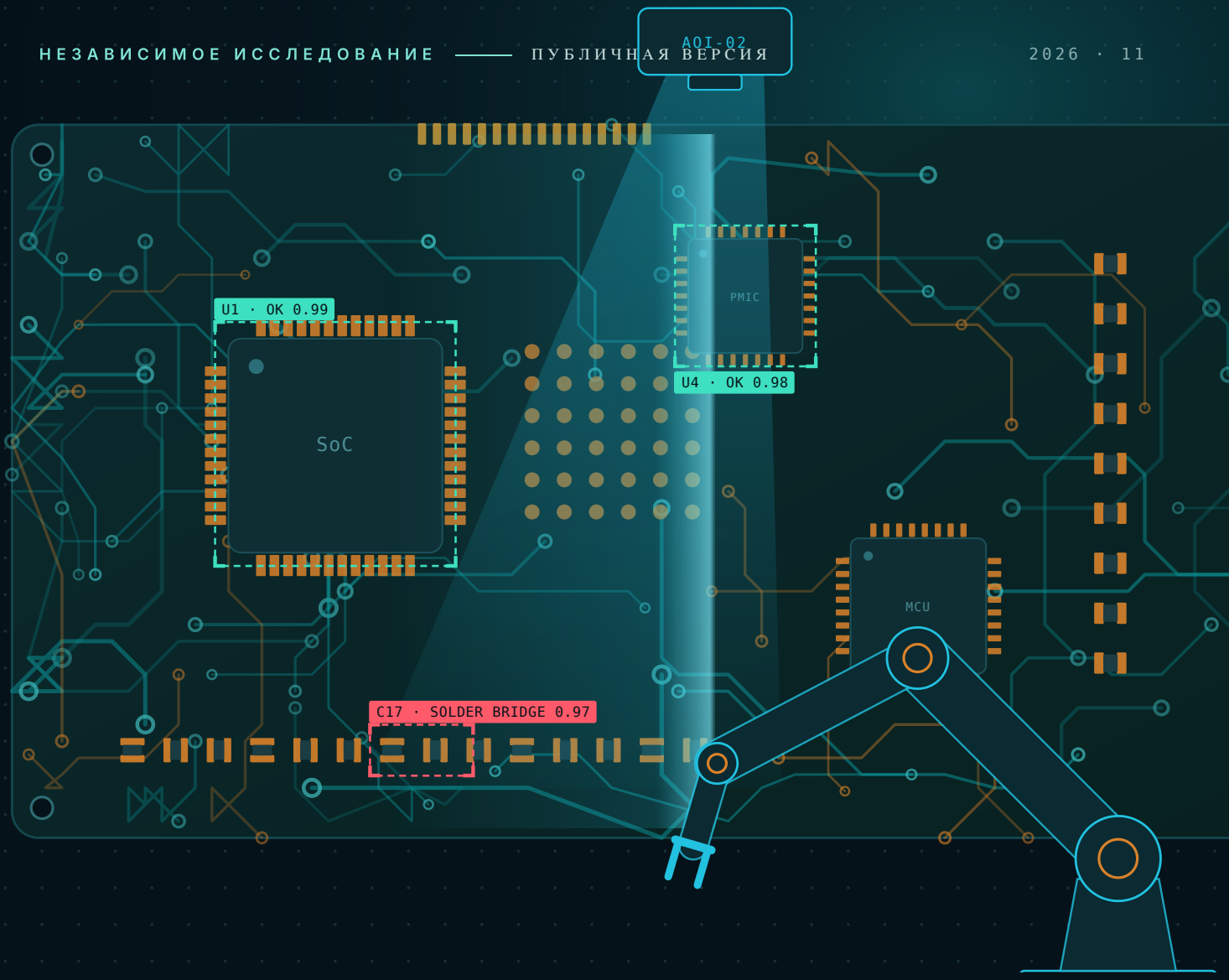


ПРОИЗВОДСТВО ЭЛЕКТРОНИКИ · ИИ-ИНСПЕКЦИЯ × РОБОТОТЕХНИКА

Визуальный контроль на основе ИИ и роботы-манипуляторы на заводах электроники

Оценка 2026 года

Основные конфигурации · Региональная применимость · Локальное развёртывание · Модели «учитель–ученик»

Окупаемость инвестиций · Конкуренция · То, на что ответит только пилот

Tan Shuai
tanshuai.comi@tanshuai.com

Данные на 11 ноября 2026 года
Материковый Китай · Тайвань · Юго-Восточная Азия · США

СОДЕРЖАНИЕ

Содержание



◆ **КРАТКОЕ СОДЕРЖАНИЕ** — Стоит делать, но узко, сейчас и с партнёрами — 5

01 **Линия сегодня** — Что сегодня запускают заводы электроники и что технически осуществимо — 9

02 **Локальное развёртывание** — Запуск ИИ внутри завода: чистые комнаты, воздушные зазоры, оборудование и квантование — 18

03 **Учитель и ученик** — Учитель и ученик: фронтир-модели могут предразмечать данные инспекции; для манипуляторов настоящие учителя — по-прежнему люди — 28

04 **Затраты и окупаемость** — Затраты и окупаемость: AI-перепроверка окупается за месяцы, обслуживание станков — за годы, манипуляторы для отбора — никогда — 37

05 **Рынок и конкуренция** — Рынок, конкуренция и как платят покупатели — 45

06 **Воплощённый ИИ** — Воплощённый ИИ: сначала он меняет то, как учат роботов, позже — как они двигаются — 54

07 **Открытые вопросы** — На что публичные источники не дают ответа и как это решает пилот — 61

◆ **Об авторе** — Tan Shuai — 64

Метки уверенности в тексте

- **verified** — совпадают два или более независимых источника либо два независимых первичных документа
- **single source** — только один источник, включая первичные документы
- **vendor claim** — опубликовано продавцом продукта или услуги
- **estimate** — арифметика или вывод автора из процитированных данных

Как читать этот отчёт

Ключевые цифры несут метку уверенности. **verified** означает, что совпадают два или более независимых источника либо два независимых первичных документа; **single source** — что источник только один (включая первичные документы); **vendor claim** — что опубликовано продавцом продукта или услуги, даже если широко повторяется; **estimate** — моя собственная арифметика или вывод из процитированных данных, с указанием исходных данных. «Первичный» описывает тип источника (годовой отчёт, проспект, нормативный акт, стандарт, официальная статистика); это не меняет класс уверенности. Там, где источники расходятся, цифры приводятся рядом и не усредняются. Конвертация валют использует центральный курс Народного банка Китая от 9 ноября 2026 года, $1 = 6.7188 RMB (Xinhua), а для Тайваня — фиксинг SMBC того же дня, $1 = 32.12 NT$ (SMBC). Данные о труде и субсидиях в материковом Китае используют дельту Жемчужной реки, где производство электроники наиболее сконцентрировано (официальные данные о зарплатах из Шэньчжэня), с дельтой реки Янцзы для сравнения. Утверждения из отраслевых интервью рассматриваются как утверждения, которые нужно проверять, а не как факты. Вопросы, на которые публичные источники не дают ответа, собраны в главе 7.

РИСУНКИ

Рисунки



01	Где ИИ вписывается в заводской цех сегодня	7
02	Серийная SMT-линия: люди сидят за AOI	10
03	Около 97–99% срабатываний AOI — ложные	11
04	Универсальная камера — это не AOI	13
05	Изолированное развёртывание	18
06	Зоны чистоты и ESD определяют, где могут стоять вычисления	19
07	Пороги скорости по применению	21
08	Локальные вычисления: память, цена и что легально купить в материковом Китае	23
09	Одна большая рабочая станция не создаёт общего ассистента	24
10	Квантование: 8-bit безопасен, 4-bit зависит от рецепта, 3-bit и ниже — вне игры	25
11	Кандидаты-учителя: универсальные баллы удвоились за поколение	29
12	«Есть ли дефект?» всё ещё сложно для универсальных моделей	30
13	Два конвейера учитель–ученик	31
14	Насколько большим должен быть золотой набор	32
15	Задержка краевой политики	34
16	Роботы-ученики: почти производственный уровень только на структурированных станциях	35
17	Сколько стоит работник инспекции или погрузки в год	38
18	Ценовая лестница	39
19	Окупаемость по регионам	41
20	Почему отбор плохих плат не окупается	42
21	Пошлины США на роботов, AOI и краевые компьютеры	43
22	Размеры рынка различаются более чем в десять раз	46
23	Снизу вверх: 40 000 линий → 16 000 обслуживаемых → ≈RMB 320M разово	47
24	Окно модернизации — около пяти лет	48
25	Клин перепроверки окружён с трёх сторон	48
26	Маржа — в продуктовом железе	50
27	Как платят покупатели	51
28	Оборачиваемость дебиторской задолженности	52
29	Стек воплощённого ИИ	54
30	Что на самом деле раскрывают внутризаводские развёртывания роботов	56
31	Чистое замещение труда на робота	56
32	Хронология стандартов и регулирования	58

КРАТКОЕ СОДЕРЖАНИЕ

Стоит делать, но узко, сейчас и с партнёрами

КЛЮЧЕВОЙ ВЫВОД

Вердикт: для заводов электроники то, что зрело в 2026 и за что заводы будут платить, — это AI-перепроверка и нишевая инспекция на изображениях AOI плюс программируемое обслуживание станков роботами, а не «большая модель, смотрящая на платы, плюс воплощённые роботы, которые их разбирают».

0.46–0.65 лет
Окупаемость AI-перепроверки в США, на Тайване и в Шэньчжэне при 50 тыс. RMB на линию

≥95% · <20 PPM
Покрытие и пропуски внутренней AI-перепроверки ведущего EMS: планка для продаж EMS первого эшелона

100–700 млн RMB
Обслуживаемый рынок сторонней модернизации в материковом Китае, разово, в окне около пяти лет

0 баллов
Опубликованные баллы флагманов ноября 2026 на промышленных дефектах: учителя, а не инспекторы

1.1–1.5 лет
Окупаемость ячейки обслуживания станков коботом в две смены в США, на Тайване и в Шэньчжэне

0.003–0.14 FTE
Труд, который экономит манипулятор для отбора за смену на линии SMT: он никогда не окупается

Инспекция
ROI

Контролируемое глубокое обучение, применённое к изображениям, которые уже производит существующий AOI, сокращает ложные срабатывания или часы перепроверки на 60–90% (vendor claims). DBG, крупная EMS-компания, сообщает в годовом отчёте, что её внутренняя система покрывает более 95% отмеченных компонентов с пропусками ниже 20 PPM и сокращением ручной перепроверки более чем на 85%; это планка, которую должен преодолеть сторонний продукт, чтобы продаваться EMS первого эшелона. При 50 000 RMB на линию AI-перепроверка окупается в пределах года в материковом Китае, на Тайване и в США (Шэньчжэнь 0.65 года, Тайвань 0.53, США 0.46), ячейки обслуживания станков коботами — за 1.1–1.5 года, а в Юго-Восточной Азии — за 2–6 лет, тогда как манипулятор, добавленный к линии SMT только для отбора бракованных плат, не окупается нигде и никогда, потому что при выходе годных с первого прохода выше 90% отбор занимает лишь 0.003–0.14 человека за смену estimate .

Модели
РОБОТЫ

Мультимодальные большие модели по-прежнему не могут быть основным инспектором. По состоянию на 11 ноября 2026 года ни один из флагманов (Claude Fable 5.2, GPT-6.x, Gemini 4, Qwen3.8 и другие) не имеет опубликованного балла ни на одном бенчмарке промышленных дефектов или PCB, а в задаче ACM MM 2026 PCBA пять лучших систем, обученных на 6 600 реальных изображениях линии, набрали лишь 77–82 в задаче «есть ли дефект». Фронтир-модели могут предразмечать и сортировать. Для манипуляторов настоящие учителя — это по-прежнему человеческие демонстрации, обучение с подкреплением на реальном роботе и симуляция: ученики достигают 90–99.4% на закреплённых структурированных станциях, но лишь 30–80% на движущихся конвейерах, незнакомых деталях и новых продуктах; фронтир-

агент выполнил 104 из 108 квазистатических задач на реальном роботе, по 280 секунд каждая.

Локально
РАЗВЁРТЫВАНИЕ

Изолированное развёртывание на чистых, контролируемых заводах полностью практично. Обычный цех SMT — это, как правило, ESD-защищённая зона, а не классифицированная чистая комната: безвентиляторный бокс Jetson или отечественный NPU-бокс ставится у линии, а GPU-серверы — в серверную. По собственным экспортным классификациям NVIDIA, Jetson Thor и DGX Spark можно продавать обычному материковому заводу без лицензии США; DGX Station — нельзя. По умолчанию используется 8-bit; вендорский quantization-aware 4-bit применим с условиями; 3-bit и ниже исключены.

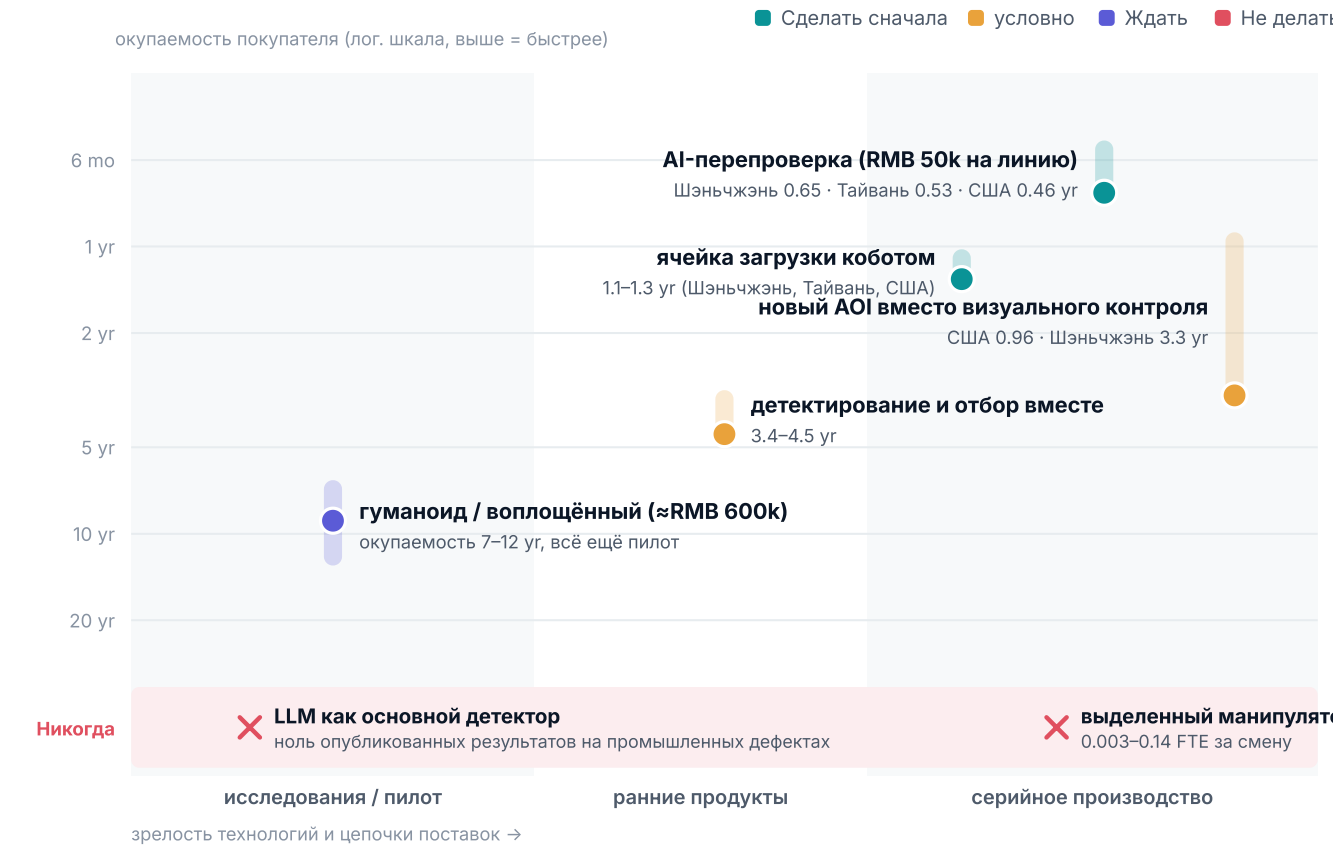
Рынок
КОНКУРЕНЦИЯ

Рынок реален, но ограничен. В материковом Китае около 35 000–46 000 линий SMT и около 26 000–66 000 встроенных AOI без ИИ; обслуживаемый рынок сторонней модернизации составляет около 100–700 млн RMB разово, в окне около пяти лет, а производители AOI начали встраивать перепроверку в новые машины estimate. Конкуренция уже сильна: клин перепроверки окружён производителями AOI, внутренними командами EMS и платформенными вендорами, а маржа — в продуктном железе.

