klassifikator destilliert und bei jedem neuen Produkt erneut validiert.

Auch Bedingungen und Datenregeln entscheiden über den Lehrer. Anthropic verbietet die Nutzung seiner Dienste zum Trainieren konkurrierender Modelle (Anthropic commercial terms); OpenAI erlaubt, mit den Ausgaben interne Klassifikatoren zu trainieren, die nicht an Dritte verkauft werden (OpenAI business terms); DeepSeek erlaubt Destillation ausdrücklich (DeepSeek terms) primary terms. Keines definiert ein »konkurrierendes Modell«, daher braucht die kommerzielle Nutzung eine rechtliche Prüfung, und die Datenregeln der Kunden halten Bilder meist vor Ort. Der sicherste Standard ist ein lokal betriebener Open-Weight-Lehrer, mit Cloud-Frontier-Modellen nur, wenn der Kunde es schriftlich erlaubt.

Inspektions-Schüler: klein und schnell; die Engpässe sind Optik, Daten und Validierung

Der Schüler ist nicht der Engpass: YOLO26n braucht 3.80 ms pro Bild in INT8 auf einem Orin Nano Super für $249–399 (Ultralytics) vendor test, NVIDIA meldet 98.51% auf drei PCB-Defektklassen nach der Anpassung eines Foundation-Modells auf etwa 600 Bildern (NVIDIA) vendor claim, und synthetische Defekte bringen nur 0.9–2.5 mAP-Punkte (arXiv 2605.04635). Die Beschränkungen sind Optik, Liniendaten, Validierung und Lizenzen: Ultralytics YOLO ist AGPL-3.0 oder unternehmenslizenziert (Ultralytics), daher brauchen geschlossene Produkte eine Lizenz oder einen Apache-2.0-Detektor wie RF-DETR.

Roboter-Schüler: Strukturierte Stationen sind nahe der Produktionsreife; flexible und dynamische Arbeit ist es nicht

Im November 2026 gibt es drei Arten von Roboter-Lehrern, und nur eine kann ein lokaler Schüler werden: offene, feinabstimmbare VLAs (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Geschlossene Robotermodelle (π0.6/π0.7 über eine Beta-API, Gemini Robotics 2 für vertrauenswürdige Tester) können nicht vor Ort bereitgestellt werden (The AI Insider, DeepMind). Frontier-Allgemeinmodelle sind langsam und schlecht bei feiner Manipulation: Im Test von Artificial Analysis erzielten sie 46–69 mit 28–45% Grifferfolg und etwa 30 Minuten pro Aufgabe, »viel zu langsam und teuer für jeden Produktionseinsatz« (AA-Robotics); GPT-6 Astra schaffte einen Präzisionsschlitz nur 2 Mal in 20 (RoboCurve); und das direkte Steuern von Gelenken ergibt 0–5.5% auf LIBERO bei 0.2–0.4 Hz gegenüber benötigten etwa 83 Hz (Anthropic) verified across three sources. Fable 5.2s 64.9% auf dem simulierten RoboDojo-Benchmark ist ein interner Lauf (Anthropic) vendor claim.

Drittstudien seit Oktober stimmen überein: etwa 33–44% auf LIBERO-10 für die drei führenden Agenten (arXiv 2610.30051); Agenten, die bei 24–39% der Teilaufgaben fälschlich Erfolg melden (arXiv 2611.08209); und 104 von 108 quasistatischen Aufgaben am echten Roboter für Opus 5.5, aber mit 280 Sekunden je Aufgabe und ohne Insertion oder Montage (AMP1) single sources. Frontier-Modelle gehören offline: Programme schreiben, Aufgaben orchestrieren und Erfolgskriterien entwerfen.

Die Schülerqualität hat einen klaren Vorteil: gut innerhalb der Trainingsverteilung, einbrechend außerhalb. π0.7 »übertrifft oft 90%« bei gesehenen Aufgaben und erreicht 60–80% bei ungesehenen (π0.7 paper) vendor research; ContactDP erreicht 97.8% bei trainierten Steckverbindern, aber 28.6–66.7% bei ungesehenen (ContactDP); leichte Kamera- oder Posenänderungen drücken 95% unter 30% in LIBERO-Plus (LIBERO-Plus, CVPR 2026); gelerntes Greifen von bewegten Förderbändern erreicht höchstens 58.89% (ForeTime-VLA) single sources, ein Job, den klassisches Vision-Tracking seit Jahren in der Produktion erledigt.

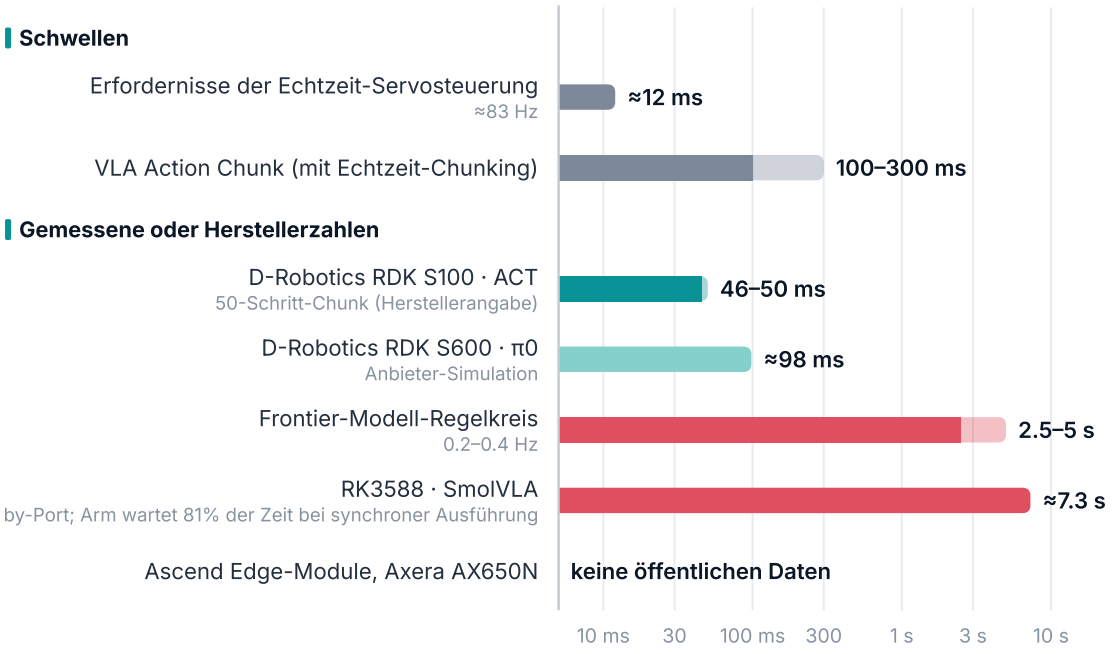
Die Belege nahe der Produktion stammen alle aus strukturierten Stationen, menschlichen Demonstrationen und Korrektur am echten Roboter. Neuromeka behielt eine Zustandsmaschine und vorgelehrte Bewegung und lernte nur Greifen, Kabeleinführen und Löten, jeweils aus unter 20 Minuten Daten: 322 von 324 in Produktion, 12.8% langsamer als ein Mensch (Neuromeka) single source, vendor. Fraunhofer erreichte 92.8% über vier Steckverbinder aus 68

Minuten Demonstrationen (Fraunhofer); Berkeleys HIL-SERL gelang 100 von 100 Mal das Einstecken von RAM, SSD und USB nach 1–2.5 Stunden RL am echten Roboter (HIL-SERL); AgiBot lief 18–20 Sekunden pro Schritt innerhalb eines 20-Sekunden-Takts auf einer Tablet-Testlinie (IThome) vendor claim .

Destillation kauft Geschwindigkeit, nicht Generalisierung: π0.5 läuft mit 0.70 Hz auf Jetson Orin roh und 10.30 Hz nach asynchronem Scheduling (EagleVLA); Schichtdestillation bringt es von etwa 3 Hz auf etwa 10 Hz (Shallow-π) verified . Ein Schüler erbt nur, was der Lehrer abdeckte; jedes neue Produkt braucht weiterhin Aufgabendaten.

FIG 15 **Edge-Policy-Latenz** ***Eine kleine ACT-Policy läuft auf einem Edge-Board in etwa 50 ms; ein auf einen Hobby-Chip portierter VLA braucht 7.3 s.***

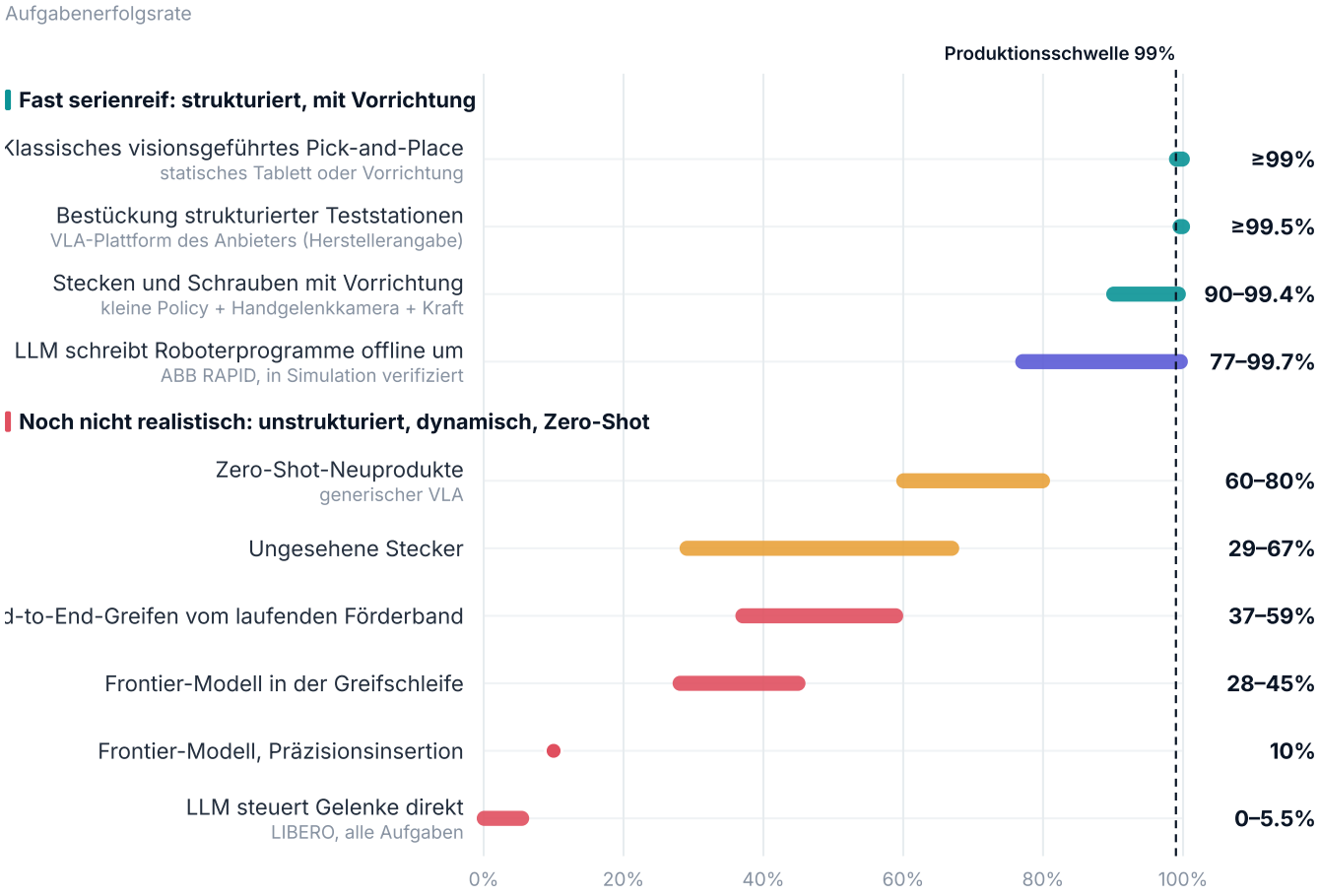
Inferenzlatenz pro Action Chunk oder Regelzyklus (log-Skala)



Quelle: D-Robotics-Dokumentation und -Forum (Herstellerangaben); GitHub-Community-Ports; Anthropic-Forschung.

Auf chinesischen Edge-Chips ist die einzige gemessene gelernte Policy eine ACT-Policy auf D-Robotics' RDK S100 mit 46–50 ms pro 50-Schritt-Chunk, ohne veröffentlichte Erfolgsrate (D-Robotics RDK S docs) vendor claim ; ein Hobby-Port von SmolVLA auf einem RK3588 braucht etwa 7.3 s pro Chunk (GitHub). Quantisierung kauft Geschwindigkeit nicht umsonst: Kein 4-Bit-π0.5-Build war bei Batchgröße 1 schneller als 16-Bit (LoopQ). Demonstrationen sind teuer (Reuters meldete etwa 300 Versuche pro nutzbarer Bewegung für einen Anfänger-Teleoperator (Notebookcheck)), und Validierung kostet mehr: Der Nachweis einer Ausfallrate ≤0.1% bei 95% Konfidenz erfordert etwa 3,000 aufeinanderfolgende Erfolge, 17–22 Roboterstunden estimate . Keine gelernte Manipulationspolicy hat eine Sicherheitszertifizierung.

FIG 16 **Roboter-Schüler: fast serienreif nur bei strukturierten Stationen** ***Stationen mit Vorrichtung erreichen 90–99%+; unstrukturierte, dynamische und Zero-Shot-Aufgaben sind noch nicht realistisch.***



Quelle: In Kapitel 3 zitierte Paper und Herstellerunterlagen; einige Zahlen sind Herstellerangaben.

Elektronikaufgabe	Praktikabler Schüler	Erwarteter Erfolg	Geschwindigkeit vs. Mensch	Kosten pro Produktwechsel	Urteil 2026–2028
High-Mix-Pick-and-Place, Kitting, Maschinenbeschickung (statisches Tray oder Vorrichtung)	Gelernte Wahrnehmung plus programmierte oder geplante Bewegung	≥99% mit klassischer Vision-Führung	Gleich wie konventionelle Stationen	Stunden	**Jetzt machbar**
Fixierte Steckverbinder-, FFC- und Kabeleinführung, Schrauben, Löten	Zustandsmaschine + kleine BC/Diffusion/ACT-Policy + Handgelenkkamera + Kraft/Drehmoment (idealerweise Berührung) + menschliche Korrektur	90–99.4%	10–15% langsamer	10–300 Demos (≈15–70 min) + 1,000–3,000 Validierungsläufe (6–22 Roboterstunden)	**Machbar, F&E-intensiv Monate für die erste Station**
Strukturierte Teststations-Beschickung und OK/NG-Sortierung	Anbieter-VLA-Plattform + digitaler Zwilling + RL am echten Roboter	≥99.5% vendor claim, conflicting	18–20 s pro Schritt, innerhalb eines 20-s-Takts	≤4 h Neutraining + 15 min Kalibrierung	**Auf einer Anbieterplattform machbar**

Elektronikaufgabe	Praktikabler Schüler	Erwarteter Erfolg	Geschwindigkeit vs. Mensch	Kosten pro Produktwechsel	Urteil 2026–2028
Produktwechsel-Programmierung	LLM schreibt oder bearbeitet Roboterprogramme offline, in Simulation verifiziert	77–99.7% bei ABB-RAPID-Umschreibungen	Kein Effekt auf den Takt	Minuten bis Stunden pro Programm	**Als Engineering-Werkzeug machbar**
End-to-End gelerntes Greifen von bewegten Förderbändern	VLA-Schüler	37–59%; ≤37% auf schnellen Bändern	—	≥458 Demos pro Setup	**Nicht vor 2028; klassisches Tracking plus gelernte Wahrnehmung nutzen**
Neue Produkte ohne neue Daten (Zero-Shot)	Generische VLA	60–80%; ungesehene Steckverbinder 29–67%	—	—	**Nicht realistisch**
Frontier-LLM im Regelkreis, das den Arm steuert	GPT-6-, Claude-, Gemini-Agenten	Greifen 28–45%; Präzisionsinsertion 10%	≈30 min oder 280 s pro Aufgabe vs. ≈20 s	—	**Nicht realistisch**
Frei geführte Kabelbäume und mehrstufige flexible Montage	—	Keine Produktionsbelege 2026	—	—	**Nicht vor 2028**

Quellen: die obigen Studien und (University of Passau on RAPID rewriting, GraspVLA, Komax) **mostly single source and vendor data**. In 2026–2028 liegt der Wert eines großen Modells für Roboter als vortrainierter Prior und Offline-Programmierer, nicht als etwas, das einen Schüler »lehrt«. Jeder produktionsnahe Fall, den ich fand (Neuromeka, AgiBot bei Longcheer, Fraunhofer, ContactDP), beruht auf menschlichen Demonstrationen, Kraft- oder Berührungssensorik, Simulation oder digitalen Zwillingen und überwachtem RL. Keiner beruht auf Labels eines Frontier-Modells.