Вопрос	Суждение	Ключевые цифры (источники и уверенность — в главах)
① **Что сегодня запускают основные заводы электроники?**	Массовая линия SMT — это «3D SPI + AOI до и после оплавления + выборочный рентген + ICT/FCT + ремонт и визуальный контроль», с ручной перепроверкой за каждым AOI. ИИ входит в перепроверку тремя путями: встроенный производителями AOI, построенный внутри ведущими EMS или добавленный третьими сторонами. Сборка и перемещение используют манипуляторы, SCARA и AMR с обученным восприятием и программируемым движением; воплощённый ИИ всё ещё в пилотировании	Такт ≈13–22 с на плату; один ODM заложил 30.6 млн RMB на AOI для 18 линий, ≈3–5 AOI на линию; ≈97–99% срабатываний AOI — ложные
② **Насколько далеко заходит технология, можно ли за деньги купить «ноль людей» и что может использовать каждый регион?**	Инспекция индустриального уровня — это «специализированный детектор + AI-перепроверка + несколько человек»; даже неограниченные деньги не достигают нуля людей. Материковый Китай использует только отечественные модели или локальные открытые веса; частные заводы на Тайване могут использовать почти всё, но фабрики и критическая инфраструктура на практике исключают материковые бренды; в Юго-Восточной Азии меньше всего ограничений; США обладают полным стеком плюс пошлины и сертификация	AI-перепроверка сокращает ложные срабатывания или часы перепроверки на 60–90% vendor claims ; камера 4K на целую панель ≈78 µm на пиксель против 8–20 µm у AOI; манипуляторы материкового производства платят ≈40% совокупной пошлины США, приборы типа AOI — ≈37.5%
③ **Можно ли запускать локально в чистых комнатах и изолированных заводах, на каком оборудовании, достаточно быстро и точно?**	Да; ограничения процедурные, а не вычислительные. Безвентиляторный бокс ≤100 Вт у линии, GPU-серверы в серверной; Thor для станций, Mac Studio или отечественный моноблок для заводских Q&A никакого DGX Station в материковом Китае. Скорость в порядке, когда размер модели соответствует задаче; используйте 8-bit, вендорский quantization-aware 4-bit с условиями, никогда 3-bit	≈0.5 с на кроп для модели зрения 8B на Thor estimate ; набор разработчика Thor 59 104 RMB в Китае; DGX Station $123–140 тыс., 139 токенов/с суммарно для 32 пользователей на модели 743B

Вопрос	Суждение	Ключевые цифры (источники и уверенность — в главах)
④ **Может ли новейший флагман выступать учителем на протяжении десятков раундов и быть дистиллирован в локальных учеников, в том числе для гибкого производства и манипуляторов?**	Инспекция: предразметка и сортировка; автоматически принимать метки учителя только для классов с «≥299 дефектов, ноль пропусков» на золотом наборе. Роботы: учителя — это по-прежнему человеческие демонстрации, RL на реальном роботе и симуляция, а ученики приближаются к производственному уровню только на структурированных станциях	Ни один ноябрьский флагман не имеет балла на промышленных дефектах; ученики 97.8% на обученных разъёмах, 29–67% на незнакомых, 37–59% на движущихся конвейерах; фронтир-модели, напрямую управляющие суставами, 0–5.5%
⑤ **Сколько это стоит и каков ROI по регионам?**	AI-перепроверка окупается в пределах года в материковом Китае, на Тайване и в США; обслуживание коботом — 1.1–1.5 года; Юго-Восточная Азия — 2–6 лет; манипулятор для отбора на линии SMT не окупается никогда	Бокс перепроверки 50 тыс. RMB на линию: Шэньчжэнь 0.65 года, Тайвань 0.53, США 0.46, Вьетнам 2.2 estimate ; приемлемая цена сторонней надстройки ≈10–30 тыс. RMB на линию разово или 5–10 тыс. RMB в год
⑥ **Достаточно ли уже сильны конкуренты, и лучше начать сейчас или подождать?**	Спрос не насыщен, но предложение переполнено; OEM-ИИ ограничивает окно модернизации примерно пятью годами. Входите сейчас с узким продуктом через канальных партнёров, не сражаясь в лоб с платформенными вендорами; ждите производственных данных 2027–2028, прежде чем ставить на воплощённый ИИ	≈26–66 тыс. встроенных AOI без ИИ в материковом Китае, сокращаются на 8–12% в год estimate ; ведущий AI-вендор зрения имел выручку 2025 года 1.086 млрд RMB и чистый убыток 991 млн RMB; выручка 3C ближайшего листингованного аналога упала на 79.9% год к году в январе–мае 2026

FIG 01 **Где ИИ вписывается в заводской цех сегодня** ***AI-перепроверка ложных срабатываний AOI и обслуживание машин коботами зрелы и быстро окупаются; выделенный манипулятор для отбора и LLM как основной инспектор не стоят того.***



Источник: Окупаемость по модели автора (глава 4); ось зрелости — оценка автора.

Что строить и что не строить

	Рекомендация
Строить	AI-перепроверка и классификация дефектов на изображениях AOI для средних EMS со старыми или разнородными AOI без AI-апгрейда от производителя AOI; инспекция того, что SMT AOI покрывает слабо (сквозные отверстия и ручная пайка, конформные покрытия, кабели, разъёмы и FPC, финальный внешний вид и винты); ячейки загрузки и выгрузки коботами, где обрабатывается каждая деталь (тестовые приспособления, лотки, упаковка), с инспекцией как дополнением; золотой набор и теневой режим на реальной линии как основа каждого утверждения
Не строить	Универсальную платформу «ИИ плюс манипулятор»; мультимодальный LLM как основной инспектор; манипулятор на линии SMT только для отбора NG-плат (NG-отводчик или буфер справляются); гуманоидов и «роботов как услугу»; фронтир-модели внутри контура управления или функции безопасности; горы GPU-памяти на станциях для запуска «полноразмерных» моделей; прямые продажи фабрикам, тайваньской критической инфраструктуре или заводам оборонной электроники США
С кем работать	Производители AOI и дистрибьюторы (канальные интерфейсы и интерфейсы изображений; дистрибьюторы принесли 46.68% выручки Sinic-Tek в 2025); производители коботов (сертифицированные манипуляторы и функции безопасности); EMS-компании с 10–40 линиями (поэтапное расширение и депозиты); локальные интеграторы в Таиланде и других странах (для соответствия правилам локальной привязки BOI)

Линия сегодня

01 Что сегодня запускают заводы электроники и что технически осуществимо

Стандартная массовая линия: несколько AOI, выборочный рентген, центральная перепроверка, 13–22 секунды на плату

Первичные документы листингованных ODM показывают, как на самом деле выглядит массовая линия 2025–2026. Проспект Longcheer для Гонконга (данные до 30 сентября 2025) перечисляет проверки качества SMT как «рентгеновский контроль, контроль паяльной пасты (SPI), автоматический оптический контроль (AOI), функциональный тест PCBA и визуальный контроль»; его технологический процесс идёт так: AOI до оплавления, AOI после оплавления, выборочный рентген, NG-платы на станцию ремонта SMT, затем финальная визуальная проверка. Его заводы в Хуэйчжоу, Наньчане и Вьетнаме имеют 34, 24 и 8 линий SMT и 67, 60 и 13 линий финальной сборки (Longcheer H-share prospectus) single source, primary filing. Завод ноутбуков Huaqin в 2023 планировал 18 линий SMT с бюджетом на AOI 30.6 млн RMB, около 1.7 млн RMB на линию, что по отечественным ценам на 3D AOI составляет три–пять AOI на линию для инспекции до и после оплавления двусторонних плат (Huaqin A-share prospectus) primary filing + estimate. Поэтому массовая линия обычно имеет от двух до четырёх точек перепроверки, а не одну.

Машины уже сильно автоматизированы: Huaqin сообщает о 96% автоматизации на своих линиях SMT (Huaqin H-share prospectus) single source. Оставшиеся люди заняты перепроверкой, ремонтом, визуальным контролем, тестированием и финальной сборкой, и ни одна компания не публикует, сколько перепроверяющих у неё на линию (глава 7).

Три независимые оценки сходятся на такте **13–22 секунды на плату** (экологическая отчётность 2026 года для пяти линий в Хэбэе (Hebei SMT project EIA), линии ноутбуков Huaqin и кейс iFactory), или 7–11 секунд на сторону для двусторонних плат estimate. Сложность линии с высокой номенклатурой — не объём, а запуск новых продуктов и переналадка, и каждая переналадка — это ещё один холодный старт для модели ИИ.

Со стороны ПО собственные станции перепроверки производителей AOI уже централизованы: ПО ремонтной станции Sinic-Tek позволяет «одному хосту управлять оборудованием восьми линий SMT» (Sinic-Tek prospectus) single source, primary, 2022, а после того как Jabil в 2026 развернул удалённое управление AOI ViTrox V-ONE, один инженер мог удалённо управлять 5–8 AOI против 1–2 локально (AtlasPCB) vendor claim. Сторонний продукт AI-перепроверки подключается ровно к этой центральной станции через экспорт изображений и результатов AOI (IPC-CFX, IPC-HERMES-9852 или собственный формат вендора). Сами AOI быстро добавляют ИИ: Yamaha заявляет, что её ИИ-оценка занимает около секунды на вызов, помогает инспекторам после 30 NG-изображений и достигает «очень высокого» качества после 1 000 (DirectIndustry/Yamaha) vendor claim.

FIG 02 **Серийная SMT-линия: люди сидят за AOI** ***Труд сосредоточен на местах перепроверки после автоматической оптической инспекции. Именно туда входит ИИ.***



Источник: Публичная отчётность и материалы вендоров (Huaqin, DBG, Sinic-Tek и др.); сводка автора.

Этап	Основные продукты и технологии	Типичные поставщики	Насколько далеко зашёл ИИ
После нанесения пасты	3D SPI (структурированный свет, муар, местами лазерная триангуляция)	Koh Young, Sinic-Tek, TRI	OEM-оценка пасты ИИ в производстве; Jutze заявляет, что некоторые модели «полностью на ИИ»
После установки и оплавления	2D/3D AOI с разноракурсными разноцветными LED и структурированным светом, 8–20 µm на пиксель, камеры 8–12 МП	Koh Young, Omron, Saki, TRI, ViTrox, Mirtec, Jutze, Sinic-Tek и другие	OEM AI-перепроверка (Jutze в производстве, ViTrox V510Ai, Sinic-Tek с Tencent); ведущие EMS строят свои (DBG); сторонние надстройки
Нижние выводы и скрытые соединения	Выборочный или встроенный рентген (AXI)	—	Hikvision заявляет, что её модель Guanlan ошибается в пустотах рентгена менее чем в 1% случаев vendor claim
Перепроверка и ремонт	Центральная станция перепроверки (один хост, восемь линий), удалённая перепроверка, станция ремонта	ПО производителей AOI, ViTrox V-ONE	ИИ закрывает ложные срабатывания, остальное достаётся людям
Визуальный контроль, IQC, FQC	В основном ручной; некоторые AI-машины внешнего вида с роботизированной загрузкой	SmartMore и другие	Машины внешнего вида в производстве; нет публичных данных о численности для IQC/FQC
Перемещение и обслуживание станков	Шестиосевые манипуляторы и коботы, SCARA, AMR, наведение по 2D/3D-зрению	FANUC, ABB, Yaskawa, KUKA, UR, Dobot, JAKA, Elite; Mech-Mind, Hikrobot	Обученное восприятие плюс программируемое движение в производстве; VLA и гуманоиды всё ещё в пилотировании
Данные и системы	MES, IPC-CFX, IPC-HERMES-9852	—	Результаты записываются обратно в MES; поддержка в старом оборудовании неравномерна

Источники к таблице: указанные выше документы и (Koh Young, Jutze 2025 annual report, Sinic-Tek 2025 annual report, ViTrox, Sohu on Hengtianyi, Xinhua on Hikvision).

Люди сидят за AOI, потому что большинство срабатываний ложные

Цепочка инспекции на массовой линии SMT/PCBA такова: 3D SPI после нанесения пасты для пропущенной, недостаточной или избыточной пасты, перемычек и смещения; AOI после установки (опционально) и после оплавления для пропущенных, смещённых, повёрнутых, перевёрнутых или «надгробных» деталей и объёма припоя; рентген при наличии BGA или других деталей с нижними выводами; затем внутрисхемный тест или тест летающим щупом, функциональный тест и ручной визуальный контроль. За каждым AOI есть место перепроверки и станция ремонта (Sinic-Tek prospectus, JLCPCB) verified. AOI после оплавления — самый распространённый и даёт лучшую ценность, но видит только форму и площадь; соединения, пустоты и смачивание под BGA требуют рентгена или электрического теста (JLCPCB).

Труд сосредоточен в перепроверке, потому что большинство срабатываний ложные. Один британский EMS зафиксировал **54 019 ложных срабатываний против 544 реальных дефектов на 917 номерах деталей, около 99% срабатываний** (arXiv 2505.03576) single source, preprint; на заводе Siemens в Раштатте **97.5% плат, отклонённых AOI, были исправными** (Siemens case study) vendor claim. Загруженная линия видит около 448 ложных срабатываний за смену по 42–55 секунд каждое, 5.2–6.8 часа за смену, примерно один штатный перепроверяющий (iFactory) vendor claim; часы — оценка.

FIG 03 **Около 97–99% срабатываний AOI — ложные** ***Сегодня человек подтверждает каждый вызов; AI-перепроверка сначала снимает ложные и передаёт остальное людям.***



Источник: Кейс Siemens, iFactory (заявление вендора), годовой отчёт DBG 2025; часы перепроверки оценены автором.

Из формулы «часы перепроверки за смену = платы × ложные срабатывания на плату × секунды на вызов» один перепроверяющий может покрыть около 0.7–1 массовой линии с правиловым AOI. Как только ИИ убирает 60–90% ложных срабатываний, одного времени на оценку хватило бы на 2–14 линий на человека, но хождение, обработка NG-буфера и передача в ремонт ограничивают локальное покрытие двумя–четырьмя соседними линиями; только центральная или удалённая перепроверка приближается к теоретическому потолку estimate; секунды на вызов — из одного источника. Поскольку массовые AOI сканируют 60–80 см² в секунду (Yunheng 2026 AOI guide), сторона 20×15 см занимает 4–5 секунд; ИИ должен обработать лишь несколько отмеченных областей в пределах одного такта, а не пересканировать плату.

Что уже работает в производстве: контролируемая классификация на собственных изображениях AOI

Проверенная схема — контролируемая классификация глубоким обучением на изображениях AOI, который уже есть у завода. Надстройка Fujitsu на AOI ADLINK на Тайване сократила ложные срабатывания **более чем на 90%** (Fujitsu case); OKI с июля 2026 использует ИИ, чтобы сократить ручное подтверждение после AOI примерно на **80%** в собственном EMS-бизнесе (Circuits Assembly); Siemens сообщает о 60% меньше ручного

анализа, €160 287 сэкономленных за первый год и окупаемости за восемь месяцев (Siemens case study) все vendor claims. У этих кейсов общее: почти никто не публикует уровень пропусков, и все оставляют людей на оставшихся срабатываниях. Единственный кейс с цифрой пропусков (0.8% до 0.3% против договорных 0.5%) внутренне противоречив (iFactory). ИИ сокращает часы перепроверки, а не саму работу перепроверки.