Kosten und Amortisation

04 Kosten und Amortisation: AI-Nachprüfung amortisiert sich in Monaten, Maschinenbeschickung in Jahren, Ausschleusarme nie

Arbeitskosten: Die USA liegen etwa beim 5.5-Fachen von Shenzhen, Taiwan etwa beim 1.5-Fachen, der größte Teil Südostasiens unter der Hälfte

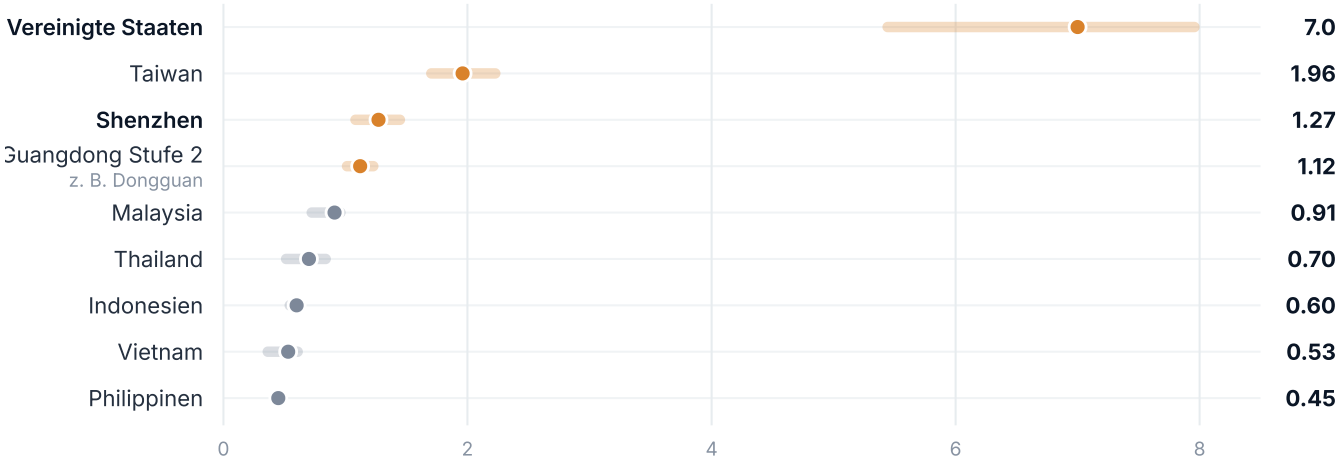
Der Ertrag aus dem Ersetzen von Arbeit beginnt mit den Arbeitskosten. Shenzhens amtliche Lohnleitlinie setzt den Median 2024 für einen Qualitätsprüfer in der Fertigung bei **74,074 RMB** pro Jahr an; Löhne machen 86.92% der Gesamtarbeitskosten in der Elektronikfertigung aus, daher liegen die voll belasteten Kosten bei etwa 85,000 RMB. JETROs Umfrage 2025 setzt die Jahreskosten eines chinesischen Fabrikarbeiters bei $12,667 an, etwa 85,100 RMB, innerhalb von 0.2% (Shenzhen wage guidance, JETRO 2025) verified .

Region	Voll belastete Jahreskosten eines Prüfers oder Bedieners (niedrig / Basis / hoch)	Grundlage
Festlandchina, Shenzhen	70k / **85k** / 100k RMB ($10.4k / **12.7k** / 14.9k)	Amtlicher Median × 1.15; stimmt mit JETRO überein verified
Guangdong Tier-2-Lohngebiete (Dongguan usw.)	65k / **75k** / 85k RMB	Guangdong-Median 2023 × 1.15 plus zwei Jahre Erhöhungen estimate
Jangtse-Delta (Suzhou, Kunshan, Wuxi)	Wie Shenzhen	Keine Grundlage für billigere Prüfarbeit; Jiangsus höchster Mindestlohn liegt etwa 5.6% über Shenzhens estimate
Taiwan	$16.6k / **19.6k** / 22.7k (534k / **630k** / 729k NT$)	Berufslohnumfrage des Arbeitsministeriums plus ≈1.18 gesetzliche Nebenkosten; JETRO als oberes Ende estimate
Malaysia	$6.8k / **9.1k** / 10.0k	JETRO und Mindestlöhne estimate
Thailand	$4.7k / **7.0k** / 8.8k	Gleich
Indonesien	$5.0k / **6.0k** / 6.3k	Gleich
Vietnam	$3.2k / **5.3k** / 6.5k	Gleich
Philippinen	$4.0k / **4.5k** / 4.9k	Gleich
Vereinigte Staaten	$54k / **70k** / 80k	BLS Mai 2025 Prüfer-Median $48,570 ÷ Löhne bei 65.5% der Vergütung estimate

Quellen: (JETRO 2025, Guangdong wage guidance, Wuxi HR bureau, Taiwan occupational wage survey, PTS News, BLS OEWS, BLS ECEC).

FIG 17 **Was eine Inspektions- oder Handhabungskraft pro Jahr kostet *Vollkosten: Die USA liegen etwa beim 5,5-Fachen von Shenzhen, Taiwan beim 1,5-Fachen.***

Jährliche Vollkosten pro QC- oder Handhabungskraft ($10k; Punkt = Basis, Balken = niedrig–hoch)



Quelle: Offizielle Lohnleitlinie Shenzhen, taiwanesisches Arbeitsministerium, JETRO, BLS; Schätzungen des Autors.

Drei Korrekturen legen die Grundlage. Erstens erhöhte Shenzhen seine Rentenbeitragsbasis von 4,775 auf 5,521 RMB pro Monat im Juli 2026 (Shenzhen Bendibao) single source, daher kostet ein Shenzhen-Prüfer 2026 eher 80–90k RMB. Zweitens ist eine Stelle keine Person: Zwei 12-Stunden-Schichten brauchen im Perlflussdelta etwa 2.0 Personen (sechs Tage die Woche) bis 2.4 (ganzjährig), in Südostasien unter Überstundenobergrenzen etwa 2.7–3.1 und in den USA und Taiwan 4.2–4.5 für 24×7 estimate. AI-Nachprüfung lässt meist einen Prüfer zwei bis vier Linien abdecken, daher ist die realistische Einsparung **0.5–1.5 Vollzeitpersonen pro Linie, nicht eine Person pro Linie pro Schicht**. Drittens jagen Käufer in Taiwan, Vietnam und Malaysia nicht nur Kosten: In JETROs Umfrage gaben 61.1% der Firmen in Taiwan, 48.2% in Vietnam und 44.9% in Malaysia an, die Einstellung sei in den letzten zwei Jahren schwieriger geworden, gegenüber 18.6% in Festlandchina (JETRO 2025) single source.

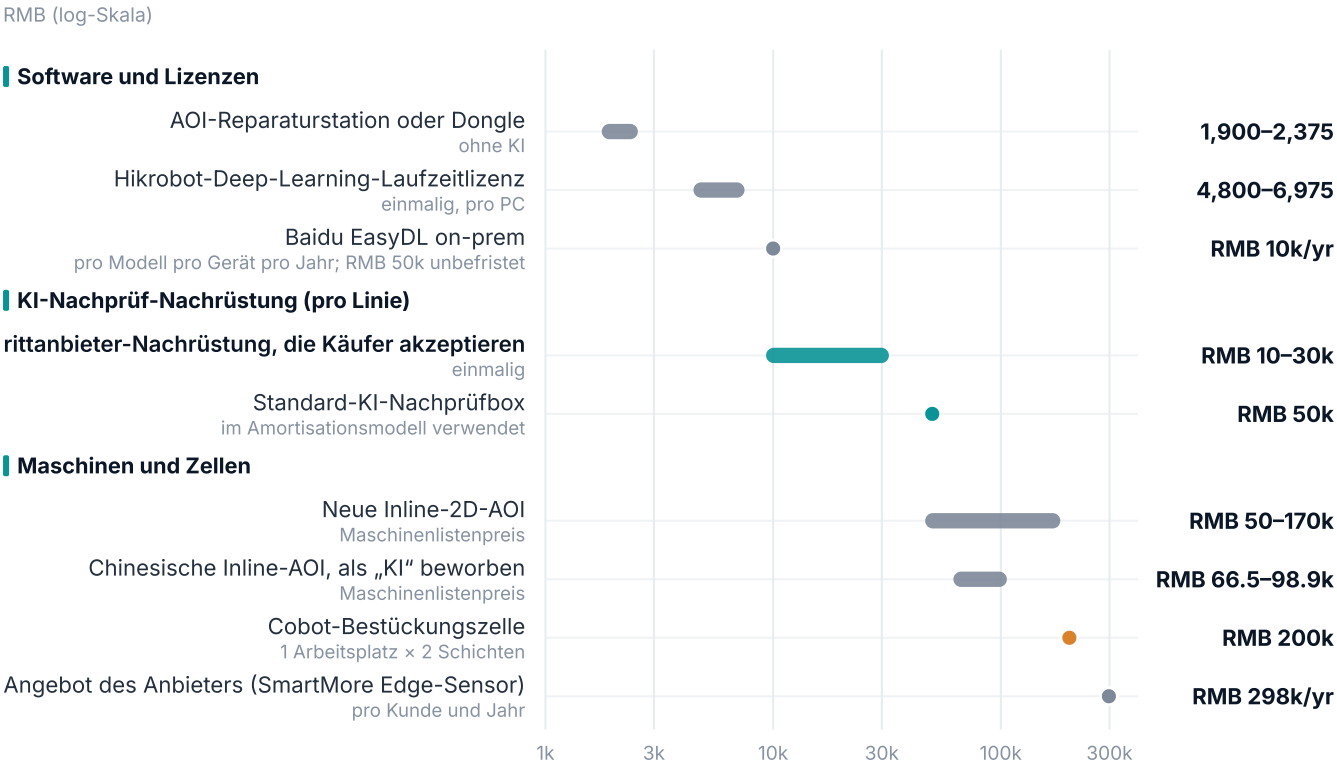
Ausrüstungspreise: vier Grundlagen, die nicht gemittelt werden sollten

AOI-Preise gehen vor allem auseinander, weil sie Unterschiedliches messen. Online-Listen (November 2026, neu, gebraucht und Miete gemischt) liegen bei etwa 17–80k RMB für Offline- oder Desktop-2D, 50–170k RMB für Inline-2D und 80–475k RMB für Inline-3D, bis zu 700k RMB für koreanisches 3D (1688 inline AOI, 1688 offline AOI, 1688 3D AOI) single source, list prices. Eine Ausschreibung vom November 2025 deckelte ein 3D-AOI plus ein 3D-SPI bei 800k RMB (tender notice). Sinic-Teks Prospekt impliziert 428–512k RMB pro 3D-AOI in 2019–2022, nach seinem Bericht 2025 etwa 340–470k RMB (Sinic-Tek prospectus, Sinic-Tek 2025 annual report) primary + estimate. Ein Übersee-Anbieter schätzt $3,000–20,000 für Desktop-2D, $30–70k Inline-2D, $50–90k Einstiegs-3D und $100–200k+ High-End-3D, mit derselben Inline-Maschine bei $30–50k in China (Averroes) vendor estimate from a seller of competing AI software.

Die AI-Nachprüfung selbst hat keinen öffentlichen RMB-Preis; Anbieter nennen Preise fallweise nach einer kostenlosen Vor-Ort-Bewertung (Sohu on Hengtianyi), und Suchen in E-Commerce-Listen, Ausschreibungen, Anbieterartikeln und taiwanischen Medien ergaben kein öffentliches Angebot (1688, 51SMT). Die Anker: etwa €27k pro Linie, impliziert durch den Siemens-Fall estimate; $21,924 Liste für das Fernprüf-Netzwerkmodul eines US-Distributors (Manncorp) vendor list price; etwa 298k RMB pro Kunde und Jahr für SmartMores Edge-AI-Sensoren 2025 (36Kr) single source. Ein als »AI« verkauftes inländisches Inline-2D-AOI listet bei 66.5k RMB (1688 inline AOI), und ein NG-Puffer für manuelle Nachprüfung bei etwa 50k RMB (1688 AOI review). Die Obergrenze für ein Drittparteien-Nachprüf-Add-on liegt also bei etwa 10–30k RMB pro Linie oder einer ähnlichen Jahresgebühr; darüber würde der Käufer die Maschine lieber durch ein inländisches AOI mit eingebauter AI ersetzen estimate.

Anker	Preis (RMB)	Was es bringt	Quelle
AOI-Reparaturstation oder Software-Dongle (älteres TRI, Zhenhuaxing)	≈1,900–2,375	Lizenzschlüssel, keine AI	1688 list price
Allgemeine Deep-Learning-Vision-Runtime (Hikrobot VisionMaster Dongle)	≈4,800–6,975 einmalig pro PC	Die Plattform; der Kunde baut und pflegt Modelle	1688 list price
Baidu EasyDL-Bereitstellung vor Ort	10,000 pro Modell pro Gerät pro Jahr oder 50,000 unbefristet	Verpackter Inferenzdienst	Baidu AI docs 2022 price page
Als »AI« verkauftes inländisches Inline-AOI	66,500–98,900 pro Maschine	Maschinenersatz	1688 list price
SmartMore Edge-AI-Sensoren	≈298,000 pro Kunde und Jahr	Anbieterlösung über mehrere Maschinen	36Kr single source

FIG 18 **Die Preisleiter** ***Von einigen tausend RMB Softwarelizenz bis 200–300k RMB für eine Roboterzelle oder ein KI-Angebot eines Anbieters.***



Quelle: 1688-Listen, Baidu-AI-Dokumentation (Preisliste 2022), 36Kr, ein Brancheninterview (unverifiziert).

Eine Jahresgebühr im Baidu-Stil, mit einem Inferenzserver für ein bis vier Linien und ein bis vier Modellen je Linie, impliziert ein Abonnement von 2.5–40k RMB pro Linie und Jahr, realistisch etwa 5–10k RMB; bei 20% einer unbefristeten Lizenz brechen Abonnement und Kauf im fünften Jahr gleich, eine nachahmenswerte Struktur estimate. Käufer sind nahezu null Softwaregebühren gewohnt: Sinic-Teks gesamter »sonstiger« Umsatz (Ersatzteile, Service, Miete, Zubehör) betrug 7.65% des Umsatzes 2025, etwa 2,300–3,100 RMB pro installierter Maschine und Jahr (Sinic-Tek 2025 annual report) primary + estimate. Ein AI-Abonnement von 5–10k RMB pro Linie und Jahr ist das Zwei- bis Fünffache dessen, was Käufer heute für AOI-Software ausgeben, muss also durch die freigesetzten Nachprüfungsstunden gerechtfertigt werden (etwa 75–90k RMB pro Vollzeitperson und Jahr), nicht durch den Vergleich mit bestehenden Softwaregebühren.

Für Roboterarme fiel der durchschnittliche Cobot-Preis in China 2025 auf etwa 58,600 RMB (GGII) estimate, und Dobots Sechsachs-Durchschnitt fiel in diesem Jahr um 19.1%, als billigere Modelle Anteile gewannen (Huxiu) verified, obwohl die Listenpreise derselben Modelle 2026 um 5–15% stiegen (Sina Finance). In den USA listete ein FANUC CRX-10iA bei $51,434 (Vention) und Zusammenstellungen setzen einen UR7e bei $30–48k an (cobotcost) estimate. Der Arm ist nur 40–50% einer US-Zelle (Standard Bots); in China laufen Kameras, Greifer, Sicherheit und Engineering leicht auf das Zwei- bis Vierfache eines 40k-RMB-Arms, daher ist absolute Engineering-Kosten statt eines Vielfachen zu kalkulieren estimate.

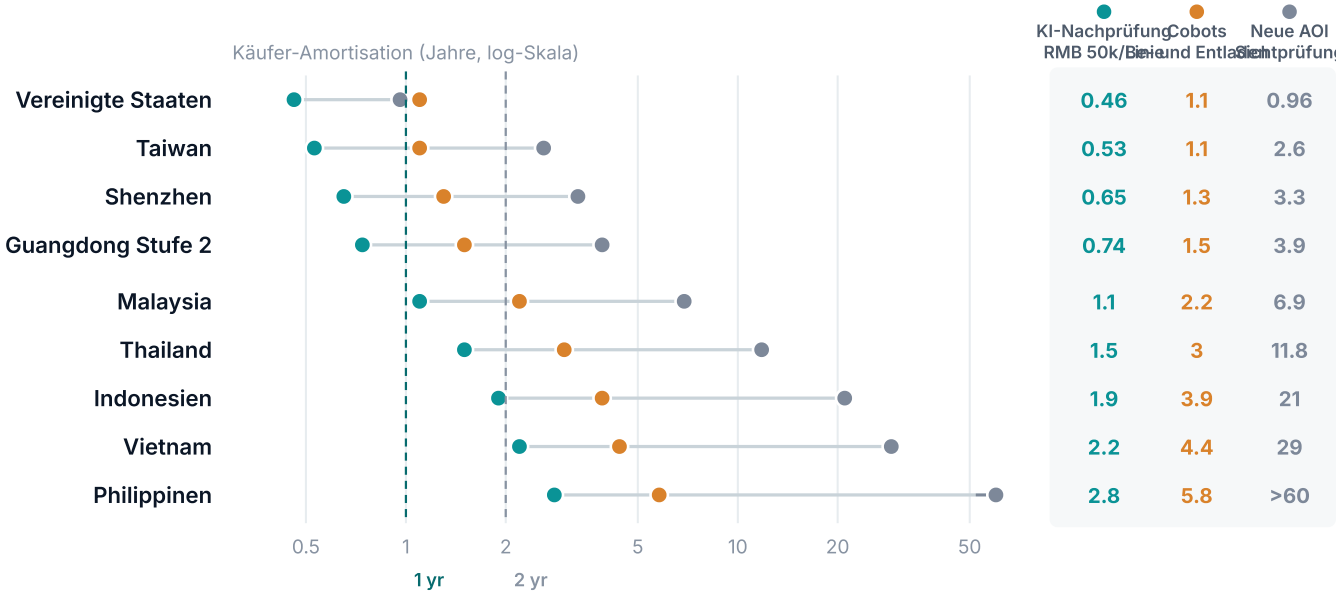
Amortisation: Nachprüfung in Monaten, Beschickung in Jahren, Ausschleusen nie

Einfache Amortisation = Investition ÷ (eingesparte Personen × belastete Arbeitskosten − jährliche Betriebskosten), für eine Zweischicht-Linie ohne Subventionen oder Zölle. Taiwan-Ausrüstung ist mit dem 1.3-Fachen Chinas angesetzt, Vietnam, Malaysia und Thailand mit dem 1.15-Fachen, Indonesien und die Philippinen mit dem 1.2-Fachen und die USA zu Überseepreisen estimate.

Szenario	Investition (China / USA)	Eingesparte Personen	Jährliche Betriebskosten
(a) AI-Nachprüfbox oder AI-Kamerabox	50k RMB / $30k pro Linie	1.0 pro Linie	15%
(b) Neues Inline-2D-AOI ersetzt manuelle Sichtprüfungen (0.5 Prüfer pro Schicht bleiben)	200k RMB / $60k	1.0	12%
(c) Cobot-Ladezelle (eine Station × zwei Schichten)	200k RMB / $140k	2.0	8%
(d) Integrierte Erkennungs-und-Ausschleus-Zelle (Inspektion 1.0 + Ausschleusen 0.07)	300k RMB / $200k	1.07	8%
(e) Ein Roboterarm nur zum Ausschleusen von NG-Platinen	150k RMB / $100k	0.07	8%

Region	(a) AI-Nachprüfung	(b) Neues AOI	(c) Cobot-Zelle	(d) Erkennen-und-Ausschleusen	(e) Ausschleusarm
Shenzhen	0.65 J. (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	Nie
Guangdong Tier-2	0.74	3.9	1.5	5.3	Nie
Taiwan	0.53	2.6	1.1	3.6	Nie
Malaysia	1.1	6.9	2.2	9.2	Nie
Thailand	1.5	11.8	3.0	15	Nie
Indonesien	1.9	21	3.9	25	Nie
Vietnam	2.2 (1.6–4.5)	29	4.4	34	Nie
Philippinen	2.8	>60 (Betriebskosten ≈ Einsparung)	5.8	≈100	Nie
Vereinigte Staaten	0.46	**0.96**	**1.1**	3.4	Nie

FIG 19 **Amortisation nach Region** ***Die KI-Nachprüfung amortisiert sich in den USA, Taiwan und Shenzhen in Monaten, in Südostasien in 1–3 Jahren; die Cobot-Bestückung in Jahren.***



Ein Roboterarm nur zum Aussortieren von NG-Platinen amortisiert sich nirgendwo
Bei Erstausbeute über 90% kostet das Aussortieren nur 0.003–0.14 FTE pro Schicht

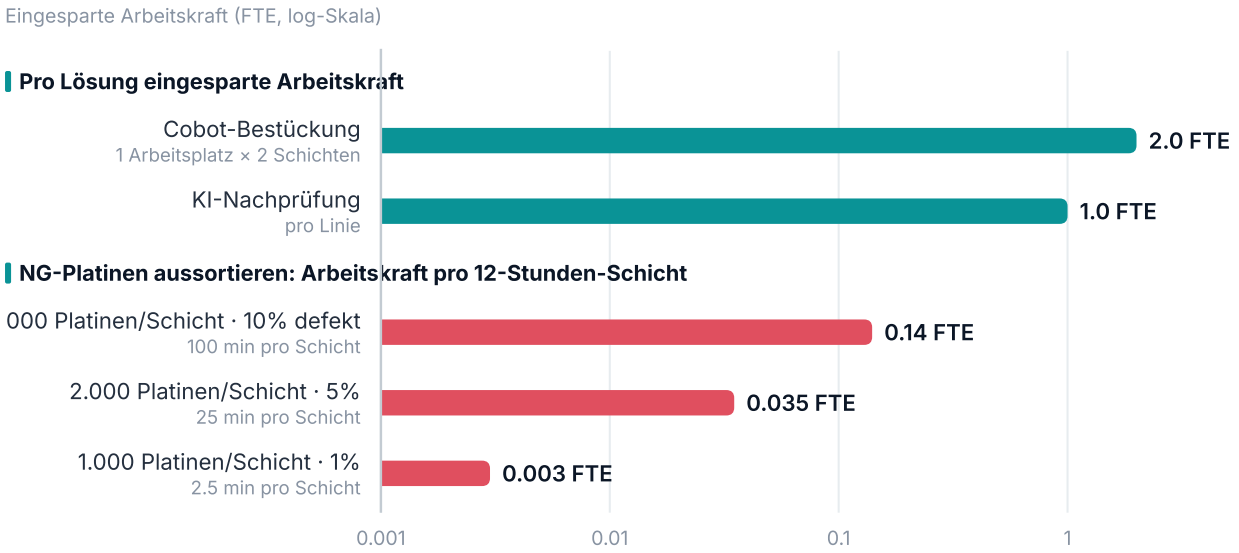
Quelle: Modell des Autors: KI-Nachprüfbox 50k RMB pro Linie; Cobot-Bestückungszelle für 200k RMB oder 140k $, zwei Schichten.

Alle Werte sind Schätzungen aus den obigen Eingaben. Sensitivitäten: Halbieren Sie die eingesparten Personen und jede Amortisation verdoppelt sich grob; fahren Sie eine Schicht und (b)–(d) verdoppeln sich; 100% Abschreibung im ersten Jahr in den USA verkürzt die Amortisation um etwa 20%, während Zölle auf chinesische, japanische oder koreanische Ausrüstung sie um etwa 10–40% verlängern; eine 15%-Lokalsubvention in China verkürzt sie um etwa 15%. Ich fand keine unabhängige Umfrage zu den Amortisationsschwellen der Käufer; Anbieter sagen, AOI amortisiere sich »typischerweise in 9–18 Monaten« (Averroes) und der Siemens-Fall nennt acht Monate vendor claims. Ich verwende ein und zwei Jahre als Screening-Linien. Das Modell lässt Qualitätsgewinne (Garantie, Retouren) sowie Fluktuations- und Schulungskosten weg, ist also konservativ.

Das Kontraintuitive an »Erkennen und Ausschleusen« ist, dass die Arbeit in der Nachprüfung liegt, nicht darin, die wenigen echten NG-Platinen von der Linie zu nehmen. Bei 15 Sekunden Handhabung pro NG-Platine estimate:

Platinen pro Schicht	Defektrate	Ausschleuszeit pro 12-Stunden-Schicht	Personen
1,000	1%	2.5 min	0.003
2,000	5%	25 min	0.035
4,000	10%	100 min	0.14

FIG 20 **Warum sich das Aussortieren schlechter Platinen nicht lohnt** ***Ab 90% Erstausbeute kostet das Aussortieren nur 0,003–0,14 Vollzeitkräfte pro Schicht.***



Quelle: Schätzungen des Autors.

Der Basisfall (2,000 Platinen pro Schicht, 5% defekt, zwei Schichten) sind 0.07 einer Person, etwa 6,000 RMB pro Jahr wert in Shenzhen, weniger als die grob 12,000 RMB jährlichen Betriebskosten des Arms, daher amortisiert er sich nirgendwo; selbst der schlechteste Fall (4,000 Platinen pro Schicht, 10% defekt) braucht in Shenzhen etwa 13 Jahre und in den USA 9 estimate. Wenn die Sorge das Anhalten der Linie zum Ausschleusen ist, löst eine Standard-NG-Ausschleusung oder ein Pufferförderer das ohne Arm. Branchenleute sagen »über 90% Erstausbeute, Linie anhalten und die Platine von Hand entnehmen«, und **das stimmt fürs Ausschleusen, aber nicht für die Inspektion**. Das Produktdesign auf einer SMT-Linie ist daher »prüfen plus markieren oder ausschleusen«, mit Roboterarmen für Orte, wo jedes Teil gehandhabt wird (Testvorrichtungs-Beladung, Tray- und Magazinhandhabung, Verpackung), wo Inspektion ein fast kostenloses Add-on wird und Szenario (c) gilt.

Subventionen, Steuern und Zölle verschieben die Amortisation um 15–25%, aber nicht die Rangfolge

Festlandchina: Förderung senkt meist die Finanzierungskosten; einzelne Maschinen qualifizieren sich selten direkt. Eine nationale Kreditsubvention zahlt 1.5 Zinspunkte für bis zu zwei Jahre in 14 Prioritätsketten (Yicai) verified, etwa 9,000 RMB auf eine 300k-RMB-Maschine; Guangdongs Upgrade-Prämie von bis zu 20% erfordert ein registriertes Projekt mit mindestens 7 Millionen RMB neuer Ausrüstung und schließt Shenzhen aus (Guangdong rules), und Dongguan zahlte im Juli 2026 tatsächlich 4.42% (Dongguan allocation) single source, primary. Ein AI-Inspektionsgerät ist nur als Ausrüstung »mit begleitender Software« innerhalb des registrierten Projekts des Käufers genehmigt, daher geht der Basisfall von keiner Subvention aus.

Taiwan: Anreize gehen an den Käufer. Nach Artikel 10-1 des Statute for Industrial Innovation, im Mai 2025 um AI erweitert, kann ein Käufer 5% der Ausgaben 2025–2029 für neue intelligente Maschinen oder AI (einschließlich CNN/DNN-basierter Inspektion) gegen die Einkommensteuer dieses Jahres anrechnen, oder 3% pro Jahr über drei Jahre (Statute, credit rules) verified; anbieterseitige Zuschüsse sind für Festlandkapital geschlossen (MOEA) single source.

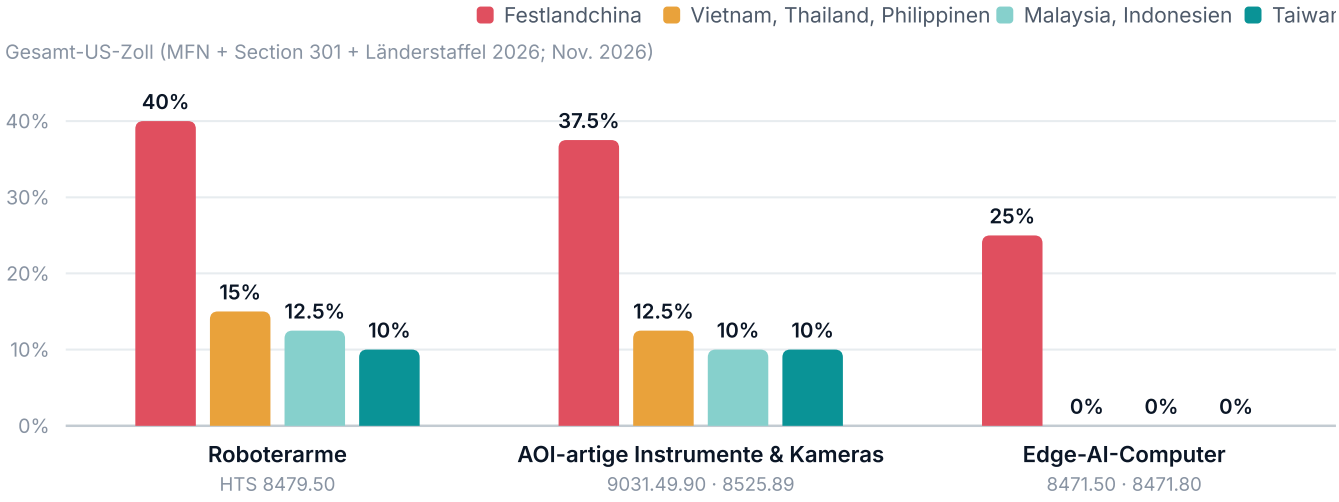
Südostasien: Nur Malaysia und Thailand belohnen den Kauf von Automatisierung. Malaysias 200%-Automatisierungs-Kapitalfreibetrag auf die ersten 10 Millionen RM ist für einen profitablen Käufer etwa 24% der Investitionsausgaben wert (MIDA, PwC Malaysia) verified; Thailands BOI gewährt eine dreijährige Steuerbefreiung bis zu 50–100% einer Upgrade-Investition, mit höherer Obergrenze bei Beteiligung thailändischer Automatisierungsanbieter (BOI) single source, primary, daher funktioniert der Markteintritt in Thailand am besten mit einem lokalen Integrator.

Vietnam, Indonesien und die Philippinen belohnen den Bau von Werken statt den Kauf von Automatisierung (Vietnam Law Magazine, PwC Philippines).

Vereinigte Staaten: Abschreibung verkürzt die Amortisation; Zölle verlängern sie. Nach dem 19. Januar 2025 gekaufte Anlagen erhalten dauerhaft 100% Abschreibung im ersten Jahr, ein etwa 21%-iger Bundessteuerschutz für einen profitablen Käufer (IRS Notice 2026-11) single source, primary ; die Section-232-Untersuchung zu Robotern und Industriemaschinen endete am 28. Oktober 2026 mit Verhandlungen statt neuer Zölle und soll innerhalb von 180 Tagen neu bewertet werden (White House, Mayer Brown) verified . Endmontage in Südostasien senkt den Zuschlag auf Arme und Instrumente vom AOI-Typ von etwa 37.5–40% auf 10–15% (Kapitel 1).

Region	Hauptinstrumente	Effekt auf die Käufer-Amortisation estimate
Festlandchina	1.5-Punkte-Kreditsubvention für bis zu zwei Jahre; Upgrade-Prämien von Provinz und Stadt (Guangdong verlangt registrierte Projekte ≥7 Mio. RMB; Dongguan zahlte tatsächlich 4.42%); Shenzhen-Maßnahmen 2026–2029 (Sätze nicht öffentlich)	Etwa 0–5% für eine einzelne Maschine; 4–15% innerhalb eines großen Projekts
Taiwan	Statute §10-1-Gutschrift, 5% oder 3% × 3 Jahre	Etwa 5% kürzer, wenn der Käufer genug Steuern zahlt
Malaysia	200% Automatisierungs-Kapitalfreibetrag (erste 10 Mio. RM)	Etwa 15–25% kürzer für profitable Käufer
Thailand	BOI dreijährige Steuerbefreiung bis zu 50–100% der Investition	Etwa 15–25% kürzer für profitable Käufer außerhalb einer Steuerbefreiung
Vietnam, Indonesien, Philippinen	Einkommensteuer-Anreize für den Werkbau; zusätzliche 20% Ausrüstungsabschreibung auf den Philippinen	Fast nichts für einzelne Automatisierungskäufe
Vereinigte Staaten	100% Abschreibung im ersten Jahr; ≈37.5–40% kombinierter Zoll auf Festland-Ausrüstung	Abschreibung ≈20% kürzer; Zölle ≈10–40% länger

FIG 21 **US-Zölle auf Roboter, AOI und Edge-Computer** ***In Festlandchina gefertigte Arme zahlen etwa 40%; aus Südostasien oder Taiwan versandt 10–15%.***



Quelle: USITC Harmonized Tariff Schedule Kapitel 99; Federal Register; USTR (November 2026).