AI-перепроверка перешла от сторонних надстроек к OEM-функциям и внутренним системам, что меняет то, с чем сравнивают новичка. Годовой отчёт Jutze за 2025 перечисляет «апгрейд 3D AOI» как серийное производство и говорит, что он «использует алгоритмы ИИ, чтобы убрать ручную перепроверку на стороне инспекции» (Jutze 2025 annual report); Sinic-Tek с 2019 развивает «мультимодальную систему ручной перепроверки с поддержкой ИИ» и совместно с Tencent разрабатывает двусторонний 3D AOI, нацеленный на покрытие дефектов 99.5%+ и точность распознавания 98% (Sinic-Tek 2025 annual report) vendor claims в первичных документах. Крупные EMS строят свои системы. Годовой отчёт DBG за 2025 перечисляет «ИИ-оценку для визуального контроля» как завершённый проект, который «покрывает более 95% компонентов, отмеченных AOI», с «уровнем пропусков ниже 20 PPM» и «более чем 85% меньше ручной перепроверки», полностью отслеживаемый (DBG 2025 annual report) single source, primary. Это единственная найденная мной цифра по заводу электроники, которая даёт покрытие, пропуски и сокращение вместе. **Сторонний продукт, продаваемый EMS первого эшелона, должен соответствовать 95% покрытия DBG, 20 PPM пропусков и 85% сокращения**; ниже первого эшелона бенчмарк — это 50–85% сокращения, которые рекламируют вендоры.

Детектирование аномалий, обученное только на исправных деталях, быстрое (EfficientAD-S занимает 2.2 мс на изображение (EfficientAD, WACV 2024)), но слабо на «логических» аномалиях, таких как отсутствующие или неверные детали, а лучшие методы в среднем ниже 60% AU-PRO на более сложном MVTec AD 2 (MVTec AD 2) single source. Отсутствующие и неверные детали лучше проверять по CAD и BOM.

Потолок при любой цене: слои машин, очень мало людей, но не ноль

Я не нашёл ни одного завода электроники, который полностью убрал ручную перепроверку и достиг нулевых пропусков. Практика для продукции IPC-A-610 Class 3 (высокой надёжности) по-прежнему предполагает, что обученные инспекторы решают пограничные случаи, которые автоматические пороги не могут классифицировать, а один сборщик PCB называет «автоматизированный контроль может заменить подпись человека» «распространённым заблуждением» (PCBCart) single source. Человеческий базовый уровень тоже нужно скорректировать: со ссылкой на исследование Sandia 2012 года, один визуальный инспектор в лучшем случае ловит около **80%** дефектов, два последовательно — около 96% (Instrumental) single source, cited by a vendor. Защитимое утверждение — «ИИ плюс малый контур ручной перепроверки лучше полностью ручного контроля», а не «ИИ заменяет контроль».

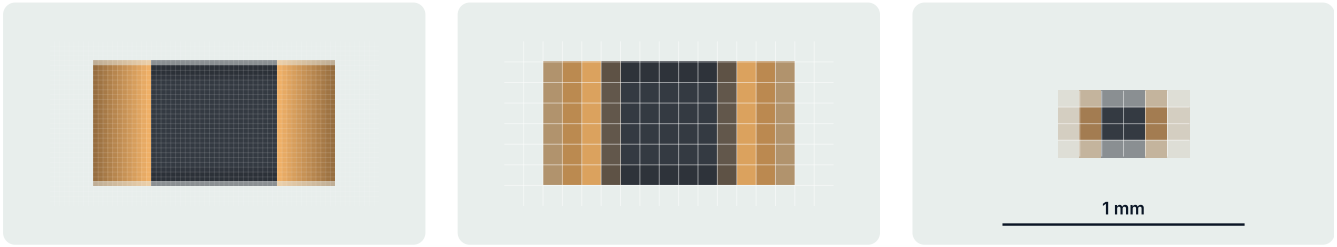
Без ограничения бюджета потолок — это 3D SPI, 3D AOI, рентген, AI-перепроверка и AI-станции внешнего вида, сложенные вместе, с людьми только для пограничных случаев, новых продуктов и переналадок, новых типов дефектов и подписи заказчика. Флагманские заводы сообщают о 85–92% меньше ручного контроля (I-Connect007, 2021, WiAdvance) vendor claims; моя оценка потолка без бюджета — труд перепроверки на уровне 8–15% от сегодняшнего estimate. Финальная сборка по-прежнему трудоёмкая: автоматизация кнопок iPhone 16 была отменена из-за высокого уровня дефектов (9to5Mac) single source.

Универсальная камера высокого разрешения — это не AOI

Производственные AOI снимают плитка за плиткой при 8–20 µm на пиксель, с камерами 8–12 МП, программируемой разноракурсной разноцветной подсветкой и структурированным светом 3D (Koh Young, Saki) verified. В отрасли давно отметили, что деталь 01005 покрывает лишь около 8×16 пикселей при 25 µm на пиксель, что «возможно, недостаточно» для надёжной оценки (Circuits Assembly, 2007). По той же арифметике estimate, камера 4K на панель 300 mm даёт около **78 µm на пиксель**: имперская деталь 0402 — это около 13×6 пикселей, достаточно для наличия и грубого смещения, но не для формы соединения, а 01005 — около 5×3 пикселей, фактически невидима. Достижение 15 µm на пиксель по всей панели потребовало бы около

330 мегапикселей, поэтому AOI перемещают камеру над платой. Когда кто-то говорит «0402», проверьте, что он имеет в виду: метрический 0402 — это имперский 01005, далеко за пределами любой широкоугольной камеры.

FIG 04 **Универсальная камера — это не AOI** ***Камера 4K на панель 300 mm даёт около 78 µm на пиксель: деталь 0402 сжимается до ≈13×6 пикселей, а 01005 исчезает.***



AOI: 8–20 µm на пиксель
0402 ≈ 100×50 px

камера 4K на панель 300 mm: ≈78 µm/px
0402 сжимается до ≈13×6 px: хватает лишь на наличие и грубое смещение

та же камера на 01005
≈5×3 px: фактически невидимо

Источник: Circuits Assembly; расчёт автора для панели 300 mm и сенсора 4K. Сетки пикселей иллюстративны.

На линиях, где уже есть AOI, лучшее место для ИИ — поверх изображений AOI. Только линии без AOI или станции в конце линии оправдывают закрытую камерную станцию с контролируемым освещением; открытая камера высокого разрешения над линией наиболее подвержена окружающему свету и теням (CSDN).

Сборка и перемещение: обученное восприятие плюс программируемое движение уже в производстве

Отчёт IFR за 2026 говорит, что «машинное зрение для инспекции и наведения роботов зрело и индустриального уровня», тогда как «универсальный физический ИИ и обучение из симуляции и опыта ещё не индустриального уровня» (IFR World Robotics 2026 executive summary) single source, primary . Электроника — крупнейший покупатель роботов (142 235 установленных единиц в мире в 2025, 59% всех установок в Китае (IFR, IFR China release)), однако в Китае лишь 166 роботов на 10 000 производственных работников против 307 в США (IFR, April 2026). Схема в производстве — обученное восприятие плюс детерминированное движение; сквозное обученное движение близко к производственному уровню только на структурированных станциях (главы 3 и 6).

Что может использовать каждый регион

Материковый Китай полагается на отечественные модели и локальные веса; на Тайване решает идентичность заказчика; в США — пошлины и сертификация.

Элемент	Материковый Китай	Тайвань (частные заводы)	Юго-Восточная Азия	США
Claude (любой канал)	● Недоступен	● Условно: поставщик не должен быть в мажоритарной собственности или под контролем лиц или организаций из материкового Китая, Гонконга или Макао	● Аналогично	● Аналогично

Элемент	Материковый Китай	Тайвань (частные заводы)	Юго-Восточная Азия	США
GPT и Gemini в облаке	🔴 Недоступны	🟢 Доступны	🟢 Доступны (списки не проверялись по странам)	🟢 Доступны
Китайские облачные модели (Qwen, DeepSeek, Doubao, GLM, Kimi)	🟢 Доступны (зарегистрированные модели; предоставление услуги требует регистрации)	🟠 Легальны, но материковый хостинг редко проходит проверку безопасности заказчика; запрещены в госсекторе, госпредприятиях и критической инфраструктуре	🟠 Легальны (международные узлы)	🟠 Легальны, строгая проверка
Локальные открытые веса китайского происхождения (Qwen, GLM, DeepSeek)	🟢 Первый выбор; частное внутризаводское развёртывание не требует регистрации модели или алгоритма (чтение юрфирм)	🟢 Доступны (госсектор ограничен)	🟢 Доступны	🟠 Условно: проверка происхождения
Локальные открытые веса некитайского происхождения (например Gemma 4)	🟢 Доступны	🟢 Доступны	🟢 Доступны	🟢 Доступны
Классические негенеративные модели зрения	🟢 Доступны (вне правил генеративного ИИ)	🟢 Доступны	🟢 Доступны	🟢 Доступны
Jetson Thor/T5000/T4000, DGX Spark	🟢 Доступны: NVIDIA классифицирует их как 5A992.c, лицензия США не нужна для обычного гражданского завода; у Spark есть китайский SKU; канальные цены примерно на 40–60% выше прайса США	🟢 Доступны	🟢 Доступны	🟢 Доступны
RTX 5090, RTX PRO 6000, DGX Station	🔴 Недоступны: первые два — 4A090.a, DGX Station — 5A992.z; легальные альтернативы — 4090D, 5090 D V2, RTX PRO 5000 72GB, RTX 6000D, L20 (4A994.l)	🟢 Доступны	🟢 Доступны, но покупатели со штаб-квартирой или конечной собственностью из Китая считаются китайскими; Малайзия требует разрешения на экспорт и перевалку	🟢 Доступны
Mac Studio M5 Ultra 512 GB	🟢 Доступен (магазин Apple China)	🟢 Доступен	🟠 Предположительно доступен	🟢 Доступен
Huawei Ascend и другие китайские ускорители	🟢 Доступны	🔴 Высокий риск	🔴 Высокий риск	🔴 На практике неприменимы

Элемент	Материковый Китай	Тайвань (частные заводы)	Юго-Восточная Азия	США
Стационарные манипуляторы, AOI, промышленные камеры материкового производства	● Доступны	● Импортируемы; с 1 декабря 2025 власти могут действовать напрямую по материковым брендам, а ведомства, госпредприятия, некоторые фонды и поставщики критической инфраструктуры могут быть ограничены или лишены права их использовать	● Доступны	● Легальны, но совокупная пошлина около 40% на манипуляторы, 37.5% на приборы и камеры типа AOI, 25% на периферийные компьютеры
Зарубежные AMR и мобильные манипуляторы	● Доступны	● Доступны	● Доступны	● Новые модели требуют условного одобрения Министерства обороны США
Камеры Hikvision и Dahua	● Доступны	● Частное использование разрешено, госсектору запрещено	● Доступны	● Новые разрешения FCC заблокированы; запрещены для федеральных подрядчиков
Безопасность и сертификация роботов	● GB 11291.1-2011 и GB 11291.2-2013 обязательны, пересматриваются с апреля 2026; сертификация CR добровольная; GB/T 39404-2026 охватывает кибербезопасность контроллеров	● Ограждение по умолчанию; совместная работа без ограждения должна соответствовать CNS 14490 или ISO 10218 с письменной оценкой	● По заказчику и месту экспорта, подтверждается письменно до отгрузки	● Электрооборудование должно быть «одобрено» (на практике — знак NRTL или полевая оценка); UL 1740 требует, чтобы робоячейки соответствовали ISO 10218-2 с 11 августа 2025
Фабрики и заводы контролируемой продукции	—	● Контракты на закупку TSMC требуют SEMI E187; материковое оборудование выводится из фабрик 2 nm	—	● Технические данные ITAR включают «фотографии»; политика отказа для Китая; материковые поставщики на практике не могут обслуживать заводы оборонной электроники

Источники к таблице: (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).

Материковый Китай: полный стек, но ни закрытые фронтир-модели, ни самые мощные одиночные машины не могут войти. Зарегистрированные отечественные облачные модели, локальные открытые веса (Qwen под Apache-2.0, GLM и DeepSeek под MIT), классические модели зрения и отечественные AOI, камеры и коботы — всё применимо. Собственные экспортные классификации NVIDIA относят набор Jetson AGX Thor, модули T5000/T4000 и DGX Spark к 5A992.c, поэтому обычному заводу EMS или PCBA не нужна лицензия США, тогда как RTX 5090 и RTX PRO 6000 (4A090.a) и DGX Station (5A992.z, TPP 58 484) нельзя купить в материковом Китае (NVIDIA export classifications, eCFR) primary; юридическое прочтение — моё и не является юрконсультацией. RS China указывает набор Thor по 59 104 RMB с налогом, примерно в 1.4 раза выше прайса США (RS China).

Четыре юрфирмы согласны, что частное внутризаводское развёртывание для собственного использования завода не требует регистрации модели или алгоритма, тогда как хостинговая многотенантная услуга нахо-

дится в серой зоне (Covington, FPF, JunHe) verified: согласованные чтения, официального решения нет. GB 11291.1-2011 и GB 11291.2-2013 остаются обязательной базой безопасности роботов и пересматриваются с 28 апреля 2026; GB/T 39404-2026 о кибербезопасности контроллеров вступил в силу 1 ноября 2026 (revision plan 20262542, GB/T 39404-2026) single source, primary. Пока не выйдет новый GB 11291 (2028–2029 estimate), проектируйте под ISO 10218:2025 и храните файлы соответствия GB 11291.

Тайвань: частные заводы могут использовать почти всё; решает идентичность заказчика. Фронтир-API, открытые веса и материковое оборудование — всё легально (data.gov.tw). Но с момента вступления в силу изменённого Закона об управлении кибербезопасностью **1 декабря 2025** власти могут действовать напрямую по продуктам, которые являются «материковыми брендами», и ограничивать их в госорганах, госпредприятиях и у поставщиков критической инфраструктуры (Laws & Regulations Database, review rules) verified; список поставщиков критической инфраструктуры конфиденциален; а TSMC вписала SEMI E187 в свои контракты на закупку (CNA) и, по сообщениям, выводит материковое оборудование из своих фабрик 2 nm (TrendForce) single source. Частные EMS, сборочные и малые PCBA-заводы открыты; фабрики, парковые коммунальные службы и связанные с правительством заказчики на практике закрыты. Материковому поставщику следует проектировать под E187, поставлять программную спецификацию материалов (SBOM), предлагать офлайн-обновления и никогда не подключаться удалённо из материкового Китая.

Юго-Восточная Азия: меньше всего ограничений; проблемы — медленная окупаемость и китайские материнские компании. Всё применимо, но руководство США от 31 мая 2026 требует лицензии на экспорт передовых вычислительных позиций организациям со штаб-квартирой в Китае или конечной собственностью из Китая, даже если получатель в третьей стране (Greenberg Traurig) single source, а у Малайзии свои разрешения на передовые чипы (Fulcrum). Спрос перемещается сюда: TrendForce ожидал, что Таиланд, Вьетнам и Малайзия приблизятся к 50% SMT-мощностей серверных ODM в 2026 (TrendForce) single source, прогноз 2023.

США: полный стек, ценой пошлин, сертификации и контролируемых данных. По главе 99 тарифного расписания США манипулятор материкового производства платит около 40% совокупной пошлины, приборы и камеры типа AOI — около 37.5%, периферийные компьютеры — 25%; произведённые в Малайзии или Индонезии — около 10–12.5%, во Вьетнаме, Таиланде или на Филиппинах — 12.5–15%, на Тайване — 10%, компьютеры — 0% вне материкового Китая (USITC HTS chapter 99, Great Lakes Customs Law) primary schedule + estimate. Уведомление USTR от 10 ноября продлило 178 старых исключений раздела 301 до 9 января 2027, ни одно из них не покрывает роботов, приборы инспекции, камеры или компьютеры (Federal Register 2026-22925) verified.

OSHA требует, чтобы рабочее электрооборудование было «одобрено» (на практике — знак NRTL или полевая оценка около $3 000–$20 000+ на ячейку), а с 11 августа 2025 UL 1740 подводит целые робоячейки под ISO 10218-2 (eCFR 1910.399, Intertek, Industrial Monitor Direct). Оборонная электроника на практике закрыта для материковых поставщиков: ITAR рассматривает производственные фотографии как технические данные и применяет к Китаю политику отказа (eCFR 22 CFR 120.33, eCFR 22 CFR 126.1). Ячейка для США должна использовать стационарный манипулятор, неограниченные камеры и локальные или US-облачные модели, с финальной сборкой в Юго-Восточной Азии, где это действительно меняет происхождение.