Je Region: In Festlandchina amortisiert sich eine AI-Nachprüfbox in Monaten und eine Cobot-Zelle in 1.3–1.5 Jahren, während ein neues AOI, das nur zur Arbeitseinsparung gekauft wird, 3–4 Jahre braucht und integriertes Erken-

nen-und-Ausschleusen 4.5–5.3, beide jenseits eines Zwei-Jahres-Screens. Taiwan schneidet etwas besser ab als Shenzhen (0.53 Jahre für die Nachprüfung, 1.1 für eine Cobot-Zelle), und die Einstellung ist dort schwieriger. In Malaysia und Thailand braucht eine AI-Box 1–1.5 Jahre und eine Cobot-Zelle 2–3, etwa 15–25% schneller mit lokalen Steueranreizen für profitable Käufer. In Indonesien, Vietnam und den Philippinen amortisiert sich nur eine sehr billige AI-Box innerhalb von 2–3 Jahren; die Nachfrage kommt dort eher aus Qualitätsanforderungen der Kunden und Arbeitskräftemangel. In den USA amortisiert sich jede Option außer einem reinen Ausschleusarm in 0.2–3.4 Jahren, mit Zöllen, der Section-232-Neubewertung und der NRTL-Zertifizierung als Hauptvariablen.

05 Markt und Wettbewerb

Markt, Wettbewerb und wie Käufer zahlen

Marktgröße: Veröffentlichte Zahlen weichen um ein Vielfaches voneinander ab; nur eine Bottom-up-Zählung ist brauchbar

Jede Marktzahl muss mit ihrem Geltungsbereich gelesen werden. Chinas Machine-Vision-Markt lag 2024 bei 18.15 Mrd. RMB (Zhiyan) verified, wobei 3C-Elektronik der größte Abnehmer ist; globale Schätzungen für Machine Vision weichen um etwa das Dreifache und für AOI um etwa das 6.5-Fache voneinander ab (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Bottom-up aus den Umsätzen von Koh Young, TRI, Jutze, Sinic-Tek und ViTrox liegt globale SMT-Inspektionsausrüstung in der Größenordnung von $1–2 Mrd. estimate .

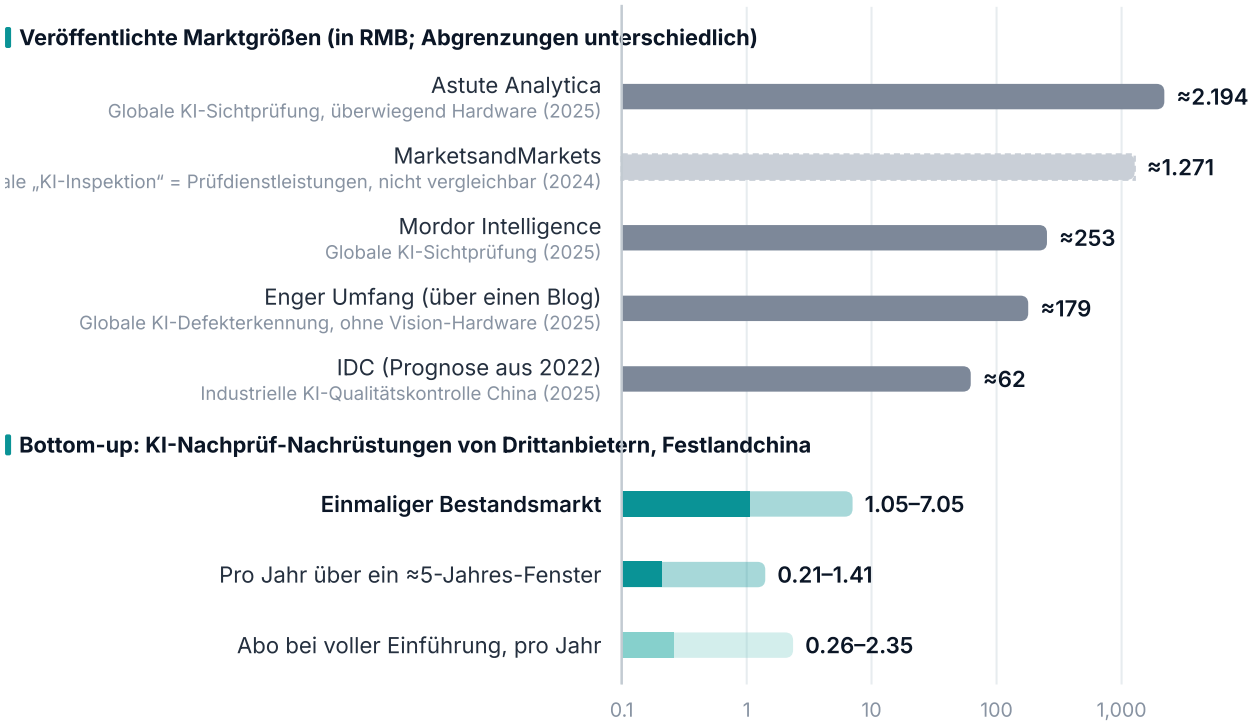
Es gibt keinen eigenständigen Markt für »AI-Sichtprüfungssoftware«. Die veröffentlichten Zahlen weichen um mehr als das Zehnfache voneinander ab und enthalten alle Hardware oder Dienstleistungen:

Geltungsbereich	Herausgeber	Zahl
Global »AI-Industriedefekterkennung« (eng, ohne Vision-Hardware)	Nicht genanntes Unternehmen, über einen iFactory-Blog	$2.66 Mrd. im Jahr 2025, 8.6% CAGR bis 2035, Elektronik 34%
Global »AI-Sichtprüfung«	Mordor Intelligence	$3.76 Mrd. im Jahr 2025, $12.8 Mrd. im Jahr 2031, 22.67% CAGR
Global »AI-Sichtprüfung« (überwiegend Hardware)	Astute Analytica	≈$32.66 Mrd. im Jahr 2025
Global »AI-Inspektion« (Prüf- und Zertifizierungsdienstleistungen; nicht vergleichbar)	MarketsandMarkets	$18.92 Mrd. im Jahr 2024
China »industrielle AI-Qualitätsinspektion«	IDC, Prognose aus dem Jahr 2022	$958 Mio. im Jahr 2025 (≈6.2 Mrd. RMB), 91.5% in Kommunikation, Elektronik und zwei weiteren Branchen

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

FIG 22 **Marktgrößen unterscheiden sich um mehr als das Zehnfache** ***Nur der enge Umfang und eine Bottom-up-Zählung sind für Entscheidungen brauchbar.***

RMB 100M (log-Skala; USD zu 6.7188)



Quelle: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; Bottom-up-Schätzung des Autors.

Bottom-up gibt es nur eine öffentliche Zählung von SMT-Linien: New Venture Research schätzte 2020 weltweit etwa 73,231 Linien, davon 47% in China (19% im Jahr 2000), etwa 70% High-End und 30% Low-End (EMSNow, NVR SMT2020 summary) single source, 2020; NVR gehört heute zu IPC, und neuere Daten sind kostenpflichtig (NVR). Bei einem Wachstum von 3–5% pro Jahr und einem China-Anteil von 40–47% ergeben sich 2026 weltweit etwa 87,000–98,000 Linien und 35,000–46,000 in Festlandchina (Basis 40,000) estimate.

Eingabe	Niedrig	Basis	Hoch	Grundlage
SMT-Linien in Festlandchina (2026)	35,000	40,000	46,000	NVR-2020-Basis × Wachstum × Anteil
Anteil ohne AI-Nachprüfung	75%	80%	85%	Nachrüstfenster, unten
Anteil, den ein Dritter erreichen kann (nicht intern, nicht an OEM-AI gebunden, zahlungsfähig)	40%	50%	60%	Annahme: ≈30% Low-End-Linien plus einige High-End-Linien mit gemischten Marken, ohne interne Eigenbauten nach DBG-Art
Adressierbare Linien	**10,500**	**16,000**	**23,500**	Produkt
Einmalpreis pro Linie	RMB 10k	RMB 20k	RMB 30k	Preisstaffel und -obergrenze (Kapitel 4)
Einmaliger adressierbarer Markt (installierte Basis)	**RMB 105M**	**RMB 320M**	**RMB 705M**	Produkt
Pro Jahr über ein ≈5-Jahres-Fenster	RMB 21M	RMB 64M	RMB 141M	Fenster schrumpft um 8–12% pro Jahr
Abonnement (RMB 2.5k / 5k / 10k pro Linie und Jahr, volle Übernahme)	RMB 26M pro Jahr	RMB 80M pro Jahr	RMB 235M pro Jahr	Anker in Baidu-artiger Jahresgebühr

FIG 23 **Bottom-up: 40.000 Linien → 16.000 bedienbar → ≈320 Mio. RMB einmalig** ***Die Nische der KI-Nachprüfung durch Drittanbieter in Festlandchina ist real, aber klein.***



Ein Neueinsteiger gewinnt in 3 Jahren 2–5% der bedienbaren Linien: ≈320–800 Linien, ≈RMB 2–5 Mio. pro Jahr
Als Abo (RMB 5k pro Linie und Jahr): RMB 1.6–4 Mio. wiederkehrend, nur 0.3–2.3% der IDC-Prognose für Chinas industrielle KI-QC

Quelle: NVR (Basis 2020); Schätzungen des Autors.

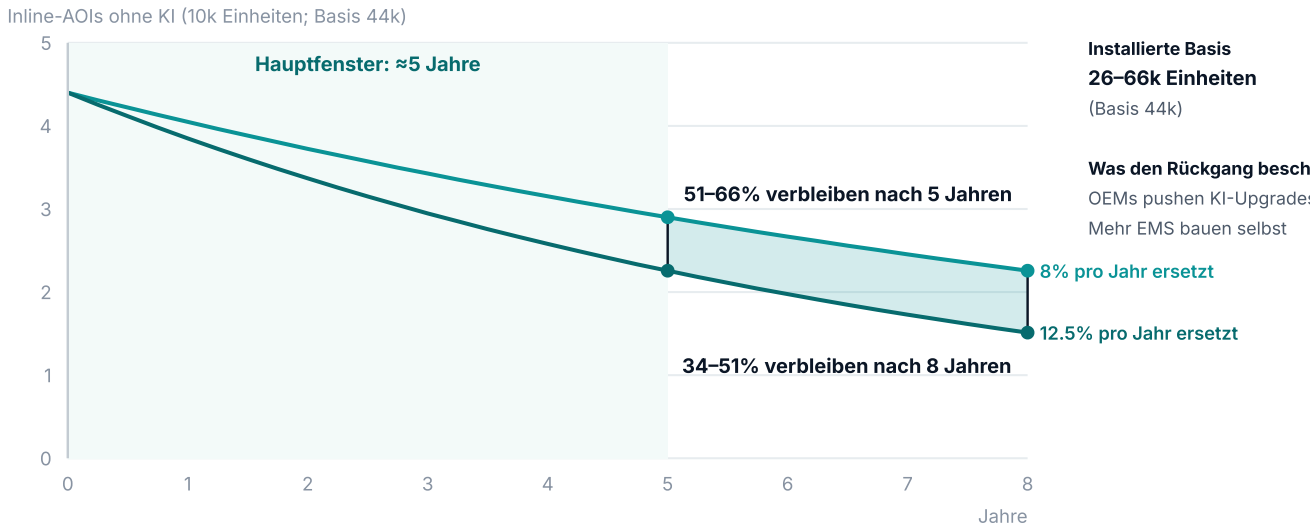
All das sind Schätzungen; weder der Anteil noch der Preis wurden öffentlich gemessen. Die Nische ist nur 0.3–2.3% pro Jahr von IDCs Prognose für die gesamte industrielle AI-Qualitätsinspektion Chinas im Jahr 2025 wert (etwa 6.2 Mrd. RMB). Ein Neueinsteiger, der in drei Jahren 2–5% der adressierbaren Basislinien gewinnt, hätte etwa 320–800 Linien: insgesamt 6.4–16 Mio. RMB bei 20,000 RMB pro Linie, etwa 2–5 Mio. RMB pro Jahr oder 1.6–4 Mio. RMB pro Jahr an wiederkehrenden Einnahmen bei 5,000 RMB pro Linie estimate. Das reicht für eine Geschäftseinheit innerhalb eines Unternehmens, nicht für ein eigenständiges Unternehmen in Venture-Größe. Jeweils 1,000 installierte Linien in Taiwan, Südostasien oder den USA sind grob 3–15 Mio. RMB einmalig zu Festlandpreisen wert; die Preise sind in Taiwan und den USA höher, aber Käufer dort nutzen häufiger OEM-3D-AOIs mit AI-Optionen estimate.

Das Nachrüstfenster beträgt etwa fünf Jahre

Kein AOI-Hersteller veröffentlicht die Maschinenlebensdauer; EMS-Unternehmen schreiben Ausrüstung über 3–10 Jahre ab (DBG 2025 annual report), und ältere AOIs verlieren nach dem End of Life Software-Updates und Ersatzteile (I-Connect007) single source. Ich verwende 8–12 Jahre (Basis 10).

Bei 1.0–1.65 Inline-AOIs pro Linie hat das Festland etwa 35,000–78,000 Inline-AOIs im Einsatz (Basis etwa 55,000), im Einklang mit bekannten Auslieferungen (Circuits Assembly, Sinic-Tek 2025 annual report). Keine Quelle misst, wie viele mit AI-Nachprüfung laufen; unter der Annahme, dass inländische Maschinen sie seit etwa 2023 anbieten und 40–60% der jüngeren Lieferungen sie nutzen, **laufen in Festlandchina etwa 26,000–66,000 Inline-AOIs (Basis etwa 44,000) nicht mit AI** estimate. Wenn jede neue Maschine ab sofort mit AI ausgeliefert wird, schrumpft dieser Bestand um 8–12.5% pro Jahr: Nach fünf Jahren bleiben 51–66% und nach acht Jahren 34–51%, schneller, wenn AOI-Hersteller AI-Upgrades auf neuere Maschinen drängen. **Der Großteil der Gelegenheit fällt in die nächsten etwa fünf Jahre**, mit mittelgroßen EMS, die ältere inländische 2D-AOIs, gemischte Marken und keinen AI-Upgrade-Pfad im Kern betreiben.

FIG 24 **Das Nachrüstfenster beträgt etwa fünf Jahre*Inline-AOIs ohne KI schrumpfen um 8–12,5% pro Jahr, da sie ersetzt oder aufgerüstet werden.***

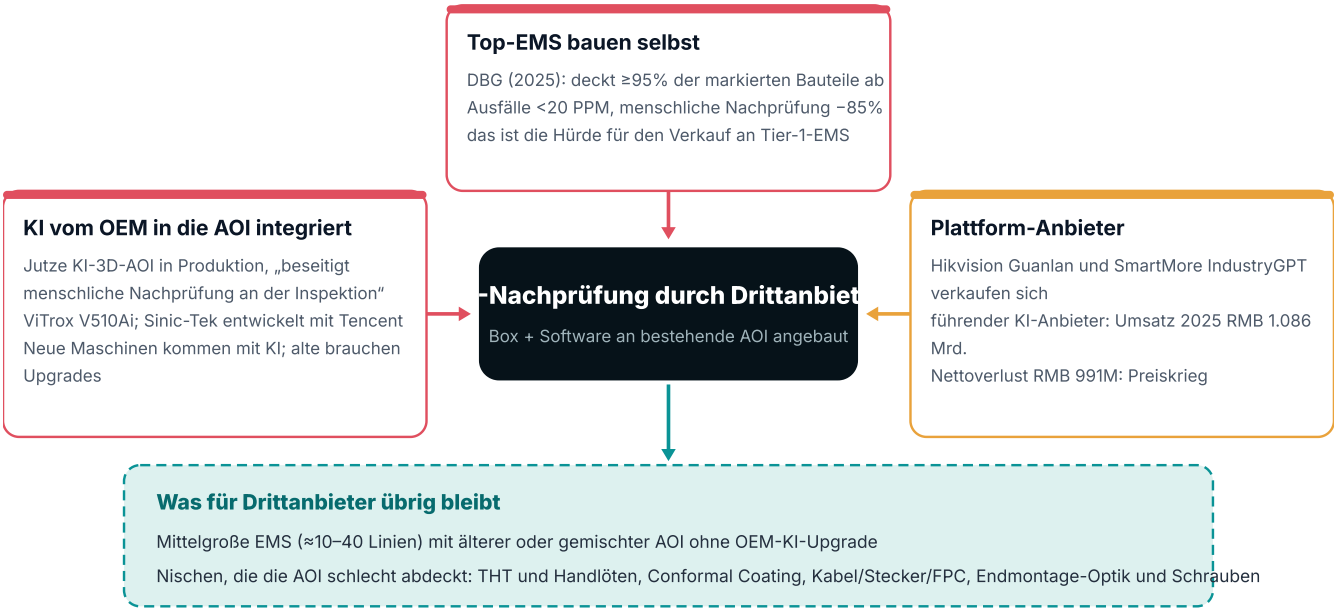


Quelle: Schätzungen des Autors; installierte Basis und Anteile nicht unabhängig verifiziert.

Die Wettbewerber sind bereits stark: Der Review-Keil ist umzingelt

Der AI-Review-Keil verengt sich aus drei Richtungen. AOI- und Vision-Marktführer bauen AI in neue Plattformen ein: Jutze hat »AI-Algorithmen vollständig übernommen« (Jutze investor relations, June 2026), OPTs industrieller AI-Umsatz wuchs 2025 um 96.77% (OPT 2025 annual report) single source, primary, und ViTrox beansprucht bis zu 95% Genauigkeit für die autonome Nachprüfung (ViTrox) vendor claim. Große Hersteller bauen eigene Lösungen (iThome). Und venturefinanzierte AI-Unternehmen verkaufen in großem Maßstab: SmartMores Edge-Sensor-Kunden wuchsen in zwei Jahren von 65 auf 251 (36Kr) single source, Hikvision verkauft sein Guanlan-Modell für die Inspektion (Xinhua), und die Integration von »Auge, Gehirn und Hand« wird bereits von SmartMore, Hikrobot und Mech-Mind angeboten (36Kr, Securities Star).

FIG 25 **Der Nachprüf-Keil ist von drei Seiten umzingelt*Raum für Drittanbieter liegt bei mittelgroßen EMS und in Inspektionsnischen, die die AOI schlecht abdeckt.***



Quelle: Geschäftsberichte von Jutze, Sinic-Tek und DBG; öffentliches Material von Hikvision und SmartMore; vom Autor zusammengestellt.

Der für einen Neueinsteiger verbleibende Raum ist eng: mittelgroße EMS mit älteren 2D/3D-AOIs und ohne AI-Upgrade vom Hersteller; Linien, die mehrere AOI-Marken mischen (ein Nachrüstanbieter unterstützt bereits Omron, VI, Koh Young und Mirtec (Sohu on Hengtianyi)); Nischen, die SMT-AOIs schlecht abdecken, wie Durchsteck- und Handlötstellen, Conformal Coating, Kabel, Steckverbinder und FPCs sowie Enderscheinung und Schrauben; und kleine Werke, die sich kein AOI für 230–280k RMB leisten können.

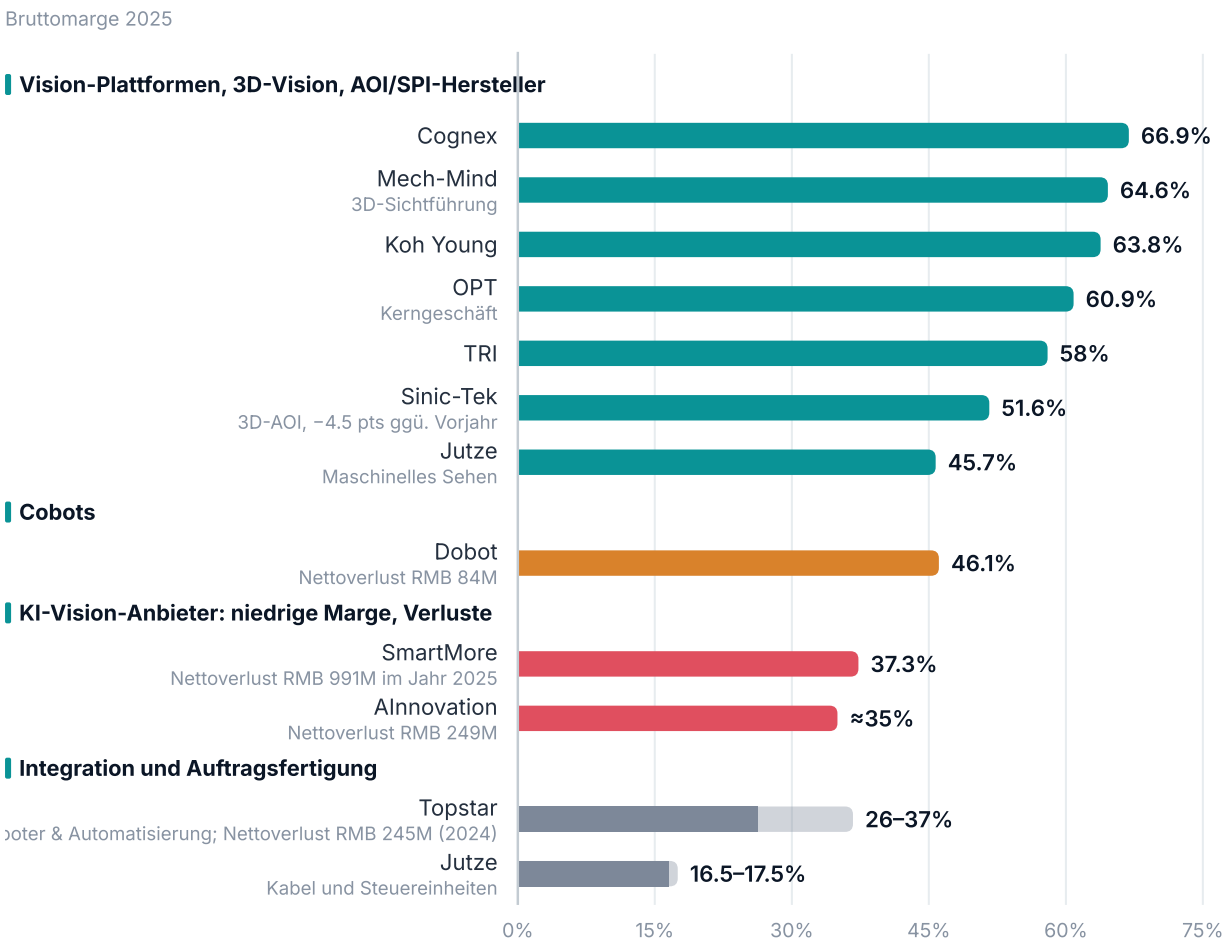
Der nächstgelegene börsennotierte Vergleich ist eine Warnung. Weiyi Zhizao, das AI-Inspektionsausrüstung und Roboter verkauft, verzeichnete im Januar–Mai 2026 einen Umsatzrückgang im 3C-Bereich von 79.9% gegenüber dem Vorjahr, räumte »zunehmenden Wettbewerb und Preisdruck« ein, verschob seine Hongkong-Notierung am 25. Oktober 2026 und legte 15 vor Abschluss beendete Projekte offen, davon 12 Projekte mit verkörperten Robotern (Weiyi Zhizao prospectus, postponement notice) single source, primary . AInnovation beendete 2025 mit 29.8% seiner Forderungen wertberichtigt (AInnovation 2025 annual report). Der 3C-Inspektionskeil ist überfüllt, und die Preissetzungsmacht pro Linie ist schwach.

Die Marge liegt bei produktisierter Hardware

Position in der Kette	Repräsentative Unternehmen und Bruttomarge 2025	Gewinn und Zahlungseingänge
Vision-Plattformen	Cognex 66.9%	Profitabel; Verkauf von ≈$22M margenschwachem Umsatz
AOI/SPI-Ausrüstung	Koh Young 63.8%; TRI 58%; OPT-Kerngeschäft 60.85%; Sinic-Tek 3D-AOI 51.62% (−4.53 Punkte ggü. Vorjahr); Jutze Machine Vision 45.74%	Profitabel; Forderungstage 139 (Jutze) bis 247–274 (OPT)
3D-Vision-Führung	Mech-Mind 64.6%	—
Anbieter von AI-Vision-Lösungen	SmartMore 37.3%; AInnovation ≈35%	Verlustreich: SmartMore-Nettoverlust 991M RMB im Jahr 2025, AInnovation 249M RMB
Cobots	Dobot 46.1%	Dobot-Nettoverlust 84M RMB; UR-Umsatz von $326M (2022) auf ≈$232M (2025, Prognose)
Systemintegration	Topstar Robotik und Automatisierung 26.2–36.7%	Nettoverlust 245M RMB (2024); Forderungstage 180–240
Auftrags-Hardware	Jutze Kabel und Steuereinheiten 16.5–17.5%	—

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

FIG 26 **Die Marge steckt in produktisierter Hardware** ***Vision-Plattformen und Gerätehersteller erzielen 46–67% Bruttomarge; KI-Lösungsanbieter 35–37% und machen Verluste.***



Quelle: Geschäftsberichte oder Prospekte 2025.

Ein Neueinsteiger behält seine Marge nur, wenn er die Inspektion in ein Produkt, eine Ausrüstung oder Software verwandelt; projektweises Nachjustieren drückt die Marge auf etwa 35% bei anhaltenden Verlusten. Auch die Mechanismen arbeiten gegen kleine Unternehmen: Umsatz wird erst nach der Abnahme vor Ort erfasst, Vorauszahlungen sind klein (Jutze, OPT), Piloten sind meist kostenlos (A3/Mindtrace), und Markeninhaber geben oft die Inspektionsausrüstung auf den Linien ihrer Auftragsfertiger vor.

Wie Käufer zahlen: 30% bei Unterzeichnung, 60% nach Abnahme, 10% ein Jahr einbehalten

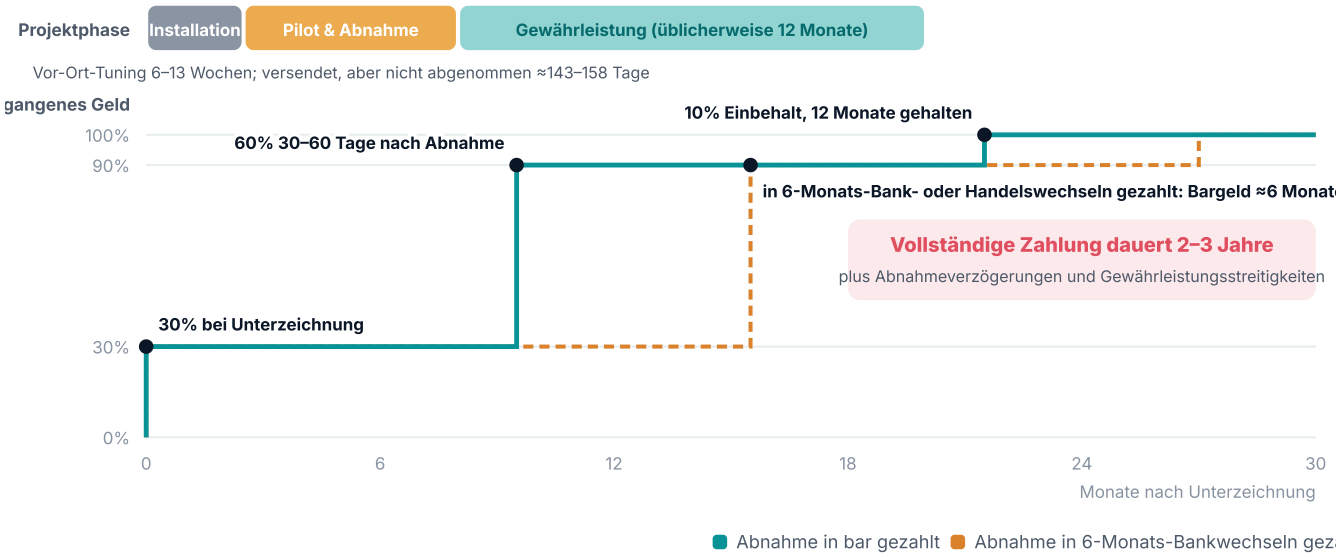
Die Zahlungsbedingungen, die Festland-Elektronikwerke für Inspektionsausrüstung verwenden, sind in 13 namentlich genannten Käuferverträgen festgehalten, die in Sinic-Teks Antwort auf die Notierungsanfrage offengelegt sind (Sinic-Tek inquiry reply 8-1-1) single source: one primary filing, several independent buyers, 2020–2022 :

Käufer (Jahr)	Unterzeichnung	Lieferung	Abnahme	Einbehalt oder Restbetrag
Eine Hikvision-Einheit (2021)	30% innerhalb von 15 Tagen	—	65% innerhalb von 30 Tagen nach Endabnahme	5% Garantieeinbehalt, 12 Monate nach Abnahme gezahlt; Teil in 6-monatigen oder längeren Bankakzepten gezahlt
Nanchang Huaqin (2020–2021)	30% innerhalb von 10 Arbeitstagen	—	60% 60 Tage nach Abnahme	10% nach einjähriger Garantie

Käufer (Jahr)	Unterzeichnung	Lieferung	Abnahme	Einbehalt oder Restbetrag
Guangdong Yinuo (2021, 3D-AOI)	30% innerhalb von 10 Arbeitstagen	—	60% innerhalb von 30 Tagen	10% nach drei Monaten störungsfreiem Lauf
Xiamen Hongxin, Jingmen Hongyi (2022)	30% innerhalb von 10 Arbeitstagen	—	60% innerhalb von 30 Arbeitstagen nach Endabnahme, 6-Monats-Bankwechsel	10% nach einjähriger Garantie
AKM Panyu, Suzhou (2021–2022)	30% innerhalb von 7–10 Arbeitstagen	—	60% einen Monat nach Abnahme und Rechnungsstellung, 180-Tage-Bankwechsel	10% nach 360 Tagen
BYD Precision, Shenzhen, Huizhou (2020)	—	30% bei Ankunft	70% nach Abnahme und Rechnungsstellung	Alles in 6-Monats-Handelswechseln
Shenzhen H&T (2020)	—	—	90% nach Abnahme und Rechnungsstellung	10% nach sechs Monaten störungsfreiem Einsatz

Andere Primärunterlagen zeigen dieselbe Struktur: Mech-Minds Großprojekte zahlen 30% bei Unterzeichnung, 30% vor Lieferung, 30% nach Abnahme und 10% Garantieeinbehalt über 12–15 Monate (Mech-Mind prospectus); GKG gibt an, dass der Einbehalt in der Ausrüstungsbranche üblicherweise 5–10% des Vertrags beträgt und ein Jahr nach Abnahme gezahlt wird (GKG inquiry reply) single sources, primary. Die Norm ist also etwa 30% bei Unterzeichnung (20–30%, bei manchen Werken wie BYD null), 60–65% nach Abnahme (30–60 Tage später gezahlt, oft in sechsmonatigen Bank- oder Handelswechseln, was das nutzbare Bargeld um ein weiteres halbes Jahr verzögert) und 5–10% Einbehalt über 3–12 Monate (12 am häufigsten) verified across several contracts.

FIG 27 **Wie Käufer zahlen *30% bei Unterzeichnung, 60% nach Abnahme, 10% ein Jahr einbehalten: Die vollständige Zahlung dauert 2–3 Jahre.***



Quelle: Antwort auf Anfrage bei Sinic-Tek, Mech-Mind, GKG, JT und weitere Unterlagen; schematisch zu üblichen Konditionen, Zeitangaben geschätzt.