При любом бюджете потолок различается по регионам, но нигде не достигает нуля людей estimate. Материковый Китай может сложить отечественные 3D SPI, 3D AOI, рентген, AI-перепроверку и внутреннюю AI-команду, чтобы снизить труд перепроверки до 8–15%, но без закрытых фронтир-моделей или DGX Station общезаводские LLM Q&A должны работать на локальных открытых весах и отечественных вычислениях (полноразмерная модель для многих пользователей требует как минимум двух восьмикарточных серверов Ascend). Тайвань и Юго-Восточная Азия технически могут использовать лучшее из всего, включая фронтир-API и DGX Station, ограниченные идентичностью заказчика, материковыми брендами и лицензированием китайской материнской компании. У США нет технических ограничений, ценой пошлин, сертификации NRTL и разделения персонала для контролируемых данных. Везде люди остаются для пограничных случаев, новых продуктов и переналадок, новых типов дефектов и подписи заказчика.

Продукт, продаваемый в разных регионах, поэтому нуждается в двух стеках моделей: материковый стек на зарегистрированных отечественных моделях или локальных открытых весах и зарубежный стек на фронтир-API плюс локальная модель некитайского происхождения. Политика Anthropic, действующая с 12 ноября 2026, оценивает по собственности и контролю, поэтому материковый поставщик не может строить на Claude даже через зарубежную дочку, облачную платформу или реселлера (Anthropic) single source, primary.

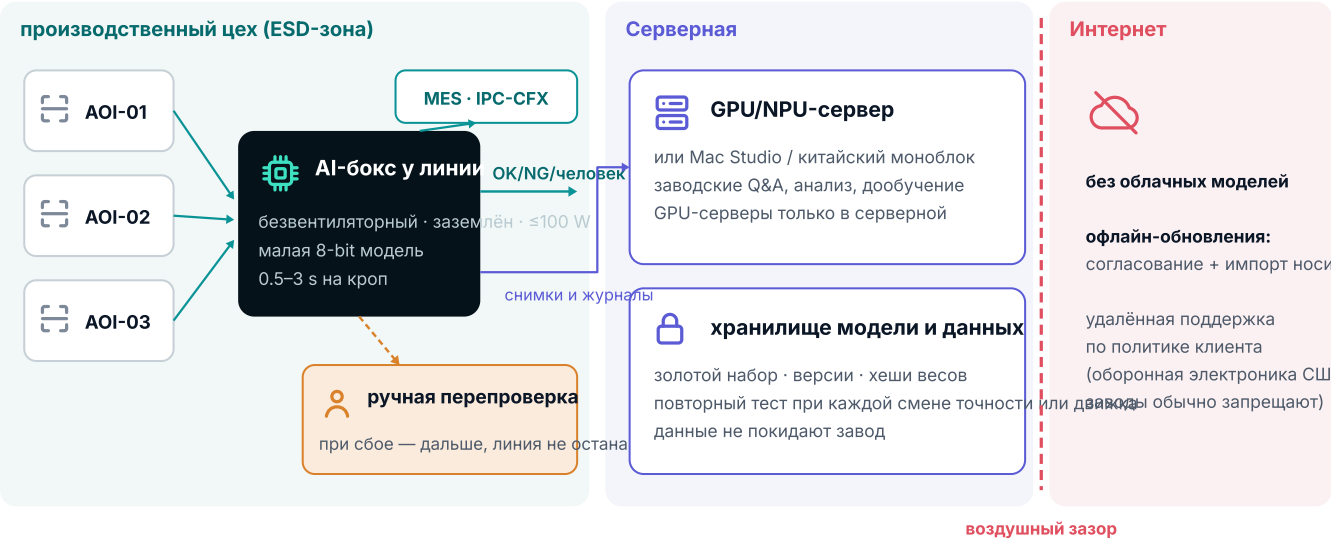
Локальное развёртывание

02 Запуск ИИ внутри завода: чистые комнаты, воздушные зазоры, оборудование и квантование

В изолированном заводе ограничение — процесс, а не вычисления

Полностью офлайн визуальная инспекция — рутина; ограничения процедурные. Руководство MIIT Китая 2024 года по безопасности промышленных систем управления требует зонирования, межсетевых экранов или data-диодов, аудируемого удалённого обслуживания с ограничением по времени, сканируемых носителей и проверенных патчей (guide text) single source ; полупроводниковые заказчики добавляют SEMI E187 и E188, которые TSMC вписала в свои контракты на закупку в 2023 (SEMI, TXOne) verified . Результаты возвращаются в MES через IPC-CFX, а идентичность платы передаётся через IPC-HERMES-9852 (Hermes standard). Распространённая ловушка — лицензирование, предполагающее сеть: офлайн-режим Roboflow кэширует веса на 30-дневной аренде (Roboflow). Офлайн-продукту нужны офлайн-активация, подписанные обновления на сканируемых носителях, программная спецификация материалов (SBOM), контроль USB, локальные журналы аудита и адаптер CFX.

FIG 05 **Изолированное развёртывание** ***Небольшой бокс у линии решает; серверная отвечает на вопросы; при сбоях — люди, а данные никогда не покидают завод.***



Источник: Сводка автора (глава 2).

Закрытые фронтир-модели на практике не входят в серверную завода. Gemini предлагается изолированно только через Google Distributed Cloud для правительств и крупных предприятий (Google Cloud); Microsoft Foundry Local открывает большие модели только «квалифицированным заказчикам» и не называет ни одной модели GPT (Microsoft); Claude предлагается только в облачных правительственных регионах, без локального развёртывания (Claude for government); ни одна из трёх не предлагается в материковом Китае. Оборудование завода должно проектироваться вокруг открытых весов; фронтир-модели могут в лучшем случае размечать или объяснять изображения в штаб-квартире, офлайн, когда заказчик разрешает выносить анонимизированные изображения.

Оборудование серверной следует считать офисным: Mac Studio рассчитан на 10–35 °C, а DGX Spark — на 5–30 °C, есть сообщения о тепловых отключениях при длительном инференсе (Apple, NVIDIA forum) verified. На линию ставятся безвентиляторные промышленные боксы, такие как промышленный ПК Jetson T4000, рассчитанный на −25–60 °C (AAEON) vendor claim. Поскольку час простоя стоит крупному заводу $36 000–$2.3 млн (Siemens True Cost of Downtime 2024) vendor survey, каждый шаг ИИ на линии должен **отказывать открыто** (fail open): при любой неисправности он откатывается к существующему процессу «AOI плюс люди».

Зоны чистоты и ESD: обычный цех SMT — это ESD-защищённая зона, а не классифицированная чистая комната

То, что ограничивает место размещения вычислений, — это статика и воздушные потоки, а не класс чистоты. Большинство зон сборки SMT «формально не контролируются по частицам», а заводы мирового уровня «могут достигать или приближаться к ISO 8» (Circuitnet expert panel) single source. Медицинские PCBA и модули камер обычно находятся в ISO 7, голый кристалл и оптическая юстировка — в локальных ISO 5–6 (PCBSync) single source; строители чистых комнат указывают ISO 7–8 для SMT в целом (Angstrom Technology) vendor claim.

Универсальное требование — контроль ESD: ANSI/ESD S20.20-2021 требует изолированных проводников ниже 35 В и держит изоляторы с сильным полем на расстоянии не менее 30 см от чувствительных устройств (ANSI, ESDA, Desco) verified. Фабрики размещают вычисления в субфабе или ЦОД, в стойках с жидкостным охлаждением, соответствующих SEMI S2/S8/F47 (Prodrive) vendor claim.

То, может ли оборудование войти в чистую зону, проверяется по эмиссии частиц по ISO 14644-14; роботы имеют такие сертификаты (Fraunhofer IPA), но я не нашёл ни одного промышленного ПК, периферийного AI-бокса или GPU-сервера с ними. CE, FCC и IEC 60068 показывают, что бокс выдерживает свою среду, а не то, что он может войти в ESD-защищённую зону; это проверяет ESD-координатор заказчика. GPU-серверы никогда не ставятся на линию: DGX Station потреблял 1 292 Вт на GPU в тестах (StorageReview).

FIG 06 **Зоны чистоты и ESD определяют, где могут стоять вычисления** ***Обычный SMT-цех — это ESD-зона, а не чистая комната. GPU- и NPU-серверы — в серверной.***



Источник: ANSI/ESD S20.20, IEC 61340-5-1, ISO 14644, стандарты SEMI и материалы вендоров; сводка автора.

Зона	Типичная среда	Вычисления у линии	GPU/NPU-серверы
Обычный цех SMT/PCBA (потребительский, промышленный)	ESD-защищённая зона по S20.20 или IEC 61340-5-1, обычно не классифицированная чистая комната (≈ISO 8); влажность 40–60%	Безвентиляторный бокс в заземлённом металлическом корпусе (Jetson T4000/Thor, китайский NPU-бокс) рядом с AOI или местом перепроверки; пластиковые мониторы и корпуса вне зоны или ≥30 см от плат	Серверная или кондиционируемая ИТ-стойка
Медицина, модули камер, электроника электромобилей (чистая комната ISO 7–8)	Чистая комната плюс защита от ESD	Безвентиляторный, герметичный IP65/66, гладкий и легко протираемый; запросите данные ISO 14644-14 или сертификат IPA	Вне чистой комнаты, с сетью внутрь
Голый кристалл, разварка, оптические станции (локальные ISO 5–6)	Мини-среда	Внутри только камеры и свет; вычислительный бокс снаружи	Серверная
Чистая комната фабрики (ISO 5 и чище)	Фаб-специфичные планы ESD и электростатического притяжения, оборудование по SEMI E78; кибербезопасность по SEMI E187/E188	Только интегрированные в инструмент, сертифицированные фабом контроллеры; никаких дополнительных ПК с вентиляторами	Стойки с жидкостным охлаждением в субфабе или ЦОД фабрики

Таблица — моя сводка указанных выше источников estimate. Для продукта это означает безвентиляторный бокс у линии мощностью менее 100 Вт, металлический корпус, соединённый с защитным заземлением, вход 12–24 В постоянного тока, работу в широком диапазоне температур при реальном потоке воздуха в шкафу и готовый отчёт об эмиссии частиц по ISO 14644-14 для заказчиков ISO 7 и чище. Любая схема, где «один GPU-сервер обслуживает четыре линии», ставит сервер в серверную и оставляет у линии лишь малый бокс или камеру.

Ограничения по скорости по применению: точность сложнее скорости

Заводы и вендоры почти никогда не публикуют целевые показатели задержки или доступности ИИ. Пороги ниже выведены из такта, времени ручной перепроверки, частоты управления роботами и исследований взаимодействия estimate:

Применение	Ограничение по скорости	Ограничение по точности	Доступность
Встроенная инспекция каждой детали	В пределах ½–1 такта: 240–1 000 кропов при такте 10–30 с, 8–100 кропов в секунду	Пропуски не хуже сегодняшних «AOI плюс люди» (пример контракта <0.5%)	Отказ открыто, никогда не останавливать линию; цель ≥99.9% (нет публичного SLA)
Встроенное второе мнение только по отмеченным AOI кропам	Отметки (обычно ≤5–20 на плату) в пределах ½ такта: 0.5–3 с на кроп	То же; обучено на собственных изображениях линии	Откат к ручной перепроверке
AI-перепроверка ложных срабатываний	2–5 с на вызов, более чем в десять раз быстрее человеческих 42–55 с	Пропуски не хуже человеческого перепроверяющего	Деградация добавляет людей, никогда не останавливает линию
Управление роботом (VLA)	100–300 мс на чанк действий с разбиением в реальном времени, вывод 10–50 Гц	≥99% успеха задачи, чтобы превзойти программируемое движение	Система безопасности независима от модели

Применение	Ограничение по скорости	Ограничение по точности	Доступность
Планирование робота и выбор навыков	1–3 с при необходимости в каждом цикле; ≤1 мин, если только при исключениях или переналадках	Вывод должен быть проверяемым кодом или ID навыка	Откат к фиксированной программе
Q&A (SOP, запросы MES)	Первый токен в идеале ≤1–2 с, максимум 10 с; ≥15–20 токенов/с на пользователя; 5–20 одновременных пользователей	Точные цитаты; верность на длинном контексте	Уровень офисного ИТ (≈99–99.5%)
Анализ и отчёты о первопричинах	Минуты допустимы (пакетно)	Наивысшая; верность на длинном контексте	Уровень офисного ИТ

Physical Intelligence измерила 97 мс инференса для VLA класса 3B (около 108–139 мс с сетью и предобработкой), установила, что разбиение в реальном времени сохраняет пропускную способность неизменной вплоть до 200 мс дополнительной задержки, и выполняла тонкие задачи, такие как зажигание спички или подключение сетевого кабеля, при более чем 300 мс (Physical Intelligence) vendor research ; исследования взаимодействия ставят предел удержания хода мысли в одну секунду, а внимания — в десять (NN/g). Точность — сложная часть: открытая Qwen3-VL-8B в среднем набирает лишь 56.8 zero-shot на пяти наборах промышленных аномалий (IndusAgent) single source . Никакое оборудование не спасёт визуально-языковую модель, не настроенную на собственных изображениях линии.

FIG 07 **Пороги скорости по применению** ***Встроенная инспекция меряется миллисекундами, Q&A — секундами, анализ — минутами. Люди тратят 42–55 s на вызов.***



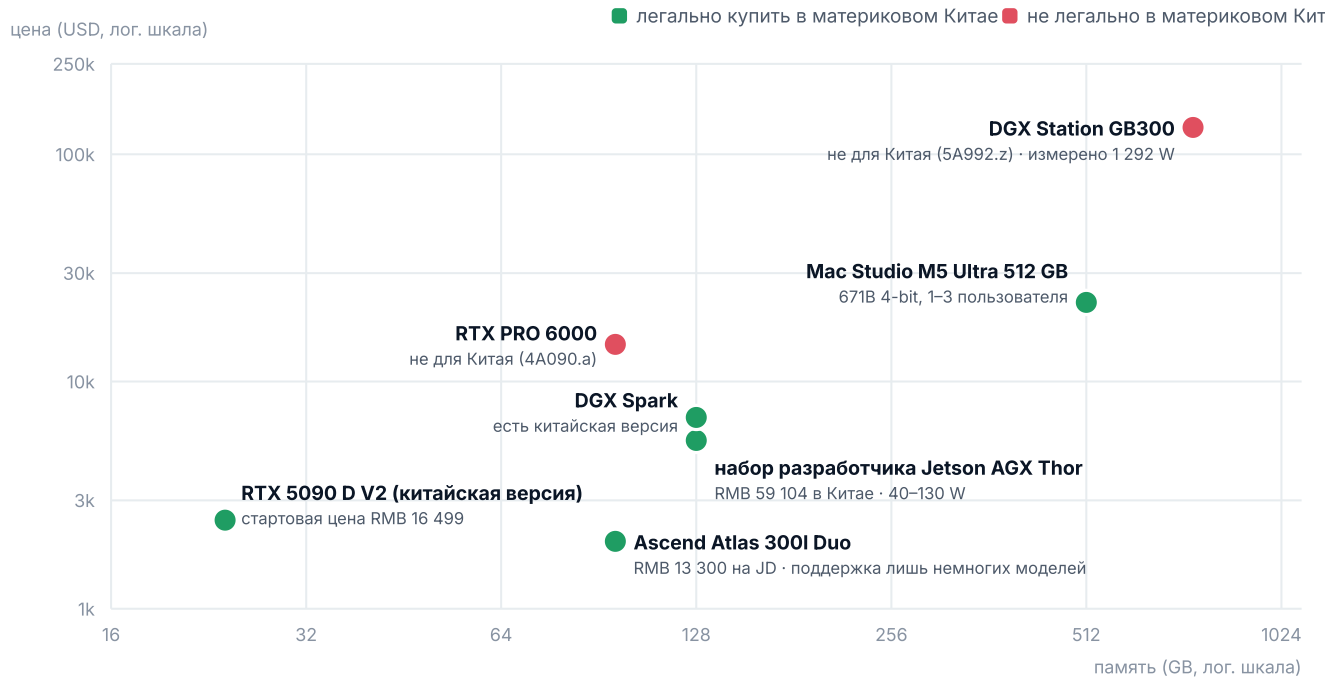
Источник: Сводка автора (главы 1–2).