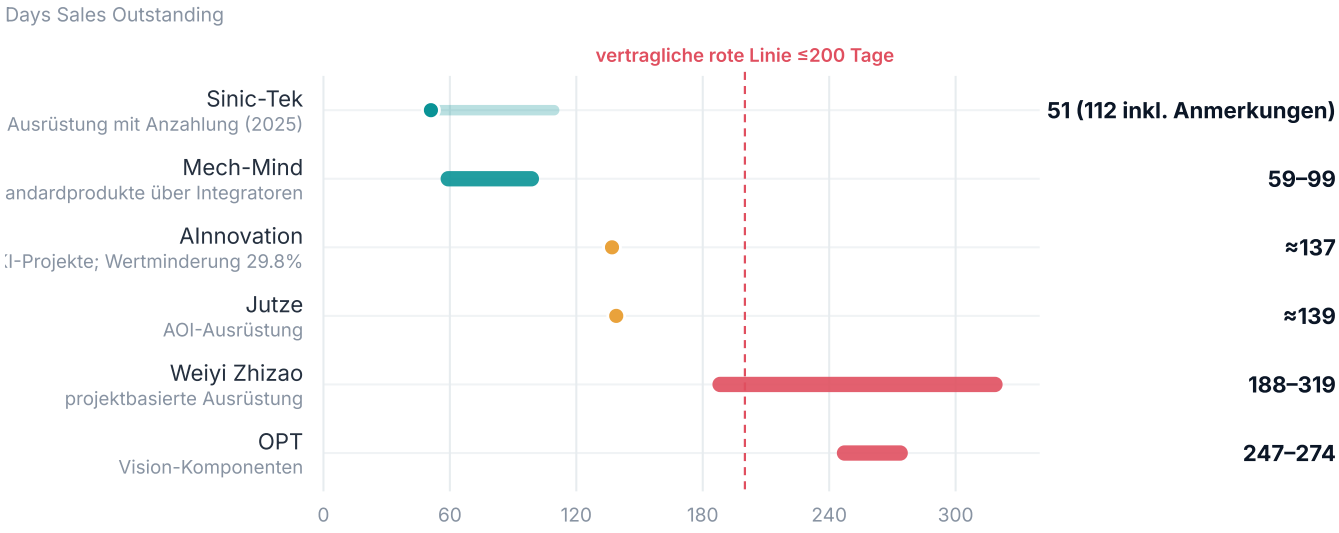
Abnahme und Tests sind langsam und werden vom Verkäufer bezahlt. Sinic-Teks gelieferte, aber nicht abgenommene Waren brauchten etwa 143–158 Tage zum Umschlagen, weil große EMS- und Markenkunden Integrationstests der gesamten Linie und einen verlängerten Probebetrieb benötigen (Sinic-Tek inquiry reply 8-1-2); Weiyis AI-Projekte brauchen 4–8 Monate von der Evaluierung bis zur Bereitstellung (Weiyi Zhizao prospectus) single sources, primary. Es gibt keine Konversionsdaten für AI-Nachprüfungs-Piloten; Sinic-Tek wandelte nur etwa 18–19% seiner

Demo-Maschinen direkt in Verkäufe um. Ich rechne mit einer von drei bis einer von fünf für kostenlose Piloten und etwa 80% für bezahlte Piloten nach der Vorauswahl estimate.

Unternehmen	Geschäftsmodell	Forderungslaufzeit (Tage)
Sinic-Tek (2025)	Ausrüstung mit Anzahlungen	≈51 (≈112 einschließlich Wechsel)
Mech-Mind	Standardprodukte über Integratoren	59–99
Jutze	AOI-Ausrüstung	≈139
AInnovation (2025)	AI-Projekte	≈137 zum Jahresende, mit 29.8% Wertberichtigung
Weiyi Zhizao	Projektbasierte Ausrüstung	188–319
OPT	Vision-Komponenten	247–274

Quellen: die Geschäftsberichte und Prospekte oben; Forderungen (einschließlich Vertragsvermögen) ÷ Umsatz × 365 primary + estimate.

FIG 28 **Days Sales Outstanding** ***Standardprodukte 51–99 Tage; projektbasierte Ausrüstung 188–319 Tage.***



Quelle: Unternehmensunterlagen; Forderungen (inkl. Vertragsvermögen) ÷ Umsatz × 365.

AI-Nachprüfungs-Lizenzen, die über AOI-Hersteller oder Integratoren verkauft werden, sollten wie Mech-Minds 60–100 Tage eingehen; direkte Projektverkäufe wie die von Weiyi 190–320 Tage. Als Cash-Zeitstrahl zusammengesetzt: kostenloser Pilot ein bis drei Monate, Qualifizierung bei einem großen ODM bis zu etwa zwölf Monate, Aufbau, Lieferung und Abnahme drei bis sieben Monate; 30% gehen mit der Bestellung ein, 60% etwa vier bis neun Monate danach und die letzten 10% etwa 16–21 Monate danach. **Vom ersten Pilot bei einem großen EMS bis zur vollständigen Zahlung können zwei bis drei Jahre vergehen**; mittelgroße EMS ohne formelle Ausschreibungen können die Qualifizierung auf Wochen verkürzen estimate. Die Abnahmekriterien in den offengelegten Verträgen decken nur die Integration der gesamten Linie, den Durchsatz (ein taiwanisches Projekt verlangte 100 Einheiten pro Stunde) und Monate störungsfreien Lauf ab; keines nennt Durchschlupf- oder Fehlalarmraten (Anda inquiry reply) single source, primary. Schriftliche Abnahmekriterien für die AI-Nachprüfung müssen vor dem Pilot vereinbart werden.

Vier Muster des Scheiterns

Dokumentierte Misserfolge fallen in vier Muster: Auslieferung, bevor das Produkt zuverlässig war (Rethink (The Robot Report)); eine Softwareschicht auf der Hardware eines Marktführers, die dessen Ökosystem später ersetzte (READY Robotics (The Robot Report)); Geldknappheit am Boden eines Zyklus (Dextrous Robotics (Robots & Star-

tups)); und Modellunternehmen, die in Acqui-Hires enden (Covariant (GeekWire)). Die Lehren: eng produktisieren statt projektweise nachzujustieren; Betriebskapital für abnahmebasierte Umsätze und lange Zahlungsfristen vorhalten; lokalen Service für Werke bieten, die 24 Stunden laufen; und nicht skalieren, bevor das Produkt bewiesen ist.

Timing: Inspektion gewinnt nichts durch Warten; Arm-Hardware wird billiger

Die Argumente fürs Warten sind schwach. Die Hardware wurde 2026 nicht billiger: Roboterhersteller erhöhten die Preise um 5–15%, manche Produkte um 25–50% (Sina Finance) single source, und NVIDIA-Kits verteuerten sich ein drittes Mal (36Kr). Produktionsreife VLAs werden innerhalb des Fensters nicht eintreffen (IFR, Instrumental). In der Zwischenzeit entfernt jeder AOI-Ersatz ein Nachrüstziel, und jede OEM-AI-Bereitstellung bindet die Daten und Prozesse einer Linie; die Modellfähigkeit wird zur Ware, und knapp ist der Zugang zu Werken und beschrifteten Fabrikdaten. Auch lokale Förderung hat Fenster: Shenzhens Regeln zur »AI-Pionierstadt« 2026, mit bis zu 2 Mio. RMB für AI-Anwendungsdemonstrationsprojekte und Modellgutscheinen über 30% der Anschaffungen, laufen bis zum **30. Juni 2027** (Yicai, May 2026) terms single source.

Was weiterhin schnell billiger wird, sind Humanoide und geschickte Hände (Hände fielen in zwei Jahren von etwa 50,000 RMB auf unter 10,000 (Sina Finance)) sowie Einstiegs-Cobots, während sich der Produktmix nach unten verschiebt; das spricht dafür, den verkörperten Teil aufzuschieben, nicht den Inspektionsteil. **Insgesamt: Inspektion jetzt, als enges Produkt; Roboterzellen jetzt, über Partner und Integration; verkörperte AI später, heute nur eine F&E-Station.**

Verkörperte AI

06 Verkörperte AI: Zuerst ändert sie, wie Roboter unterrichtet werden, später wie sie sich bewegen

Was sich ändert, sind Wahrnehmung und Programmierung; die Bewegung bleibt deterministisch

Jeder Roboterarm-Marktführer vermarktet heute »Physical AI«, doch was ausgeliefert wurde, fällt in drei Ebenen: gelernte Wahrnehmung ohne Anlernen (FANUCs AI Masterless Vision, Auslieferung im Januar 2027 (FANUC); KUKAs neuronale Griffauswahl (KUKA)); große Modelle, die Programme schreiben und Aufgaben planen (KUKAs iiQWorks.Copilot, das warnt, dass AI-generierter Code in einem digitalen Zwilling verifiziert werden sollte (KUKA); Yaskawa mit Gemini Robotics-ER, wo »generative AI entscheidet, was zu tun ist« und der Roboter entscheidet, »wie er sich bewegt« (Yaskawa)); sowie Rechen- und Datenplattformen (URs AI Accelerator mit einem Jetson Orin (Universal Robots)) vendor claims. Kein Marktführer veröffentlicht Produktionserfolgsraten für seine AI-Funktionen, und die japanischen Marktführer kooperieren mit Google, SoftBank, Microsoft und anderen, statt eigene Frontier-VLAs zu bauen (FANUC news index): Die gelernte Policy-Ebene wird von Modellunternehmen kommen, die zertifizierten Arme und Steuerungen von den Armherstellern.

FIG 29 **Der Embodied-AI-Stack** ***Große Modelle planen offline, kleine Policys laufen in der Zelle, Bewegung und Sicherheit bleiben deterministisch.***

Ebene	Was es tut und wie weit die Evidenz reicht	Zeitskala · wer sie besitzt
Planung L4	**Frontier-Modelle (offline)** Roboterprogramme offline schreiben oder ändern, Umrüstungen, Fehler erklären Schreibt ABB-RAPID-Programme zu 77–99.7% um, nach Simulation bereitgestellt Regelkreis nur 0.2–0.4 Hz: bleibt aus der Echtzeit	**Sekunden bis Minuten** Modell-Giganten
Fähigkeiten / Policy L3	**ACT, Diffusionspolicies, kleine VLAs** Strukturierte Stationen mit Vorrichtung 90–99.4% Edge-Chips: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**Action Chunks 100–300 ms** Chip-Anbieter · Roboterherstell
Bewegungssteueru L2	**Robotersteuerung** Trajektorien, Servo, Kraftregelung, Kollisionsprüfung deterministischer Code; gelernte Policys setzen nur Ziele	**Millisekunden (Echtzeit brauc >83 Hz)** Roboterhersteller
Sicherheit L1	**Sicherheits-PLC, Scanner, Kraft- und Geschwindigkeitsbegrenzung** zertifiziert nach ISO 10218-1/-2:2025 und lokalen Vorschriften Kein ML in Sicherheitsfunktionen: EU-Bewertung durch benannte Stelle ab 20. Januar 2027	**harte Echtzeit** Zertifizierter Roboter + Integrat

Kleine Firmen gewinnen bei Integration und Sensorik-Hardware: gelernte Wahrnehmung und kleine Policys in zertifizierte Roboterzellen bringen

Quelle: Zusammenfassung des Autors (Kapitel 3 und 6); ISO 10218:2025; EU-Maschinenverordnung.

Wie weit entfernt von unstrukturierter, flexibler Kleinserienarbeit: strukturierte Stationen in 1–3 Jahren, flexible Montage 5+

Echte Fabrikzahlen tauchen langsam auf, aber sie sind anbietergemeldet, eng und stammen aus strukturierten Stationen. AgiBots acht Humanoide auf Longcheers Tablet-Endtestlinie protokollierten 64,828 Aufgaben bei 99.99%

über mehr als 64 Stunden (AgiBot, Yicai Global) vendor and customer, doch das ist Beschickungsarbeit, die ein konventioneller Arm mit Vision erledigen kann. UBTech gibt an, dass sein Walker S2 mit 30–50% der menschlichen Effizienz arbeitet (Irish Times/FT) single source. Figure 02s elf Monate bei BMW Spartanburg ergaben Volumenzahlen (1,250+ Stunden, 90,000+ Teile), während seine Zyklus-, Erfolgs- und Eingriffszahlen Ziele sind, keine Ergebnisse (Figure, BMW Group) verified, numerators only. Nucleus sagt, etwa 40% seiner Fabrikvideos seien teleoperiert (Interesting Engineering) vendor claim, und eine Reuters-Recherche nannte etwa 20% der Arbeitsleistung eines Arbeiters für manche chinesische Humanoide (Notebookcheck) single source. Kein Einsatz veröffentlicht Erfolgsrate, Zykluszeit und Eingriffsrate gemeinsam, und keiner wurde geprüft, doch genau diese drei sind es, was eine Amortisationsrechnung braucht.

Einsatz	Erfolgsrate	Takt	Eingriffe oder Teleoperation	Verfügbarkeit oder Ausfallintervall	Laufumfang
Figure 02 @ BMW Spartanburg	Nur Ziel (>99% pro Schicht)	Nur Ziel (84 s)	Nur Ziel (0 pro Schicht)	Nicht offengelegt (Unterarm als häufigste Ausfallstelle)	1,250 h
Agility Digit @ GXO	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt	100,000 Gebinde
Apptronik Apollo @ Mercedes-Benz	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt
Agile ONE @ Audi	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt	Nicht offengelegt
Nucleus (ungenanntes Werk)	Nicht offengelegt	Nicht offengelegt	≈40% teleoperiert (Methode nicht offengelegt)	Nicht offengelegt	≈2 h Video
AgiBot G2 @ Longcheer	99.99% von 64,828 Aufgaben	18–20 s pro Schritt	Nicht offengelegt	Ein Kommunikationsfehler, ≈2 min zur Wiederherstellung	64+ h
Neuromeka Hybridzelle	322/324	12.8% langsamer als ein Mensch	Nicht offengelegt	Nicht offengelegt	324 Zyklen
Physical Intelligence π0.6 an einem Kundenstandort	—	—	96.4% autonom über eine Schicht	Nicht offengelegt	Eine Schicht

Quellen: dieses Kapitel und Kapitel 3 mostly vendor claims.

FIG 30 **Was In-Plant-Roboterbereitstellungen tatsächlich offenlegen** ***Die meisten legen nur das Volumen offen; keiner Erfolgsrate, Takt und Interventionsrate gemeinsam.***

	Erfolgsrate	Takt	Eingriffe / Teleoperation	Verfügbarkeit / Fehler	Serienlauf
Figure 02 @ BMW	nur Ziel >99%/Schicht	nur Ziel 84 s	nur Ziel 0/Schicht	k. A.	1.250 h
Agility Digit @ GXO	k. A.	k. A.	k. A.	k. A.	100.000 Behälter
Apptronik Apollo @ Mercedes	k. A.	k. A.	k. A.	k. A.	k. A.
Agile ONE @ Audi	k. A.	k. A.	k. A.	k. A.	k. A.
Nucleus (ungenanntes Werk)	k. A.	k. A.	≈40% teleoperiert*	k. A.	≈2 h Video
AgiBot G2 @ Longcheer	99.99%	18–20 s pro Schritt	k. A.	ein Fehler, 2 min	64+ h
Neuromeka Hybridzelle	322/324	12.8% langsamer	k. A.	k. A.	324 Zyklen
π0.6 @ Kundeneinsatzort	—	—	96.4% autonom	k. A.	1 Schicht

* Methode nicht offengelegt

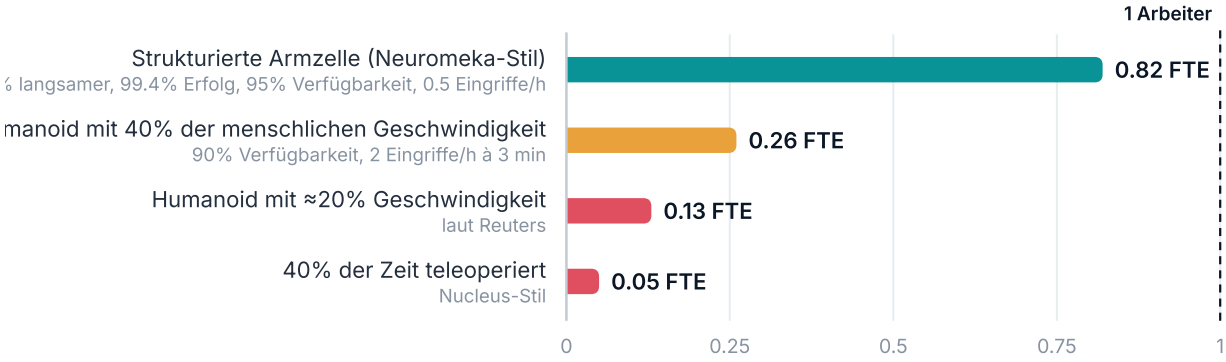
Offengelegt (meist Herstellerangaben) nur Ziele k. A

Quelle: Öffentliches Material von BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neuromeka, Physical Intelligence u. a.

Teleoperation wiegt für die Wirtschaftlichkeit schwerer als die Erfolgsrate. Netto ersetzte Personen pro Roboter pro Schicht ≈ (menschlicher Zyklus ÷ Roboterzyklus) × Verfügbarkeit × Erfolg, minus Eingriffszeit und Teleoperation estimate, my formula. Eine strukturierte Station nach Neuromeka-Art (12.8% langsamer, 99.4% Erfolg, 95% Verfügbarkeit, 0.5 Eingriffe von 2 Minuten pro Stunde) ersetzt etwa 0.82 Personen; ein Humanoid bei 40% der menschlichen Geschwindigkeit, 90% Verfügbarkeit und zwei 3-minütigen Eingriffen pro Stunde etwa 0.26; bei Reuters' 20% etwa 0.13; und wenn, wie bei Nucleus, ein Bediener ihn 40% der Zeit teleoperiert, etwa 0.05, fast nichts. Bei etwa 600k RMB pro Humanoid und 85k RMB pro Person und Jahr auf dem Festland amortisieren sich die letzten drei über Jahrzehnte; nur die strukturierte Armstation hat eine realistische Amortisation (Kapitel 4) estimate.

FIG 31 **Netto ersetzte Arbeitskraft pro Roboter** ***Eine strukturierte Armzelle ersetzt etwa 0,82 einer Arbeitskraft; ein zu 40% der Zeit teleoperierter Humanoider etwa 0,05.***

Netto ersetzte Arbeitskraft pro Roboter und Schicht



Quelle: Formelbasierte Schätzungen des Autors.

Abseits strukturierter Stationen erreichen allgemeine Modelle bei neuen Aufgaben weiterhin 40–90%: Gemini Robotics 2 erreicht 74.2% beim Labor-Pick-and-Place, 78.9% beim Werkzeug-Kitting und 44% beim Zubinden eines Beutels (eesel/DeepMind); bei RoboChallenge Table30, 30 Echtroboter-Aufgaben, erreicht π0.5 43.7% und das beste offene Modell 50.3% (RoboChallenge, QbitAI).

Externe Prognosen weichen voneinander ab, aber keine stützt einen breiten Einsatz innerhalb von zwei Jahren. Die IFR sagt, allgemeine Physical AI sei »noch nicht industrietauglich« und die meisten Humanoide seien »weiterhin

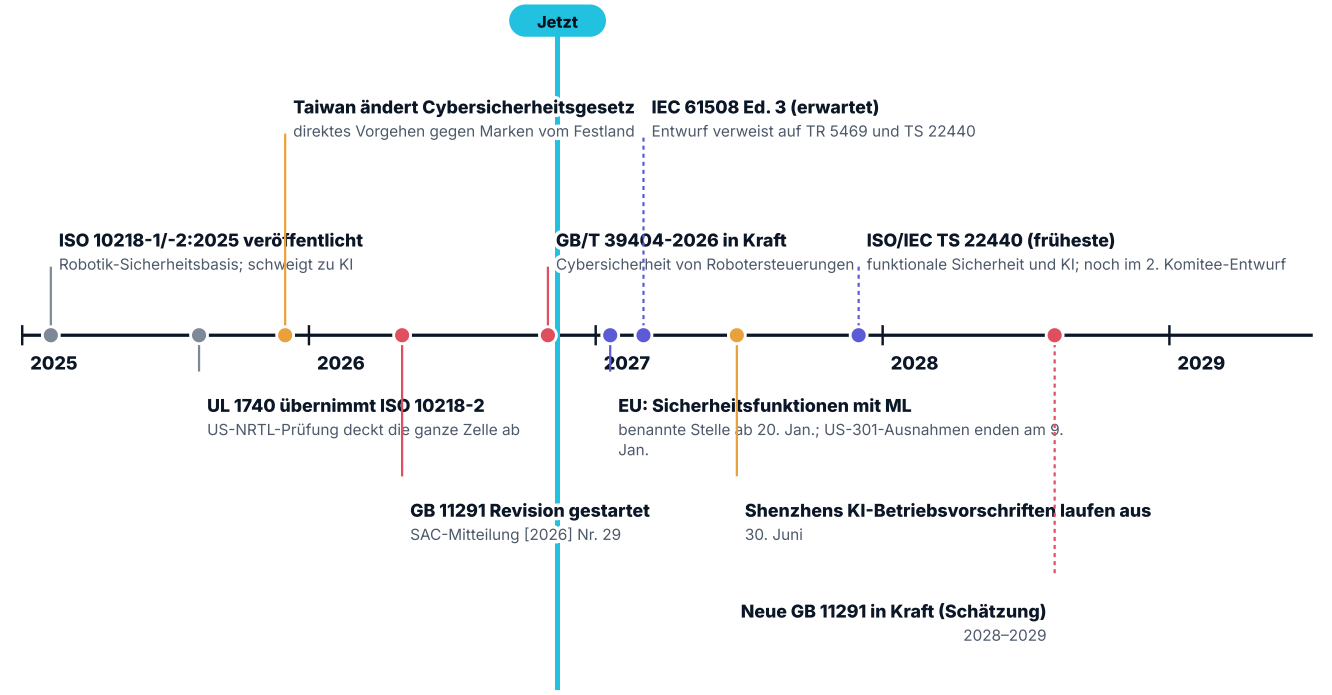
Prototypen oder Piloten« (IFR); Interact Analysis schätzt, dass nur etwa 10% der 2025 produzierten Humanoide in den echten Einsatz gingen, mit dem kommerziellen Wendepunkt »nach 2032« (automation.com); UBS hält einen Humanoid-»EV-Moment« vor 2030 für unwahrscheinlich (Yicai Global); chinesische Insider schätzen zwei bis drei weitere Jahre bis zum Einsatz im Fabrikmaßstab (Gasgoo) single sources. Mein Urteil estimate: In gut ausgestatteten Elektronik- und Automobilwerken mit dem Anbieter vor Ort werden **Aufgaben an strukturierten Stationen** (Beschicken, OK/NG-Sortierung, fixierte Insertion) in etwa 1–3 Jahren ≥99% bei annähernd menschlicher Geschwindigkeit erreichen; **Mixed-Part-Kitting** mit allgemeinen VLAs in etwa 3–5 Jahren; **frei geführte Kabelbäume und mehrstufige flexible Montage** in mehr als 5. Die Humanoid-Amortisation funktioniert heute nicht: Ein Walker S2 zu etwa 600k RMB (10jqka) single source entspricht bei 30–50% der menschlichen Effizienz über zwei Schichten etwa 0.6–1.0 Arbeitern, eine einfache Amortisation von 7–12 Jahren zu Festland-Arbeitskosten vor Integration, Wartung und Ausfallzeiten estimate.

Wo es hakt

Glied	Belege im Jahr 2026
Daten	Die Branche schätzt, dass mindestens 10 Millionen Stunden hochwertiger Echtdaten nötig sind und bis Anfang 2026 erst etwa 500,000 gesammelt wurden (Gasgoo); Reuters berichtet von etwa 500,000 Stunden gegenüber etwa 100 Millionen benötigten (Notebookcheck); ein Anfänger-Teleoperator braucht etwa 300 Versuche pro nutzbarer Bewegung; Werke teilen keine Prozessdaten; simulations-trainierte Policies fallen auf echten Robotern stark ab
Zuverlässigkeit	Strukturierte Stationen 98–99.99% (vendor claims, Fehler undefiniert); allgemeine Modelle 40–90% bei neuen Aufgaben; ungesehene Steckverbinder 29–67%
Geschwindigkeit	Humanoide bei 30–50% der menschlichen Geschwindigkeit (UBTech), manche bei etwa 20% (Reuters); Frontier-Agenten bei etwa 280 s pro quasistatischer Echtroboter-Aufgabe; LLM-Regelkreise bei 0.2–0.4 Hz gegenüber benötigten etwa 83 Hz (Anthropic)
Kosten	Walker S2 liegt im Schnitt bei etwa 600k RMB; 7–12-jährige Amortisation estimate
Sicherheitszertifizierung	ISO 10218-1/-2:2025 macht die funktionale Sicherheit expliziter und übernimmt kollaborative Anforderungen, sagt aber nichts über AI (The Robot Report); ISO/IEC TS 22440 (funktionale Sicherheit und AI) war im November 2026 noch ein zweiter Komitee-Entwurf (ISS, ISO Open Data); die IEC 61508-3-Leitlinie rät von AI oberhalb SIL 1 ab (Analog Devices); unter der EU-Maschinenverordnung benötigen Sicherheitsfunktionen mit maschinellem Lernen ab dem 20. Januar 2027 eine Bewertung durch eine notifizierte Stelle (EUR-Lex)
Hände und Werkzeuge	Geschickte Hände schaffen etwa 30 N (≈3 kg) pro Finger gegenüber etwa 10 kg bei einer menschlichen Hand (Gasgoo); Werkzeugwechsel sind weiterhin manuell
Integration	Sensor-Fehlablesungen, Protokoll-Mismatches und MES-Anbindungen vor Ort; Eingriffsraten werden fast nie offengelegt

Die Sicherheitsimplikation ist direkt: Die Normen verbieten gelernte Policies weder noch erkennen sie sie an. ISO 10218-1/-2:2025 sagt nichts über AI; ISO/IEC TS 22440 (funktionale Sicherheit und AI) ist weiterhin im Komitee-Entwurf und vor Ende 2027 oder 2028 unwahrscheinlich estimate; ISO/IEC TR 5469:2024 ist nur eine Leitlinie und empfiehlt Nicht-AI-Fallbacks, überwachende Einschränkungen und redundantes Abstimmen; IEC 61508 Ausgabe 3 wird Anfang 2027 erwartet (ISO Open Data, BSI, 61508 Association) verified. Noch keine veröffentlichte Norm bietet einen akzeptierten Weg, maschinelles Lernen als Sicherheitsfunktion zu zertifizieren. Auch regionale Regeln bewegen sich: China überarbeitet GB 11291 (in Kraft etwa 2028–2029) (revision plan 20262543); UL 1740 hat ISO 10218-2 seit August 2025 in die US-amerikanische NRTL-Prüfung aufgenommen; die EU verlangt ab dem 20. Januar 2027 die Bewertung von Sicherheitsfunktionen mit maschinellem Lernen durch eine notifizierte Stelle; und kein Humanoid ist für den zaunlosen Fabrikeinsatz zertifiziert, solange ISO 25785-1 noch im Entwurf ist (i-SCOOP). Die vertretbare Architektur: Maschinelles Lernen trifft nur Prozessentscheidungen (»welches Teil«, »ist diese Platine gut«), während eine zertifizierte deterministische Sicherheitsebene (typischerweise PL d / Kategorie 3) und eine Sicherheits-SPS über Geschwindigkeit, Kraft und Stopps entscheiden. Behandeln Sie jede Behauptung von »Kamera-AI für zaunlose Sicherheit« als unbewiesen, bis ein PL-d- oder SIL-2-Zertifikat die Komponente des maschinellen Lernens benennt.

FIG 32 **Zeitstrahl zu Normen und Regulierung***Ein Zertifizierungspfad für KI innerhalb von Sicherheitsfunktionen wird sich vor 2027–2029 nicht stabilisieren.*



Quelle: ISO, IEC, UL, EUR-Lex, SAC und nationale Regulierungsdatenbanken.

Welche Ebene die Chance ist: Modelle gehören den Giganten, Edge-Boxen den Chipherstellern, kleinen Firmen Integration und Sensor-Hardware

Die Modellebene braucht Kapital und Daten und gehört den Giganten (AgiBot trainierte GO-2 auf »Zehntausenden von Stunden« (iFeng Tech); die stärksten Modelle sind geschlossen oder nur in der Cloud). Die Edge-Box-Ebene gehört den Chipherstellern: Jetson Thor ist der De-facto-Standard, und auf dem Festland konkurrieren D-Robotics' RDK-Boards (D-Robotics) und Rockchip-Koprozessoren (CSDN), meist ohne veröffentlichte VLA-Ergebnisse.

Kleine Firmen haben zwei Öffnungen. Eine ist Sensor-Hardware, die direkt Zuverlässigkeit erkauft: Berührung hebt ContactDPs Insertionserfolg von 58.7% auf 97.8%, und Kraft-/Drehmomentsensoren sowie Handgelenkkameras leisten Ähnliches; auch Datenerfassungs-Kits verkaufen sich (10jqka), obwohl Hände, Sensoren und Arme alle im Preis fallen. Die andere ist die Integrationsebene zwischen Armherstellern und Modellunternehmen, die die Aufgabendaten, die Sicherheitshülle und die MES-Anbindung jeder Station besitzt, was Armhersteller heute nicht als Produkt verkaufen. Wettbewerber fallen in vier Gruppen: Arm-Marktführer (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot und andere), Vision-plus-Arm-Unternehmen (Mech-Mind, Hikrobot, SmartMore), Unternehmen für verkörperte AI (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) und Neueinsteiger im Automotive-Chip-Bereich.

Käufer-Akzeptanz: keine Zahl ohne Nenner

Da kein Anbieter Eingriffsraten, Ausfallintervalle oder langfristige Erfolgszahlen veröffentlicht, gehören diese in die Abnahmebedingungen des Vertrags. Statistisch belegen 3,000 aufeinanderfolgende Produktionszyklen mit null Fehlern eine Ausfallrate von höchstens 0.1% bei 95% Konfidenz; 10 Fehler in 3,000 belegen nur ≤0.57%; der Nachweis von ≤1%, ≤0.5%, ≤0.1% und ≤0.01% mit null Fehlern erfordert 299, 598, 2,995 und 29,956 Zyklen estimate, exact binomial. Nach diesem Maßstab hat Neuromekas 322/324 ein zweiseitiges 95%-Intervall von 97.8–99.9% und belegt nicht ≥99%; AgiBots grob 6 Fehler in 64,828 Aufgaben implizieren 99.98–100%, aber nur, wenn »Fehler« so

definiert ist, wie der Käufer ihn definiert estimate. Das sind die Abnahmeanforderungen, die ich an jede Roboterarm- oder verkörperte Zelle stellen würde; die Schwellen stützen sich auf Figures eigene Ziele bei BMW und auf die Fälle Neuromeka und AgiBot, und sie sind Urteil, keine Norm estimate:

Punkt	Anforderung	Vorgeschlagene Bestehensschwelle
Definitionen	Unterzeichnete Definitionen von Zyklus, Erfolg, Fehler (Teilerfolg, erneutes Greifen, Fallenlassen, Qualitäts-Durchschlupf), Eingriff (Berührung, Fernteleoperation, Reset, Not-Halt) und wie die Autonomie berechnet wird	Definitionen kommen vor jeder Metrik; kein Prozentsatz ohne Formel
Eingriffe und Teleoperation	Pro Roboterstunde, nach Art und Dauer; mittlere Zeit zwischen Eingriffen; Teleoperationsminuten als Anteil der Laufzeit; Bediener pro Roboter	Strukturierte Stationen ≤0.5 Eingriffe pro Roboterstunde (≥2 h Abstand); null Teleoperation während der Abnahme
Langfristiger Erfolg	≥3,000 aufeinanderfolgende Produktionszyklen über ≥3 Schichten, ≥2 Teilenummern und normale Beleuchtungs- und Versorgungsschwankungen; Fehler klassifiziert; exakte Untergrenze	Untergrenze ≥99.5% für linienstoppende Inline-Stationen, ≥99.0% offline; Qualitäts-Durchschlüpfe separat gemeldet
Zykluszeit	P50/P90/P99 pro Teil einschließlich Wiederherstellung, gegenüber ≥30 gestoppten menschlichen Zyklen an derselben Station	≤1.15 × menschliches P50 oder innerhalb des Linientakts
Verfügbarkeit und Fehler	Laufstunden, Ausfallzeitprotokoll, Fehler nach Komponente, Teiletauschzeit	≥95% Verfügbarkeit über ≥2 Wochen; MTTR ≤20 min mit Ersatzteilen vor Ort
Wiederherstellung	Skriptgesteuerte Fehlerinjektion: Fallenlassen, Verdeckung, Beleuchtungswechsel, Förderbandstopp, MES- oder Netzwerkausfall, Stromausfall, Not-Halt	≥90% autonome Wiederherstellung, zurück in Produktion innerhalb von 2 min
Produktwechsel	Stunden, Demonstrationen, Bedienerfertigkeit und Datenerfassungskosten pro neuer Teilenummer oder Vorrichtung	Strukturierte Stationen: ≤1 Schicht und ≤100 Demos pro neuer Teilenummer
Sicherheit	Arm und Zelle nach ISO 10218-1/-2:2025, Zellen-Risikobewertung, lokale Regeln (GB 11291 auf dem Festland; CNS 14490 plus schriftliche Bewertung in Taiwan; ANSI/A3 R15.06-2025 und UL 1740 in den USA)	Keine Sicherheitsfunktion durch maschinelles Lernen umgesetzt; eine zertifizierte Sicherheitsebene (typischerweise PL d / Kat. 3) begrenzt Geschwindigkeit und Kraft; keine zaunlosen Laufroboter, bis ISO 25785-1 veröffentlicht und bewertet ist
Rechenleistung und Offline-Verhalten	Inferenz-Hardware, P99-Latenz, Verhalten ohne Netzwerk, Modellversionierung, erneute Validierung nach jedem Update	Taktkritische Inferenz hängt nicht von der Cloud ab; vollständige erneute Abnahme nach jeder Modell-, Quantisierungs- oder Engine-Änderung
Belege	Ungeschnittenes Video und Rohprotokolle für das gesamte Abnahmefenster; bezeugt vom Qualitätsteam des Kunden oder einem Dritten	Keine Highlight-Zusammenschnitte oder rein anbieterseitige Zusammenfassungen
Kommerzielle Bedingungen	Zahlungsmeilensteine an Abnahmemetriken gekoppelt; Verfügbarkeitsstrafen bei Abrechnung pro Roboterstunde; Dateneigentum und Regeln für den Fernzugriff	Schlusszahlung und Einbehalt erst nach der 3,000-Zyklen-Abnahme und ≥4 Wochen stabilem Lauf freigegeben
Wirtschaftlichkeit	Netto ersetzte Personen = Geschwindigkeitsverhältnis × Verfügbarkeit × Erfolg − Eingriffszeit − Teleoperationsanteil	Lehnen Sie jeden Vorschlag ab, dessen Nettoersatz seine behauptete Amortisation nicht stützen kann

Quellen: (Figure, Neuromeka, Yicai Global, i-SCOOP); die Schwellen sind meine Vorschläge.