Оборудование: Thor для станции, Mac Studio или отечественный моноблок для заводских Q&A, и никакого DGX Station в материковом Китае

Устройство	Память и пропускная способность	Цена (ноябрь 2026)	Для чего годится	Материковый Китай
Набор разработчика Jetson AGX Thor / T5000 / T4000	128 GB LPDDR5X, 273 GB/s; 40–130 Вт	$3 499 на старте → $5 499 (июль 2026); T5000 $4 999, T4000 $2 999; набор разработчика RS China 59 104 RMB с налогом	Модели зрения 2–8B, перепроверяющие кропы; VLA класса 3B в реальном времени	● Да (5A992.c, без лицензии для гражданских заводов)
DGX Spark	128 GB, 273 GB/s	$3 999 → $4 699 → $6 950; китайская модель ASUS GX10 на том же чипе 32 999 RMB на JD (окт 2025)	Многопользовательская MoE 8–35B	● Да (5A992.c, китайская модель, авторизованные каналы)
RTX PRO 6000	96 GB GDDR7	≈$8 565 → $13 250 (июнь 2026) → ≈$16 005 средний листинг eBay в ноябре	Многопользовательская 27–35B FP8; MoE класса 120B в 4-bit	● Нет (4A090.a)
Mac Studio M5 Ultra 512 GB	512 GB unified, 1.2 TB/s; 480 Вт макс. длительно	От $22 299 в США; ≈173–186 тыс. RMB оценочно в Китае	DeepSeek 671B в 4-bit для 1–3 пользователей	● Да
DGX Station GB300	748 GB (252 GB HBM3e при 7.1 TB/s + 496 GB LPDDR5X при 396 GB/s); 1 292 Вт измерено на GPU	$123–140 тыс. рыночная цена в ноябре (≈830–940 тыс. RMB до пошлины); $95–117 тыс. в августе	Высокий параллелизм на MoE 120–235B, помещающейся в HBM; лишь несколько пользователей на моделях 433–743B	● Нет (5A992.z, TPP 58 484)
Huawei Atlas 300I Duo	96 GB LPDDR4X, 408 GB/s	13 300 RMB у продавца на JD	Пакетная перепроверка малого объёма; немного поддерживаемых моделей	● Да (отечественный)
Huawei Atlas 800I A2 (8 × Ascend 910B 64 GB)	8 × 64 GB HBM	Не получено; полный DeepSeek требует двух при W8A8, четырёх при BF16	Общезаводские многопользовательские Q&A и анализ	● Да (отечественный)
RTX 5090 D V2 (китайская модель)	24 GB	16 499 RMB на старте (авг 2025), с тех пор выросла из-за цен на память	Модели зрения 8B, малые VLA	● Да

Источники: (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). Цены на 512 GB Mac Studio в материковом Китае экстраполированы из надбавок за конфигурацию в США и имеют низкую уверенность estimate.

FIG 08 **Локальные вычисления: память, цена и что легально купить в материковом Китае** ***Экспортные классификации определяют короткий список раньше бенчмарков.***



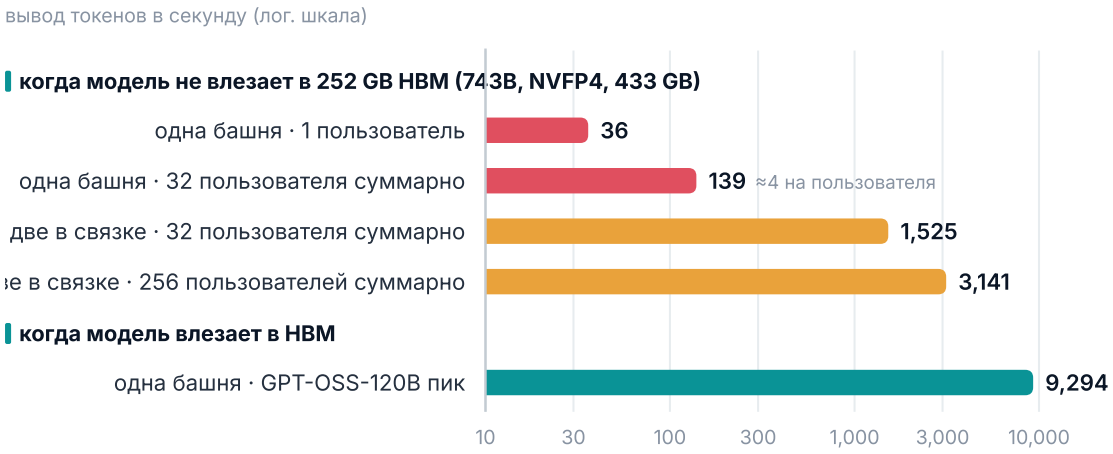
Источник: Экспортные классификации NVIDIA, Apple, RS China, JD.com и обзорные цены (ноябрь 2026).

Jetson Thor делает развёртывание на уровне станции рабочим. Тесты NVIDIA TensorRT Edge-LLM показывают, что Qwen3-VL-8B (NVFP4) префиллит 2 048 токенов за 104.5 мс и декодирует 44.4 токена/с на Thor (TensorRT Edge-LLM) vendor claim ; для кропа AOI и ответа в 20 токенов это около 0.5 с с моделью 8B и 0.2 с с 2B, или 8.5–21 кропа в секунду пакетно estimate . Задержку задаёт длина вывода, а не число параметров, поэтому ограниченное декодирование в классовый токен или короткий JSON — часть дизайна. GR00T N1.7 (3B) работает на 25.1 Гц на Thor (Jetson AI Lab) vendor claim . Orin NX занимает около 1.8 с на кроп, а у Atlas 300I Duo p95 выше 23 с, годится только для пакетной работы (CSDN) single source . В материковом Китае закладывайте NVIDIA-продукты на 40–60% выше прайса США estimate .

Mac Studio 512 GB делает полноразмерную модель на одной машине возможной для нескольких пользователей. M5 Ultra (август 2026) предлагает до 512 GB при 1.2 TB/s от примерно $22 299 (Apple, TechRadar); DeepSeek 671B в 4-bit (≈380–404 GB) декодирует 17–20 токенов/с на M3 Ultra (Pinggy) verified и оценочно 25–29 на M5 Ultra. Длинные промпты — слабое место: промпт поиска на 10 000 токенов занимает примерно 17–70 с estimate , за пределом 10-секундного ограничения взаимодействия.

DGX Station — самая мощная одиночная машина, но недоступна в материковом Китае и в пять–шесть раз дороже Mac. Поставляемый GB300 имеет **748 GB**, из которых лишь 252 GB — быстрая HBM (Wccftech, Tom's Hardware, Aug 2026) verified . Независимые тесты показывают около 9 300 выходных токенов/с на GPT-OSS-120B, которая помещается в HBM, но модель 743B, выходящая в LPDDR5X, дала 32 пользователям лишь 139 токенов/с суммарно, около 4 каждому; две связанные единицы достигли 1 525 (StorageReview, Aug 2026, StorageReview, Oct 2026) single source . Рыночные цены выросли до $123–140 тыс. к ноябрю (CyberPowerPC, monitors.com). В материковом Китае реалистичные варианты для той же задачи — Mac Studio, серверы CPU+GPU в стиле KTransformers (KTransformers) vendor test или отечественные серверы Ascend.

FIG 09 **Одна большая рабочая станция не создаёт общего ассистента** ***Когда модель не влезает в высокоскоростную память, 32 одновременных пользователя получают около 4 токенов в секунду каждый.***



Источник: StorageReview (авг и окт 2026); ServeTheHome. Независимые тесты из одного источника.

Дефицит памяти сдвинул те же позиции на 20–100% за год (DGX Spark вырос с $3 999 до $6 950 (TechRadar)), поэтому котировки оборудования следует индексировать, а не фиксировать.

Отечественные ускорители запускают полноразмерные модели; узкое место — сопровождение ПО. Ascend 910B имеет 64 GB HBM, но vLLM-Ascend помечает DeepSeek V3/R1 как «больше не поддерживаемые» и поддерживает лишь несколько моделей на меньших картах (vLLM-Ascend) single source, primary, а адаптация каждой новой модели занимает один–два месяца (Caijing/Anue); поставляйте замороженную, проверенную версию модели. Обслуживание полного DeepSeek 671B для 20–100 пользователей требует как минимум двух восьмикарточных серверов Atlas 800I A2 (China Telecom Cloud) verified, однако публичные тендеры на LLM-аппараты в Китае составили лишь 27.5 млн RMB в январе–сентябре 2025 (C114) single source: заводы платят за интеграцию и услуги, а не за коробку вычислений.

Квантование: 8-bit безопасен, 4-bit зависит от рецепта, 3-bit и ниже исключены

8-bit (FP8 или INT8) практически без потерь для визуально-языковых моделей (Red Hat, Hugging Face) verified. В 4-bit потери зависят от размера и рецепта: сборка Qwen 72B AWQ теряет менее балла на MMMU, тогда как 7B теряет 2.8 (Qwen AWQ) verified, а задачи, насыщенные деталями, падают сильнее (наивный 4-bit GPTQ стоит Qwen3-VL-30B-A3B 19 баллов на InfoVQA (VEQ)). В 3-bit и ниже средняя потеря 7–15 баллов, а диаграммы теряют 20–34 (VLMQ) verified; за пределами 64k токенов 4-bit теряет до 59%, хуже на неанглийском входе (arXiv 2505.20276). Держите визуальный энкодер в FP16/BF16.

FIG 10 **Квантование: 8-bit безопасен, 4-bit зависит от рецепта, 3-bit и ниже — вне игры** ***Инженерные ошибки в стеке обслуживания стоят больше точности, чем само квантование.***



Источник: Карта модели NVIDIA Nemotron, VLMQ, VEQ, arXiv 2609.05881, 2610.19498, 2609.04748 и др.

Два предупреждения важнее. Во-первых, **по состоянию на 11 ноября 2026 ни одно исследование или карточка модели не измеряло квантование на изображениях дефектов PCB/SMT**, поэтому перепроверяйте на золотом наборе. Во-вторых, ошибки стека обслуживания часто больше ошибки квантования: та же модель набрала примерно на 30 баллов ниже под Ollama, чем под эталонной реализацией (SiliconBench); 7 из 305 официальных тегов код-моделей Ollama были по сути сломаны, хотя файлы были целы (arXiv 2609.05881); ни один из 198 популярных репозиториев GGUF не записал свою сборку (arXiv 2610.19498); а при повторном использовании префиксного кэша 82.5% эпизодов разошлись в 4-bit против 37.4% в 16-bit (arXiv 2609.04748) single sources. Вендорский quantization-aware 4-bit близок к без потерь на задачах с деталями (NVIDIA Nemotron 3.5) vendor claim, но 4-bit VLA теряют 1.5–4.3 балла при возмущении и работают медленнее на Thor (LoopQ) single source. Управляйте квантованной моделью как новым измерительным прибором: закрепите хеши весов, записывайте сборку, сбрасывайте префиксный кэш и перепроверяйте после любого изменения точности, движка или шаблона.

Точность	≥70B плотная или ≥100B MoE	7–35B	≤4B или задачи, насыщенные деталями	Вердикт
FP8 / INT8	Безопасно (≤1 балл)	Безопасно	Безопасно; дефолт для инспекции	Использовать
4-bit AWQ/GPTQ (рецепт вендора, визуальный энкодер FP16)	0–1.5 балла	1–3 балла; наивный GPTQ на MoE с малой активацией −5.8 в среднем, −19 на InfoVQA	2–3 балла; сначала проверьте задачи с деталями	Условно
NVFP4 (рецепт вендора)	≈0.3–1 балл ниже FP8 (данные вендора)	Quantization-aware сборки почти без потерь на V*, OCRBench	Нет публичных данных	Условно (предпочитайте quantization-aware сборки)
GGUF Q4_K_M	Не тестировалось	Высокая дисперсия на доменных задачах	Не для мелкого текста или дефектов	Не для оценки деталей
GGUF Q5_K_M и выше	Предположительно безопасно	Почти без потерь в двух тестах	Нижняя граница для деталей на llama.cpp	Использовать
≤3-bit (включая 1T-модели, сжатые до 1.8–2 бит на одной машине)	Не для решений	Нет	Нет	Нет

Какая конфигурация для какой задачи

Конфигурация	Встроенная инспекция	AI-перепроверка ложных срабатываний	Планирование и управление роботом	Q&A и анализ
Специализированные детекторы или модели аномалий (YOLO, RF-DETR, класс PatchCore), INT8/FP16, любой NPU или GPU-периферийный бокс	● Да: основной судья, 2–12 мс на кроп	● Да (классификатор, настроенный на линии)	○ Только восприятие	—
Модели зрения 2–4B, NVFP4/AWQ/FP8 с визуальным энкодером FP16, на Thor/T4000 или DGX Spark	● Условно: перепроверять отмеченные AOI кропы по 0.2–0.3 с каждый	● Условно, менее 1 с после настройки	● Магистраль VLA (2–3B VLA при ≈40–130 мс на чанк)	○ Слабо
Модели зрения 7–9B (Qwen3-VL-8B, MiniCPM-V 4.5, GLM-4.6V-Flash), FP8 или вендорский 4-bit, на Thor, Spark или карте 24–32 GB	● Условно: ≈0.5 с на отмеченный кроп	● Да: 0.5–1 с на вызов, лучшая ценность	● Планирование и выбор навыков (1–3 с), не в контуре управления	● Базовые SOP Q&A для одного или нескольких пользователей
MoE 26–35B (Gemma-4-26B-A4B, Qwen3.6-35B-A3B), NVFP4/FP8	● Нет (слишком тяжело)	● Да, как объяснитель или для сложных случаев	● Планирование	● Да (8 пользователей на Spark, ≈0.1 с до первого токена)
MoE 106–241B в 4-bit на Mac Studio, RTX PRO 6000 или Spark	● Нет	● Нет	● Офлайн-планирование	● Да, для одного или нескольких пользователей

Конфигурация	Встроенная инспекция	AI-перепроверка ложных срабатываний	Планирование и управление роботом	Q&A и анализ
DeepSeek 671B или Kimi K2.x 1T в 4-bit на Mac 512 GB, сервере KTransformers или DGX Station	● Нет	● Нет	● Только офлайн-генерация программ	● Условно: 1–3 пользователя и короткие промпты; длинные промпты занимают 16–170 с
Что угодно в ≤3-bit; GGUF Q4_K_M для детальной работы малых моделей; Ollama для многих пользователей	● Нет	● Нет	● Нет	● Нет

Источники: данные TensorRT Edge-LLM, Jetson AI Lab, Apple ML Research и KTransformers выше, а также (DGX Spark tests, RTX PRO 6000 tests, Hardware Corner on Atlas, Ultralytics YOLO26, RF-DETR) **тесты сообщества и вендоров, в основном single source** .

Итак, о двух обычных опасениях: **это не обязано быть медленным**, если модель соответствует задаче (малые и средние модели зрения на Thor или Spark укладываются в каждый бюджет встроенной инспекции и перепроверки; модели 671B–1T остаются вне линии), и **квантование не обязано делать это неверным** с FP8/INT8 по умолчанию или вендорским 4-bit выше 7B, проверенным на золотом наборе. Большие риски — это примерно 57-балльная zero-shot промышленная точность универсальных моделей и ошибки стека обслуживания. Давайте запросам инспекции фиксированные размеры пакетов и собственную очередь (temperature 0 не воспроизводим по битам при пакетировании (vLLM documentation)), держите горячий резерв N+1 и делайте каждый шаг на линии отказывающим открыто.