Offene Fragen

07 Was die öffentliche Dokumentation nicht beantworten kann und wie ein Pilot es klärt

Die Fragen, die öffentliche Quellen nicht beantworten können, fallen in fünf Gruppen: wofür sich die AI-Nachprüfung tatsächlich verkauft und was Käufer zahlen; wie AOI-Hersteller AI bündeln und alte Maschinen aufrüsten; wie alt die installierten AOIs sind und wie viele bereits mit AI laufen; wie viele Personen und Stunden die Nachprüfung pro Linie wirklich kostet; und wie Piloten konvertieren und wie Käufer zahlen. Dazu kommt, wie die neuesten Modelle, quantisierten Builds und Roboterarme unter den Bedingungen eines bestimmten Werks tatsächlich abschneiden. Jede Frage lässt sich in einen Test mit einer klaren Schwelle verwandeln. Die Schwellen unten sind Entscheidungsregeln, die ich aus den Belegen in diesem Bericht abgeleitet habe estimate; berechnen Sie sie neu, sobald die ersten Werkdaten eintreffen.

Offene Frage	Wie man sie klärt	Wie ein gutes Ergebnis aussieht
Preis und Zahlungsbereitschaft für die AI-Nachprüfung	Eine standardisierte Angebotsanfrage (zum Beispiel vier Linien mit gemischten älteren 2D- und neueren 3D-AOIs, zwei Schichten, etwa 2,000 Platinen pro Schicht, 12 Defektklassen, nur vor Ort, MES- oder IPC-CFX-Integration) an mehrere Drittanbieter; eine Preissensitivitätsumfrage mit vier Fragen an 20–30 EMS-Werke	Die meisten Angebote innerhalb von 40–120k RMB einmalig oder 20–40k RMB pro Jahr für vier Linien; mindestens die Hälfte der Werke akzeptiert 10–30k RMB pro Linie bei Amortisation innerhalb von 12 Monaten
AI-Bündelung und Upgrades der AOI-Hersteller für ältere Maschinen	Kurze strukturierte Interviews mit etwa einem Dutzend AOI-Herstellern und Distributoren	Ob die AI-Nachprüfung bei Modellen 2025–26 Standard, optional oder nicht vorhanden ist; ob Maschinen vor 2020 aufgerüstet werden können und zu welchem Preis
Alter der installierten AOIs und Anteil, der mit AI läuft	Vor-Ort-Umfrage in etwa 30 EMS-Werken, geschichtet nach Größe und Region, mit unter NDA geprüften Ausrüstungslisten	Medianes AOI-Alter ≥5 Jahre und ≥60% ohne AI bedeuten, dass der Nachrüstbestand real ist
Nachprüfungs-Arbeitskräfte und -Stunden	Interviews und 2–4 Wochen AOI-Alarm-, Nachprüfungs-, Reparatur- und Schichtprotokolle in zwei oder mehr Werken, mit Arbeitsstichproben	≥1 dedizierter Prüfer pro Linie und Schicht und ≥3 Nachprüfungsstunden pro Linie und Schicht, mit mindestens 0.5 FTE pro Linie, die das Werk bereit ist umzuverteilen
Leistung der AI-Nachprüfung selbst	Golden Set plus Schattenmodus auf einer echten Linie (Kapitel 3): Datenvereinbarung in Woche 0, Baseline und Beschriftung in Wochen 1–2, Schattenbetrieb in Wochen 2–4, optionaler assistierter Betrieb in Wochen 4–8	Fehlalarme oder Nachprüfungsstunden um ≥50–60% gesenkt; Durchschlupf-Grenzen je Klasse nicht schlechter als die menschliche Baseline; null Übersehene bei kritischen Klassen; p95-Latenz ≤½ Takt; automatisches Fail-Open zu Menschen
Lehrer-Label-Qualität der Frontier-Modelle	Zwei oder drei Kandidaten auf demselben Golden Set laufen lassen, je drei Prompt-Formulierungen, ein gutes Referenzbild, Einzelkomponenten-Ausschnitte ≥42–48 px	Automatische Übernahme nur für Klassen mit null Übersehenen bei ≥299 Defekten und Übereinstimmung auf Menschenniveau; sonst nur Vorab-Beschriftung und Triage

Offene Frage	Wie man sie klärt	Wie ein gutes Ergebnis aussieht
Quantisierte Builds	Nur die Präzision ändern; Gewichts-Hashes, Vorlage, Vorverarbeitung, Decoding, Engine und Batchgröße festlegen; Prefix-Caching deaktivieren oder zurücksetzen; prüfen, ob der Referenz-Build die Anbieterergebnisse reproduziert; Paare auf dem Golden Set und dem Drift-Slice bewerten	Keine neuen Übersehenen bei kritischen Klassen; ≤1 neues Übersehen bei 1,000 Defekten; Abstand um ≤2 Punkte gesunken; ≥99.5% identische Entscheidungen über drei Läufe; sonst auf FP8/INT8 zurückfallen
Linienseitige Hardware, Sauberkeit und ESD	≥2 Einheiten jeder Kandidatenbox unter doppelter erwarteter Alarmrate belastungsgetestet; 72-Stunden-Soaktest bei realem Schrankluftstrom; ESD-Prüfung durch den Koordinator des Kunden; ISO-14644-14-Test für Kunden mit ISO 7 und sauberer	p95 ≤½ Takt; keine Abstürze oder Drosselung über 72 Stunden; ESD-Freigabe; Partikelbericht in der richtigen Klasse
Zahlungsbedingungen, Pilotregeln und Konversion	Zahlungsklauseln aus jüngeren Inspektionsausrüstungs-Verträgen in den untersuchten Werken; Interviews zu Pilotregeln, Abnahmekriterien und Qualifizierungszeit	Etwa 30/60/10 mit Bargeld innerhalb von 60 Tagen nach Abnahme; Piloten ≤8–12 Wochen mit einem namentlich benannten Freigabeinhaber; ≥1 von 3 Versuchen konvertiert; ein Rahmenvertrag über ≥5–10 Linien im ersten Jahr
Daten, IT und Haftung	Interviews und Vertragsprüfung mit Rechtsberatung	Eine lokale Box darf sich verbinden und auf Daten zugreifen; Haftung auf Softwarefehler beschränkt
Regionale Compliance vor dem Auslandsgang	Taiwan: schriftlicher Status des Kunden nach dem Cyber Security Management Act, E187-Selbstbewertung; USA: Zollentscheidung zur Klassifizierung einer Zelle, NRTL-Angebote, ITAR-Prüfung; Südostasien: BOI-Local-Linkage und malaysische Freibetragsberechtigung, Muttergesellschaft des Kunden; Festland: CCC-Umfang für gebündelte PCs und Netzteile	Der Kunde ist keine kontrollierte Stelle, und die Amortisation bleibt ≤2 Jahre einschließlich Zoll und Zertifizierung
Abnahme eines Roboters oder einer verkörperten Zelle	Die 12-Punkte-Abnahmetabelle in Kapitel 6: unterzeichnete Definitionen, ≥3,000 aufeinanderfolgende Produktionszyklen über ≥3 Schichten und ≥2 Teilenummern, skriptgesteuerte Fehlerinjektion, ≥30 gestoppte menschliche Zyklen, bezeugt	Untergrenze ≥99.5% inline (≥99.0% offline); Zyklus ≤1.15 × Mensch; ≤0.5 Eingriffe pro Stunde und null Teleoperation; ≥95% Verfügbarkeit; MTTR ≤20 min; ≥90% Wiederherstellung
ACT- und VLA-Policies auf chinesischen Edge-Chips	Vier Wochen: Portierung und Mikro-Benchmark (≥1,000 Inferenzen, Latenzverteilung, Leistung, 60-minütiger Soaktest); Closed-Loop-Simulation mit injizierter gemessener Latenz; Echtroboter-A/B gegen eine GPU; 8–24 Stunden Dauerlauf (≥3,000 Zyklen)	ACT-Chunks p95 ≤100 ms, VLA ≤300 ms; Echtroboter-Erfolg innerhalb von 3 Punkten der GPU und Zyklus ≤1.2 × Mensch; sonst die Policy auf Jetson oder einer GPU behalten
Cobot-, AMR- und Integrationspreise	Modellweise Angebote von mehreren Cobot-Distributoren und Vollzellen-Angebote von zwei oder drei Integratoren für die Testvorrichtungs-Beladung (eine Station, zwei Schichten)	Vollzelle ≤200k RMB und Amortisation ≤1.5 Jahre über zwei Schichten in Shenzhen
SMT-Linienzahlen in Taiwan, Südostasien und den USA	IPC-Daten (früher New Venture Research) oder die installierte Basis lokaler Ausrüstungsdistributoren	Genügend adressierbare Linien pro Region, um lokales Anwendungs-Engineering zu rechtfertigen

Von all diesen entscheiden vier über die sensibelsten Zahlen in diesem Bericht: der Preis, den Käufer zahlen, die Arbeitskraft, die die AI-Nachprüfung tatsächlich freisetzt, ob die AI-Nachprüfung wirklich nicht mehr übersieht als Menschen und wie schnell das Bargeld zurückkommt. Es lohnt sich, sie vor jeder größeren Bindung zu klären — von einem Werk, das einen Kauf erwägt, oder von einem Anbieter, der ein Produkt erwägt.

Eine Pilotabfolge, die für beide Seiten funktioniert

Für ein Werk und für einen Anbieter, der eine glaubwürdige Referenz möchte, ist die Reihenfolge dieselbe:

1. **Wochen 0–2: Daten und Definitionen vereinbaren.** Die Datenvereinbarung unterzeichnen, die Defektklassen und was als Durchschlupf gilt definieren und eine Baseline erfassen: Fehlalarme pro Platine, Nachprüfungssekunden pro Aufruf, nachgelagert gefundene Durchschlüpfe.
2. **Wochen 1–4: das Golden Set aufbauen** aus dem AOI-Archiv und zurückbehaltenen NG-Platinen, blind beschriftet und beurteilt (Kapitel 3).
3. **Wochen 2–4 (oder länger): Schattenmodus.** Das Modell protokolliert seine Entscheidungen, während Menschen wie zuvor entscheiden; je Klasse vergleichen.
4. **Wochen 4–8: assistierter Betrieb in zufälligen Schichten,** wobei das Modell nur Klassen freigibt, die die Hürde bestanden haben, und alles andere an Menschen geht.
5. **Abnahme, schriftlich, vor der Skalierung:** die Ergebnisse je Klasse, die Latenz, das Fail-Open-Verhalten und wem Modelle und Labels gehören. Linie für Linie skalieren und bei jedem neuen Produkt, jeder Modellversion, jeder Quantisierungs- oder Engine-Änderung erneut validieren.

Grenzen dieses Berichts

Dies ist Schreibtischrecherche. Für diese Ausgabe habe ich keine Primärinterviews mit Werken geführt; viele Liniendaten stammen von Anbietern und wurden nicht unabhängig geprüft. Mehrere Regierungswebsites, Ausschreibungsdatenbanken und Rekrutierungsplattformen waren nicht zugänglich, daher sind Shenzhens Upgrade-Subventionssätze 2026, die Lohnleitlinie des Jangtse-Deltas, Transaktionspreise für die AI-Nachprüfung, Prüfer pro Linie und Pilot-Konversionsraten Schätzungen oder Analogien; jede erscheint oben als offene Frage. Wechselkurse sind der Zentralkurs der People's Bank of China vom 9. November 2026 ($1 = RMB 6.7188) und, für Taiwan, die Fixierung von SMBC am selben Tag ($1 = NT$32.12); der vietnamesische Dong verwendet Daten vom Mai 2026. Alle Schwellen sind Entscheidungsregeln, abgeleitet aus den zitierten Belegen, keine Normen.

Über den Autor

Tan Shuai

Tan Shuai ist Technologe und Gründer mit mehr als zwanzig Jahren Arbeit über Chips, Hardware und Systemsoftware hinweg und 20 erteilten Patenten. Seine Arbeit hat ihn durch das Silicon Valley, New York, Chicago, Taiwan, die Greater Bay Area (Shenzhen und Hongkong), Vietnam und Malaysia geführt, mit einem langen Fokus auf fortschrittliche Fertigung und Lieferketten. Heute arbeitet er als AI-Forscher und -Entwickler und konzentriert sich darauf, wie AI tatsächlich in der Fabrikhalle eingesetzt wird.

Er freut sich über Gespräche mit Menschen, die an AI in der Fertigung arbeiten, ob beim Aufbau eines Teams, bei der Partnersuche oder beim Erfahrungsaustausch: i@tanshuai.com · tanshuai.com



Über den Autor

Tan Shuai

Dieser Bericht ist unabhängige Schreibtischrecherche auf Basis öffentlicher Dokumente, die bis zum 11. November 2026 verfügbar waren. Er wird nur zu Informationszwecken bereitgestellt und ist keine Anlage-, Rechts-, Steuer- oder Engineering-Beratung. Produktnamen, Preise, Modellveröffentlichungen, Exportklassifizierungen, Zölle und Vorschriften ändern sich schnell; verifizieren Sie alles, worauf Sie sich verlassen, anhand der Primärquelle. Zahlen mit der Kennzeichnung vendor claim oder estimate sind entsprechend zu lesen. Unternehmens- und Produktnamen gehören ihren Eigentümern und werden nur zur Identifikation verwendet; kein Unternehmen hat diesen Bericht geprüft oder gesponsert.

© 2026 Tan Shuai. Sie dürfen diesen Bericht unverändert und mit Quellenangabe teilen.

ELEKTRONIKFERTIGUNG · AI-INSPEKTION × ROBOTIK

AI-Sichtprüfung und Roboterarme in Elektronikfabriken: Bewertung 2026

Jede Schlüsselzahl trägt ihre Quelle und einen Vertrauensgrad: verified, single source, vendor claim oder estimate. Was die öffentliche Dokumentation nicht beantworten kann, ist in Kapitel 7 gesammelt, zusammen mit der Frage, wie ein Pilotprojekt jede Frage klärt.

Tan Shuai
tanshuai.comi@tanshuai.com

ÖFFENTLICHE AUSGABE · DATENSTAND 11. NOVEMBER 2026