Учитель и ученик

03 Учитель и ученик: фронтир-модели могут предразмечать данные инспекции; для манипуляторов настоящие учителя — по-прежнему люди

Кандидаты-учителя ноября 2026: универсальное зрение удвоилось, баллы на промышленных дефектах отсутствуют

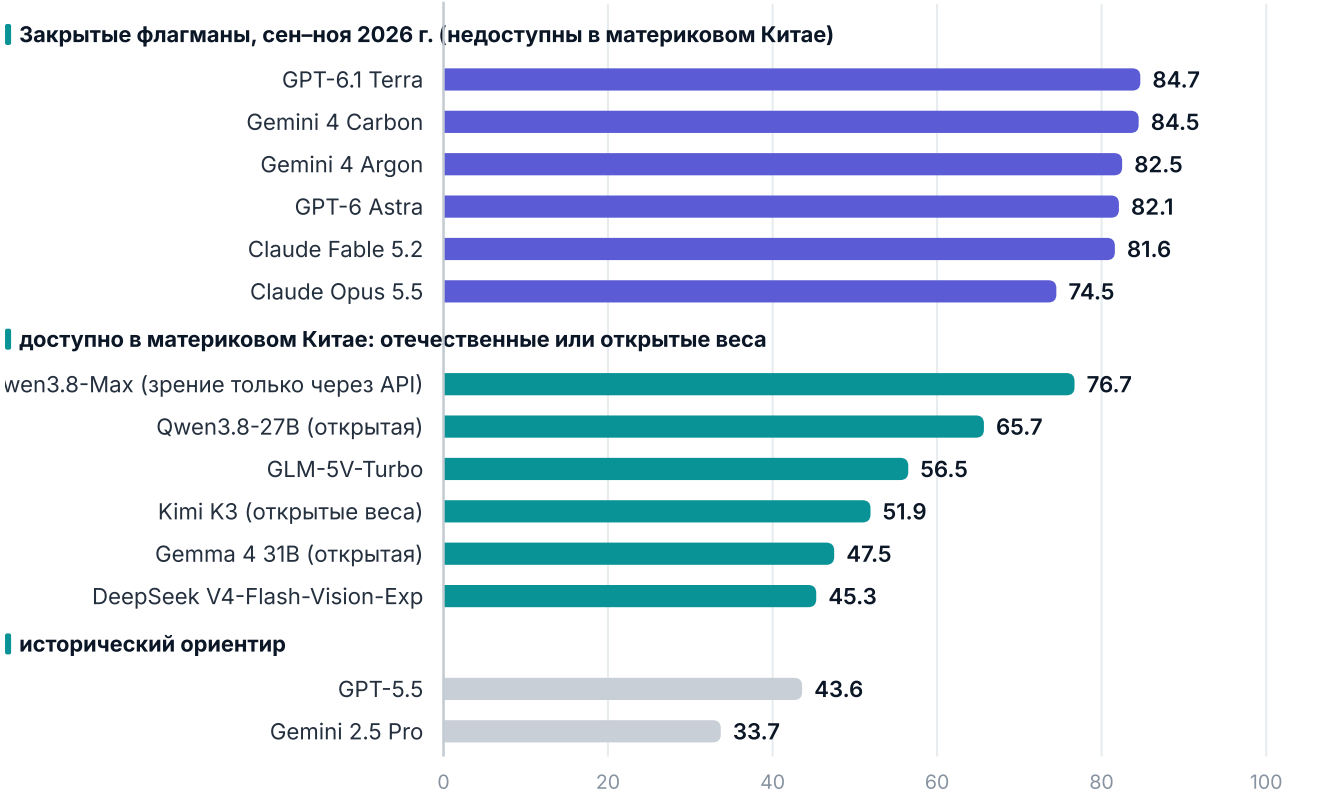
Каждый закрытый флагман, который мог бы выступать учителем в ноябре 2026, принимает изображения: Claude Fable 5.2 (общедоступна 9 ноября, $8/$40 за миллион токенов) и Opus 5.5 ($4/$20) (Anthropic, Claude models overview) verified ; GPT-6 Astra ($10/$50) и GPT-6.1 Terra ($1/$5) (OpenAI pricing); а также Gemini 4 Argon и Carbon, открытые поэтапно (Google Carbon). В материковом Китае применимы Qwen3.8-Max (зрение через API) и открытые модели, такие как Qwen3.8-27B, Kimi K3, GLM-5V-Turbo, DeepSeek V4-Flash-Vision-Exp и Gemma 4 (Roboflow blog, Hugging Face).

Ключевой пробел нужно назвать прямо: ни у одной из этих моделей нет опубликованного балла на MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB или любом наборе данных PCB/SMT. У MMAD по-прежнему нет лидерборда (MMAD GitHub), а две новейшие статьи по промышленным аномалиям (август–сентябрь 2026) всё ещё используют Gemini 2.5 Pro и Claude Sonnet 4 как свои новейшие универсальные базовые линии (Inspector-GPT, Anomaly-LR). Полнотекстовый поиск по arXiv 11 ноября дал ноль совпадений для «Fable 5.2» или «Gemini 4 Carbon» в этом контексте (arXiv API) single source, отсутствие доказательств . Ближайшие заменители:

- Новейшая фронтир-модель с баллом на MMAD — GPT-5.5: 69.95% на «дефект или нет» и 76.20% в среднем по семи задачам на подвыборке из 500 изображений, против 95.24% и 86.65% у людей-экспертов (AnomalyClaw) single source ; GPT-4o набрала 68.63% и 74.92% в 2024 (MMAD, ICLR 2025), около 1.3 балла прогресса за два года.
- На LDU-Bench, реальных изображениях SEM-контроля из литографии ИС и единственном электронном бенчмарке с моделями 2026 года, сортировка «годен/не годен» прилична (макро F1 0.932 для GPT-5.4, 0.884 для GLM-5V-Turbo, 0.876 для Claude Opus 4.6, 0.504 для Gemma 4 31B), но классификация 11 морфологий дефектов достигает лишь 0.20–0.41 (LDU-Bench) single source .
- На независимых визуальных оценках Roboflow (обновлены 10 ноября 2026), детектирование объектов mAP@50 равно 84.5 для Gemini 4 Carbon, 84.7 для GPT-6.1 Terra и 81.6 для Fable 5.2, против 33.7, 43.6 и 38.6 для Gemini 2.5 Pro, GPT-5.5 и Claude Opus 4.8 (Roboflow Vision Evals) single source, независимый оценщик, частные наборы данных .

FIG 11 **Кандидаты-учителя: универсальные баллы удвоились за поколение*Но ни у одного флагмана 2026 г. нет опубликованного балла на бенчмарке промышленных дефектов.***

Roboflow универсальное детектирование mAP@50



бенчмарки промышленных дефектов (MMAD, MVTec AD 2, VisA, Real-IAD, UniPCB): ноль опубликованных результатов у ф
универсальные баллы детектирования удвоились за поколение, но это не суждение о дефектах; тестируйте на своём золотом наборе

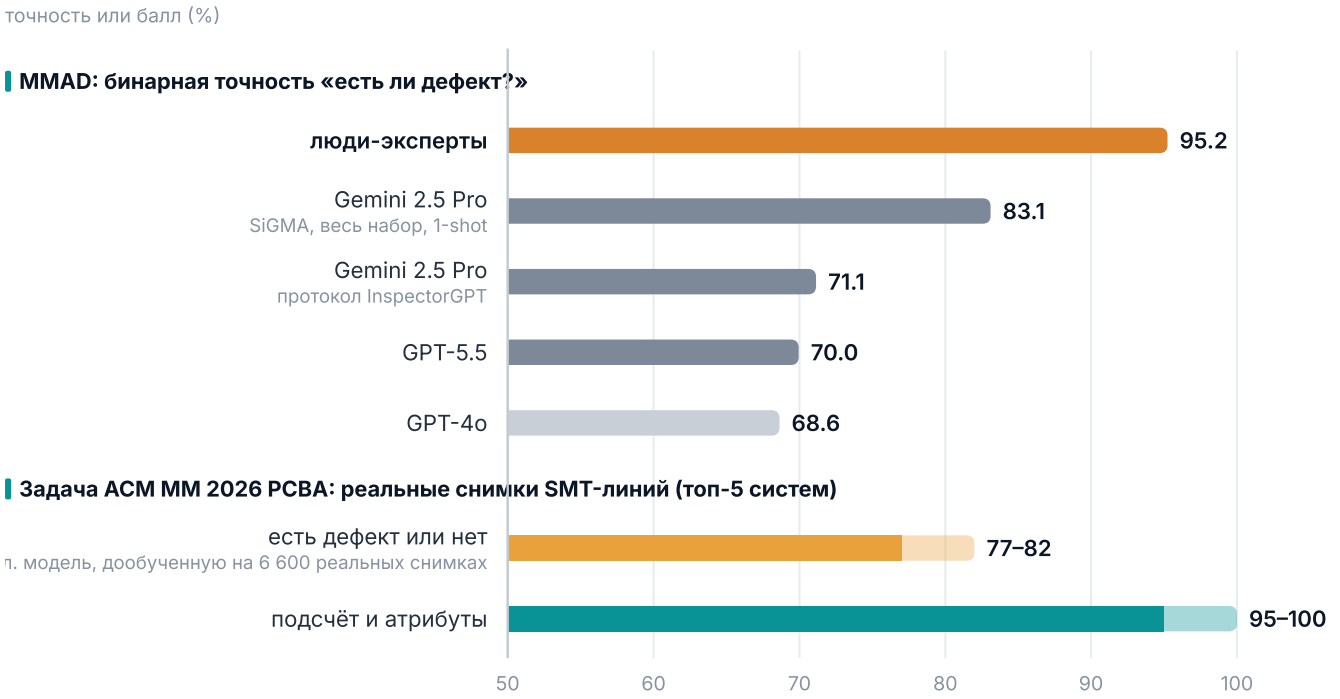
Источник: Roboflow Vision Evals; заметки о релизах моделей; страницы бенчмарков MMAD, MVTec AD 2 и др. (на 11 ноя 2026).

Кандидат-учитель (ноя 2026)	Цена за M токенов (вход/выход)	Roboflow детектирование mAP@50 / подсчёт	Бенчмарки промышленных дефектов	Применим в материковом Китае
Claude Fable 5.2	$8 / $40	81.6 / 86.5	Нет	Нет
Claude Opus 5.5	$4 / $20	74.5 / 80.6	Нет	Нет
GPT-6 Astra	$10 / $50	82.1 / 80.2	Нет	Нет
GPT-6.1 Terra	$1 / $5	84.7 / 80.2	Нет	Нет
Gemini 4 Carbon	$2 / $10 (промо)	84.5 / 90.1	Нет	Нет
Qwen3.8-Max (зрение только через API)	$2 / $6	76.7 / 81.1	Нет	Да
Qwen3.8-27B (открытая)	Локально	65.7 / 64.9	Нет	Да
GLM-5V-Turbo	$1.20 / $4.00	56.5 / 48.6	LDU-Bench сортировка 0.884	Да
Gemma 4 31B (открытая)	Локально	47.5 / 51.4	LDU-Bench сортировка 0.504	Да (локальные веса)

Кандидат-учитель (ноя 2026)	Цена за M токенов (вход/выход)	Roboflow детектирование mAP@50 / подсчёт	Бенчмарки промышленных дефектов	Применим в материковом Китае
Справка: GPT-5.5 / GPT-4o	—	43.6 / 68.0 (GPT-5.5)	MMAD бинарная 69.95% / 68.63%	● Нет
Люди-эксперты	—	—	MMAD бинарная 95.24%	—

Новые статьи всё ещё тестируют предыдущее поколение, но указывают в ту же сторону. На полном наборе MMAD люди-эксперты набирают 95.2% на «дефект или нет», а Gemini 2.5 Pro — 83.1% (SiGMA), однако та же модель набрала лишь 71.14% в другой статье (InspectorGPT): протокол теста сдвигает балл так же сильно, как и поколение модели. В задаче CVPR 2026 VAND 4.0 готовые визуально-языковые модели достигли пика 61.88 AP против 90.84 у контролируемых методов, и организаторы заключили, что они «в настоящее время не могут заменить дообученные классификаторы» (VAND 4.0) single source. В задаче ACM MM 2026 PCBA пятёрка лучших набрала 76.92–82.47 на «дефект или нет», тогда как тип компонента и подсчёт выводов достигли 95–99.5 (arXiv 2609.21276) single source : распознавать гораздо легче, чем судить.

FIG 12 **«Есть ли дефект?» всё ещё сложно для универсальных моделей** ***Лучшая универсальная модель отстаёт от людей-экспертов более чем на 12 пунктов на MMAD; дообученные системы лучше на реальных снимках линий.***



Источник: MMAD; SiGMA; InspectorGPT; задача ACM MM 2026 PCBA.

Следуют три вывода. Видение и подсчёт удвоились за поколение, но оценка аномальности движется медленно, сильно зависит от протокола и несёт систематические смещения; насколько выигрывают новые модели, можно измерить только на собственных изображениях завода. Внутри своего обучающего распределения дообученные специалисты 7B по-прежнему превосходят универсальные модели (Anomaly-LR, UniPCB). А реальные данные линии обваливают баллы: на собранных электронных панелях AUROC на уровне изображения упал с 97–98 на MVTec AD до 31.5–78.6 (arXiv 2606.01992) single source . Для материковых компаний сильнейший доступный учитель, Qwen3.8-Max, отстаёт примерно на 8 баллов от глобального фронтира.

Запуск десятков или сотен раундов: деньги — не проблема, качество разметки — проблема

Деньги — не проблема. По прайсам ноября 2026, 1 000 изображений в один мегапиксель стоят около $3 (GPT-6.1 Terra) до $32 (GPT-6 Astra) за раунд, а сто раундов на 100 000 изображениях — около $32 000–135 000 до оптовых скидок estimate (Claude vision docs, OpenAI vision guide, Roboflow Vision Evals). Для простых меток человек стоит примерно как Terra; преимущество учителя — скорость и масштаб, а не цена. NDA заказчиков обычно держат изображения продукта внутри завода, поэтому золотой набор размечается на месте.

Реальное препятствие — качество, и больше раундов не исправляют систематические ошибки. Roboflow запускает каждую задачу трижды, и баллы движутся лишь на ±0.1–1.4 балла (Roboflow), так что модель в основном повторяет свои ошибки, и голосование большинством её не спасёт. Раунды помогают только тогда, когда каждый добавляет информацию: одно–два хороших эталонных изображения добавляют около 3–4 баллов, больше — ничего (AD-Copilot); если модель спорит с хорошим эталоном на протяжении нескольких ходов, это добавляет 6.23 балла кросс-доменного AUROC для GPT-5.5 и 2.95 на MMAD, с пиком примерно после трёх вызовов (AnomalyClaw) single sources. «Ученик, обученный на метках учителя, приближается к человеческим меткам» показано только на естественных изображениях: ResNet-101, обученный на метках Qwen3-VL для COCO и Objects365, достиг 95.7–99.4% результата человеческих меток, и авторы говорят, что доменные данные не тестировались (arXiv 2602.20972) single source.

FIG 13 **Два конвейера учитель–ученик** ***Для инспекции фронтир-модель предразмечает, а люди решают; для манипуляторов настоящие учителя — по-прежнему человеческие демонстрации.***



Источник: Сводка автора (глава 3); стоимости и объёмы выборок — оценки.

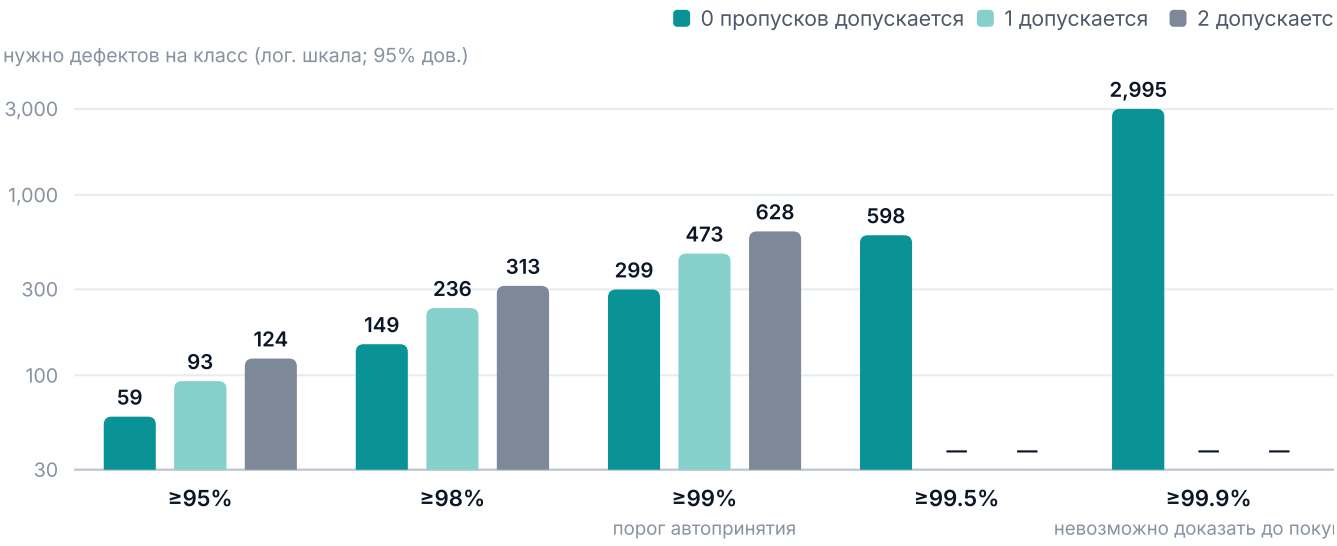
Рабочий протокол estimate:

1. **Соберите золотой набор из собственного архива AOI завода и сохранённых NG-плат**, стратифицированный на 12 классов истинных дефектов (отсутствующий, смещение, «надгробный», полярность, неверная деталь, перемычка/замыкание, недостаточный припой, избыточный припой, шарик припоя, поднятый вывод/обрыв, повреждённая деталь, посторонний материал) и четыре причины ложных срабатываний (тень, флюс, шелкография, допуск), плюс «срез дрейфа» из как минимум 300 дефектов со второй линии или при другом освещении, цвете платы или номерах деталей.

2. **Размечайте вслепую двумя или тремя инспекторами, обученными по IPC-A-610**, с решением старшего инженера и разрешением неоднозначных образцов рентгеном, электрическим тестом или поперечными срезами. Принимайте определение класса только при согласии инспекторов к ≥0.75 (≥0.90 для критических классов); человеческое согласие по 11 морфологиям электронных дефектов в LDU-Bench было лишь 0.727, так что некоторые классы потребуют слияния или переопределения. Атрибутное согласие в стиле AIAG (эффективность ≥90%, пропуск ≤2%, ложное срабатывание ≤5%, к >0.75; >0.90 для релиза) — разумный минимум (Fabrico, iFactory).
3. **Определите размер точным биномиальным (Clopper–Pearson) односторонним 95% интервалом** (binomial confidence intervals, rule of three:

Recall к доказательству	0 пропусков допускается	1 допускается	2 допускается
≥95%	59 дефектов	93	124
≥98%	149	236	313
≥99%	299	473	628
≥99.5%	598	—	—
≥99.9%	2 995	—	—

FIG 14 **Насколько большим должен быть золотой набор*Доказать recall 99% с нулём пропусков — 299 дефектов на класс; 99.9% нельзя доказать до покупки.***



Источник: Точные биномиальные доверительные интервалы (95%).

Шестидесяти дефектов на класс достаточно для экрана «годен/не годен»; 300 нужны, чтобы решить вопрос об автоматическом принятии. Полный набор, около 3 600 дефектов плюс 3 500 кропов ложных срабатываний, занимает около 37–57 человеко-часов на разметку и арбитраж, а прогон кандидата-учителя по нему стоит $23–196; реальная стоимость — от двух до шести недель сбора, заданная самым редким классом estimate. Отправляйте одиночные AOI-кропы компонентов, а не целые платы, с целью не менее 42–48 пикселей в поперечнике (arXiv 2610.09313).

1. **Принимайте метки учителя автоматически только для класса с нулём пропусков на ≥299 золотых дефектах**, с согласием модель–истина не хуже согласия человек–человек и с закрытием как минимум 50% ложных срабатываний. Критическим классам, таким как замыкания, отсутствующие детали, полярность и неверные детали, нужны 99.5–99.9% (598–2 995 дефектов); иначе всё, что ИИ не может

закрыть с высокой уверенностью, идёт человеку. Уровень пропусков ниже 0.1% дефектов нельзя доказать до покупки (2 995 дефектов с нулём пропусков; 0.01% потребовал бы 29 956), поэтому линия сохраняет человеческую выборку закрытых ИИ вызовов, нижестоящий ICT/FCT как страховку и SPC по пропускам. Проверенные метки затем дистиллируются в обычный детектор или классификатор и перепроверяются на каждом новом продукте.

Условия и правила данных тоже решают выбор учителя. Anthropic запрещает использовать свои сервисы для обучения конкурирующих моделей (Anthropic commercial terms); OpenAI разрешает использовать выходы для обучения внутренних классификаторов, не продаваемых третьим сторонам (OpenAI business terms); DeepSeek явно разрешает дистилляцию (DeepSeek terms) primary terms. Никто не определяет «конкурирующую модель», поэтому коммерческое использование требует юридической проверки, а правила данных заказчиков обычно держат изображения на месте. Самый безопасный дефолт — локально запускаемый учитель на открытых весах, с облачными фронтир-моделями только при письменном разрешении заказчика.

Ученики инспекции: малые и быстрые; узкие места — оптика, данные и валидация

Ученик — не узкое место: YOLO26n занимает 3.80 мс на изображение в INT8 на $249–399 Orin Nano Super (Ultralytics) vendor test, NVIDIA сообщает о 98.51% на трёх классах дефектов PCB после адаптации базовой модели примерно на 600 изображениях (NVIDIA) vendor claim, а синтетические дефекты добавляют лишь 0.9–2.5 балла mAP (arXiv 2605.04635). Ограничения — оптика, данные линии, валидация и лицензии: Ultralytics YOLO имеет лицензию AGPL-3.0 или корпоративную (Ultralytics), поэтому закрытым продуктам нужна лицензия или детектор Apache-2.0, такой как RF-DETR.

Ученики-роботы: структурированные станции близки к производственному уровню; гибкая и динамичная работа — нет

В ноябре 2026 есть три вида учителей-роботов, и только один может стать локальным учеником: открытые, дообучаемые VLA (π0/π0.5, GR00T N1.7, X-VLA, RDT2, SmolVLA). Закрытые модели роботов (π0.6/π0.7 через бета-API, Gemini Robotics 2 для доверенных тестировщиков) нельзя развернуть на месте (The AI Insider, DeepMind). Фронтир-универсальные модели медленны и плохи в тонких манипуляциях: в тесте Artificial Analysis они набрали 46–69 с 28–45% успеха захвата и около 30 минут на задачу, «слишком медленно и дорого для любого производственного использования» (AA-Robotics); GPT-6 Astra справилась с точной вставкой лишь 2 раза из 20 (RoboCurve); а прямое управление суставами даёт 0–5.5% на LIBERO при 0.2–0.4 Гц против необходимых около 83 Гц (Anthropic) verified по трём источникам. 64.9% Fable 5.2 на симулированном бенчмарке RoboDojo — внутренний прогон (Anthropic) vendor claim.

Сторонние исследования с октября согласны: около 33–44% на LIBERO-10 для трёх ведущих агентов (arXiv 2610.30051); агенты ложно заявляют об успехе на 24–39% подзадач (arXiv 2611.08209); и 104 из 108 квазистатических задач на реальном роботе для Opus 5.5, но по 280 секунд каждая и без вставки или сборки (AMP1) single sources. Фронтир-модели принадлежат офлайну: написание программ, оркестрация задач и черновики критериев успеха.

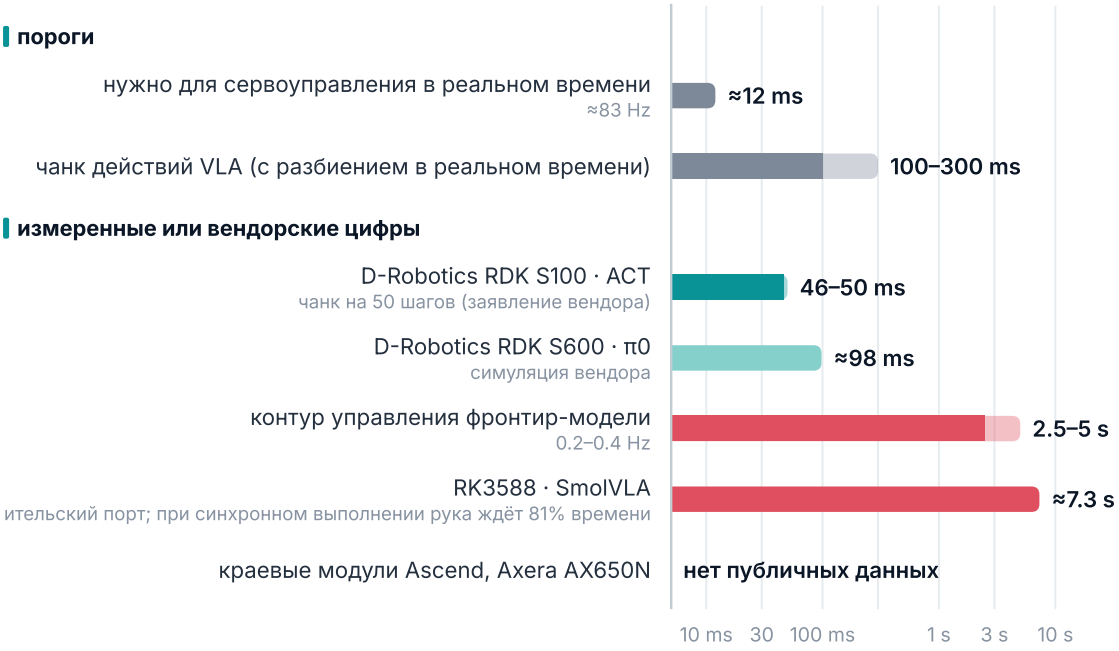
Качество учеников имеет чёткую границу: хорошо внутри обучающего распределения, обваливается вне его. π0.7 «часто превышает 90%» на виденных задачах и достигает 60–80% на невиденных (π0.7 paper) vendor research; ContactDP достигает 97.8% на обученных разъёмах, но 28.6–66.7% на незнакомых (ContactDP); лёгкое изменение камеры или позы опускает 95% ниже 30% в LIBERO-Plus (LIBERO-Plus, CVPR 2026); обученный захват с движущихся конвейеров достигает пика 58.89% (ForeTime-VLA) single sources, работу, которую классическое визуальное отслеживание выполняет в производстве годами.

Близкие к производству доказательства — все со структурированных станций, человеческих демонстраций и коррекции на реальном роботе. Neuromeka сохранил конечный автомат и предварительно обученное движение и обучал только захват, вставку кабеля и пайку, каждый менее чем за 20 минут данных: 322 из 324 в производстве, на 12.8% медленнее человека (Neuromeka) single source, vendor. Fraunhofer достиг 92.8% на че-

тырёх разъёмах за 68 минут демонстраций (Fraunhofer); HIL-SERL Беркли преуспел 100 раз из 100 на вставке RAM, SSD и USB после 1–2.5 часов RL на реальном роботе (HIL-SERL); AgiBot работал 18–20 секунд на шаг в пределах 20-секундного такта на линии тестирования планшетов (IThome) vendor claim .

Дистилляция покупает скорость, а не обобщение: π0.5 работает на 0.70 Гц на Jetson Orin «сыром» и 10.30 Гц после асинхронного планирования (EagleVLA); дистилляция слоёв ускоряет её с примерно 3 Гц до примерно 10 Гц (Shallow-π) verified . Ученик наследует только то, что покрыл учитель; каждому новому продукту всё равно нужны данные задачи.

FIG 15 **Задержка краевой политики** ***Малая политика ACT работает за ≈50 ms на краевой плате; VLA, портированный на любительский чип, занимает 7.3 s.***



Источник: Документация и форум D-Robotics (цифры вендора); порты сообщества GitHub; исследования Anthropic.

На китайских периферийных чипах единственная измеренная обученная политика — политика ACT на D-Robotics RDK S100 при 46–50 мс на чанк из 50 шагов, без опубликованного уровня успеха (D-Robotics RDK S docs) vendor claim ; любительский порт SmolVLA на RK3588 занимает около 7.3 с на чанк (GitHub). Квантование не даёт скорость бесплатно: ни одна сборка π0.5 в 4-bit не была быстрее 16-bit при размере пакета 1 (LoopQ). Демонстрации дороги (Reuters сообщает о примерно 300 попытках на одно пригодное движение для новичка-телеоператора (Notebookcheck)), а валидация дороже: доказательство уровня отказов ≤0.1% при 95% уверенности занимает около 3 000 последовательных успехов, 17–22 робото-часа estimate . Ни одна обученная политика манипуляций не имеет сертификации безопасности.

FIG 16 **Роботы-ученики: почти производственный уровень только на структурированных станциях** ***Станции с оснасткой достигают 90–99%+; неструктурированные, динамичные и zero-shot задачи пока нереальны.***



Источник: Статьи и материалы вендоров из главы 3; часть цифр — заявления вендоров.

Задача электроники	Рабочий ученик	Ожидаемый успех	Скорость против человека	Стоимость переналадки	Вердикт 2026–2028
Многономенклатурный pick-and-place, комплектование, обслуживание станков (статичный лоток или приспособление)	Обученное восприятие плюс программируемое или планируемое движение	≥99% с классическим визуальным наведением	Как у обычных станций	Часы	**Осуществимо сейчас**
Закреплённый разъём, вставка FFC и кабеля, завинчивание, пайка	Конечный автомат + малая политика BC/диффузия/ACT + камера на запястье + сила/момент (в идеале касание) + коррекция человека	90–99.4%	На 10–15% медленнее	10–300 демо (≈15–70 мин) + 1 000–3 000 прогонов валидации (6–22 робото-часа)	**Осуществимо, тяжело в R&D месяцы на первую станцию**
Обслуживание структурированной тестовой станции и сортировка OK/NG	Платформа VLA вендора + цифровой двойник + RL на реальном роботе	≥99.5% vendor claim, противоречиво	18–20 с на шаг, в пределах такта 20 с	≤4 ч дообучения + 15 мин калибровки	**Осуществимо на платформе вендора**

Задача электроники	Рабочий ученик	Ожидаемый успех	Скорость против человека	Стоимость переналадки	Вердикт 2026–2028
Программирование переналадки	LLM офлайн пишет или редактирует программы робота, проверяемые в симуляции	77–99.7% на переписывании ABB RAPID	Без влияния на такт	Минуты–часы на программу	**Осуществимо как инженерный инструмент**
Сквозное обученное взятие с движущихся конвейеров	Ученик VLA	37–59%; ≤37% на быстрых лентах	—	≥458 демо на установку	**Не раньше 2028; используйте классическое отслеживание плюс обученное восприятие**
Новые продукты без новых данных (zero-shot)	Универсальный VLA	60–80%; незнакомые разъёмы 29–67%	—	—	**Нереалистично**
Фронтир-LLM в контуре, управляющий рукой	Агенты GPT-6, Claude, Gemini	Захват 28–45%; точная вставка 10%	≈30 мин или 280 с на задачу против ≈20 с	—	**Нереалистично**
Свободно прокладываемые жгуты и многошаговая гибкая сборка	—	Нет производственных доказательств 2026	—	—	**Не раньше 2028**

Источники: исследования выше и (University of Passau on RAPID rewriting, GraspVLA, Komax) **в основном single source и данные вендоров** . В 2026–2028 ценность большой модели для роботов — как предобученный априор и офлайн-программист, а не как то, что «учит» ученика. Каждый близкий к производству кейс, который я нашёл (Neuromeka, AgiBot у Longcheer, Fraunhofer, ContactDP), опирается на человеческие демонстрации, датчики силы или касания, симуляцию или цифровые двойники и контролируемый RL. Ни один не опирается на метки фронтир-модели.

Затраты и окупаемость

04 Затраты и окупаемость: AI-перепроверка окупается за месяцы, обслуживание станков — за годы, манипуляторы для отбора — никогда

Стоимость труда: США примерно в 5.5 раза дороже Шэньчжэня, Тайвань примерно в 1.5 раза, большая часть Юго-Восточной Азии — менее половины

Окупаемость замены труда начинается с того, сколько стоит труд. Официальный ориентир зарплат Шэньчжэня ставит медиану 2024 года для инспектора качества производства на **74 074 RMB** в год; зарплаты составляют 86.92% полной стоимости труда в производстве электроники, поэтому полная стоимость около 85 000 RMB. Опрос JETRO 2025 ставит годовую стоимость китайского заводского рабочего в $12 667, около 85 100 RMB, в пределах 0.2% (Shenzhen wage guidance, JETRO 2025) verified .

Регион	Полная годовая стоимость одного инспектора или грузчика (низ / баз / высок)	Основа
Материковый Китай, Шэньчжэнь	70 тыс. / **85 тыс.** / 100 тыс. RMB ($10.4 тыс. / **12.7 тыс.** / 14.9 тыс.)	Официальная медиана × 1.15; совпадает с JETRO verified
Гуандун, вторая зона зарплат (Дунгуань и др.)	65 тыс. / **75 тыс.** / 85 тыс. RMB	Медиана Гуандуна 2023 × 1.15 плюс два года повышений estimate
Дельта реки Янцзы (Сучжоу, Куньшань, Уси)	Как Шэньчжэнь	Нет оснований для более дешёвого труда инспекции; максимальная минимальная зарплата Цзянсу примерно на 5.6% выше шэньчжэньской estimate
Тайвань	$16.6 тыс. / **19.6 тыс.** / 22.7 тыс. (534 тыс. / **630 тыс.** / 729 тыс. NT$)	Опрос зарплат Министерства труда плюс ≈1.18 уставных начислений; JETRO как верхняя граница estimate
Малайзия	$6.8 тыс. / **9.1 тыс.** / 10.0 тыс.	JETRO и минимальные зарплаты estimate
Таиланд	$4.7 тыс. / **7.0 тыс.** / 8.8 тыс.	То же
Индонезия	$5.0 тыс. / **6.0 тыс.** / 6.3 тыс.	То же
Вьетнам	$3.2 тыс. / **5.3 тыс.** / 6.5 тыс.	То же
Филиппины	$4.0 тыс. / **4.5 тыс.** / 4.9 тыс.	То же
США	$54 тыс. / **70 тыс.** / 80 тыс.	Медиана инспектора BLS за май 2025 $48 570 ÷ зарплаты при 65.5% компенсации estimate

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

FIG 17 **Сколько стоит работник инспекции или погрузки в год** ***Полная стоимость: США ≈5.5× Шэньчжэня, Тайвань ≈1.5×.***

полная годовая стоимость работника QC или погрузки ($10k; точка = база, отрезок = мин–макс)



Источник: Официальные ориентиры зарплат Шэньчжэня, Минтруд Тайваня, JETRO, BLS; оценки автора.

Три поправки задают основу. Во-первых, Шэньчжэнь поднял нижнюю границу пенсионных взносов с 4 775 до 5 521 RMB в месяц в июле 2026 (Shenzhen Bendibao) single source, поэтому инспектор в Шэньчжэне в 2026 скорее стоит 80–90 тыс. RMB. Во-вторых, пост — это не человек: две 12-часовые смены требуют около 2.0 человек (шесть дней в неделю) до 2.4 (весь год) в дельте Жемчужной реки, около 2.7–3.1 в Юго-Восточной Азии при лимитах сверхурочных и 4.2–4.5 для 24×7 в США и на Тайване estimate. AI-перепроверка обычно позволяет одному перепроверяющему покрывать две–четыре линии, поэтому реалистичная экономия — **0.5–1.5 штатных человек на линию, а не один человек на линию за смену**. В-третьих, покупатели на Тайване, во Вьетнаме и Малайзии гонятся не только за ценой: в опросе JETRO 61.1% фирм на Тайване, 48.2% во Вьетнаме и 44.9% в Малайзии сказали, что нанимать стало труднее за последние два года, против 18.6% в материковом Китае (JETRO 2025) single source.

Цены оборудования: четыре базы, которые не следует усреднять

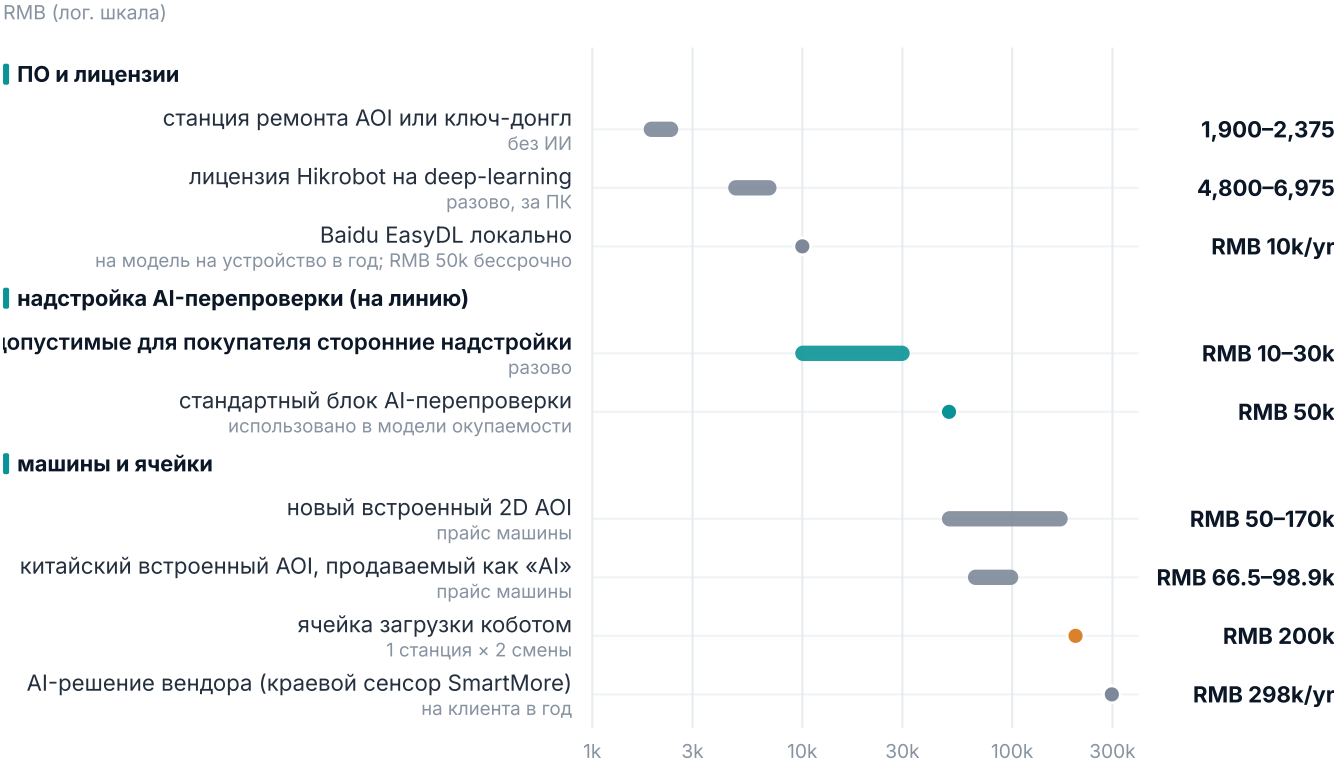
Цены на AOI расходятся в основном потому, что измеряют разное. Онлайн-листинги (ноябрь 2026, новые, б/у и аренда вперемешку) дают около 17–80 тыс. RMB за офлайн или настольный 2D, 50–170 тыс. RMB за встроенный 2D и 80–475 тыс. RMB за встроенный 3D, до 700 тыс. RMB за корейский 3D (1688 inline AOI, 1688 offline AOI, 1688 3D AOI) single source, прайсы. Тендер ноября 2025 ограничил один 3D AOI плюс один 3D SPI суммой 800 тыс. RMB (tender notice). Проспект Sinic-Tek подразумевает 428–512 тыс. RMB за 3D AOI в 2019–2022, около 340–470 тыс. RMB по её отчёту 2025 (Sinic-Tek prospectus, Sinic-Tek 2025 annual report) primary + estimate. Зарубежный вендор оценивает $3 000–20 000 за настольный 2D, $30–70 тыс. за встроенный 2D, $50–90 тыс. за начальный 3D и $100–200 тыс.+ за высококлассный 3D, с той же встроенной машиной по $30–50 тыс. в Китае (Averroes) vendor estimate от продавца конкурирующего AI-ПО.

Сама AI-перепроверка не имеет публичной цены в RMB; вендоры котируют в каждом случае после бесплатной оценки на месте (Sohu on Hengtianyi), а поиски по листингам электронной коммерции, тендерам, статьям вендоров и тайваньским СМИ не дали публичной котировки (1688, 51SMT). Якоря: около €27 тыс. на линию, подразумеваемые кейсом Siemens estimate; $21 924 по прайсу за модуль сети удалённой перепроверки дистрибьютора в США (Manncorp) vendor list price; около 298 тыс. RMB на заказчика в год за периферийные AI-сенсоры SmartMore в 2025 (36Kr) single source. Отечественный встроенный 2D AOI, продаваемый как «AI», стоит 66.5 тыс. RMB (1688 inline AOI), а NG-буфер для ручной перепроверки — около 50 тыс. RMB (1688 AOI review). Итак, потолок для сторонней надстройки перепроверки — около 10–30 тыс. RMB на линию или анало-

гичная годовая плата; выше этого покупатель скорее заменит машину отечественным AOI со встроенным ИИ estimate.

Якорь	Цена (RMB)	Что покупается	Источник
Станция ремонта AOI или ключ-донгл ПО (старые TRI, Zhenhuaxing)	≈1 900–2 375	Лицензионный ключ, без ИИ	1688 [прайс]
Универсальная среда deep-learning зрения (донгл Hikrobot VisionMaster)	≈4 800–6 975 разово за ПК	Платформа; заказчик строит и обслуживает модели	1688 [прайс]
Локальное развёртывание Baidu EasyDL	10 000 за модель за устройство в год, или 50 000 бессрочно	Упакованный сервис инференса	Baidu AI docs [страница цен 2022]
Отечественный встроенный AOI, продаваемый как «AI»	66 500–98 900 за машину	Замена машины	1688 [прайс]
Периферийные AI-сенсоры SmartMore	≈298 000 на заказчика в год	Решение вендора на нескольких машинах	36Kr single source

FIG 18 **Ценовая лестница** ***От нескольких тысяч RMB за лицензию ПО до RMB 200–300k за робоячейку или AI-решение вендора.***



Источник: Листинги 1688, документация Baidu AI (прайс 2022 г.), 36Kr, отраслевое интервью (не подтверждено).

Годовая плата в стиле Baidu, с одним сервером инференса на одну–четыре линии и одной–четырьмя моделями на каждый, подразумевает подписку 2.5–40 тыс. RMB на линию в год, реалистично около 5–10 тыс. RMB; при 20% от бессрочной лицензии подписка и покупка выходят на безубыточность на пятый год — структура, которую стоит скопировать estimate. Покупатели привыкли к почти нулевым платам за ПО: вся «прочая» выручка Sinic-Tek (запчасти, услуги, аренда, аксессуары) составила 7.65% продаж 2025, около 2 300–3 100 RMB на установленную машину в год (Sinic-Tek 2025 annual report) primary + estimate. Подписка на ИИ в 5–10 тыс. RMB на линию в год — это в два–пять раз больше того, что покупатели сейчас тратят на ПО AOI, поэтому её

надо обосновывать высвобождаемыми часами перепроверки (около 75–90 тыс. RMB на штатного человека в год), а не сравнением с существующими платами за ПО.

Для манипуляторов средняя цена кобота в Китае упала до примерно 58 600 RMB в 2025 (GGII) estimate, а средняя цена шестиосевого Dobot упала на 19.1% за тот год по мере роста доли дешёвых моделей (Huxiu) verified, хотя прайсы тех же моделей выросли на 5–15% в 2026 (Sina Finance). В США FANUC CRX-10iA стоит по прайсу $51 434 (Vention), а подборки ставят UR7e в $30–48 тыс. (cobotcost) estimate. Манипулятор — лишь 40–50% ячейки в США (Standard Bots); в Китае камеры, захваты, безопасность и инжиниринг легко достигают двух–четырёх раз от RMB 40 тыс. манипулятора, поэтому закладывайте абсолютную стоимость инжиниринга, а не кратность estimate.

Окупаемость: перепроверка за месяцы, обслуживание за годы, отбор — никогда

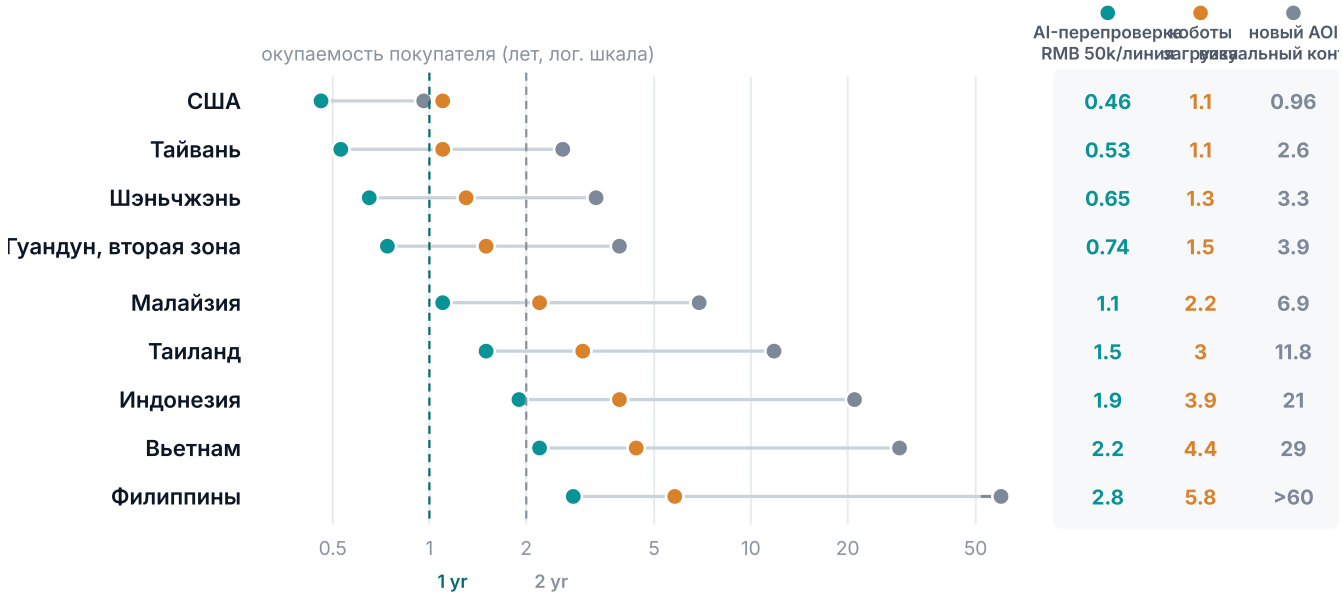
Простая окупаемость = инвестиции ÷ (сэкономленные люди × полная стоимость труда − годовые эксплуатационные расходы), для двухсменной линии без субсидий и пошлин. Оборудование на Тайване оценено в 1.3 раза от Китая, Вьетнам, Малайзия и Таиланд в 1.15 раза, Индонезия и Филиппины в 1.2 раза, а США — по зарубежным ценам estimate.

Сценарий	Инвестиции (Китай / США)	Сэкономленные люди	Годовые эксплуатационные расходы
(a) Бокс AI-перепроверки или AI-камерный бокс	50 тыс. RMB / $30 тыс. на линию	1.0 на линию	15%
(b) Новый встроенный 2D AOI вместо ручного визуального контроля (остаётся 0.5 перепроверяющего за смену)	200 тыс. RMB / $60 тыс.	1.0	12%
(c) Ячейка загрузки коботом (одна станция × две смены)	200 тыс. RMB / $140 тыс.	2.0	8%
(d) Интегрированная ячейка «детектирование и отбор» (инспекция 1.0 + отбор 0.07)	300 тыс. RMB / $200 тыс.	1.07	8%
(e) Манипулятор, добавленный только для отбора NG-плат	150 тыс. RMB / $100 тыс.	0.07	8%

Регион	(a) AI-перепроверка	(b) Новый AOI	(c) Ячейка кобота	(d) Детектирование и отбор	(e) Манипулятор для отбора
Шэньчжэнь	0.65 года (0.54–0.80)	3.3 (2.6–4.4)	**1.3** (1.1–1.6)	4.5 (3.6–5.9)	Никогда
Гуандун, вторая зона	0.74	3.9	1.5	5.3	Никогда
Тайвань	0.53	2.6	1.1	3.6	Никогда
Малайзия	1.1	6.9	2.2	9.2	Никогда
Таиланд	1.5	11.8	3.0	15	Никогда
Индонезия	1.9	21	3.9	25	Никогда
Вьетнам	2.2 (1.6–4.5)	29	4.4	34	Никогда

Регион	(a) AI-перепроверка	(b) Новый AOI	(c) Ячейка кобота	(d) Детектирование и отбор	(e) Манипулятор для отбора
Филиппины	2.8	>60 (эксплуатационные расходы ≈ экономия)	5.8	≈100	Никогда
США	0.46	**0.96**	**1.1**	3.4	Никогда

FIG 19 **Окупаемость по регионам*AI-перепроверка окупается за месяцы в США, на Тайване и в Шэньчжэне и за 1–3 года в Юго-Восточной Азии; обслуживание коботом — за годы.***



манипулятор только для отбора NG-плат никогда не окупается нигде
при выходе годных выше 90% отбор занимает лишь 0.003–0.14 FTE за смену

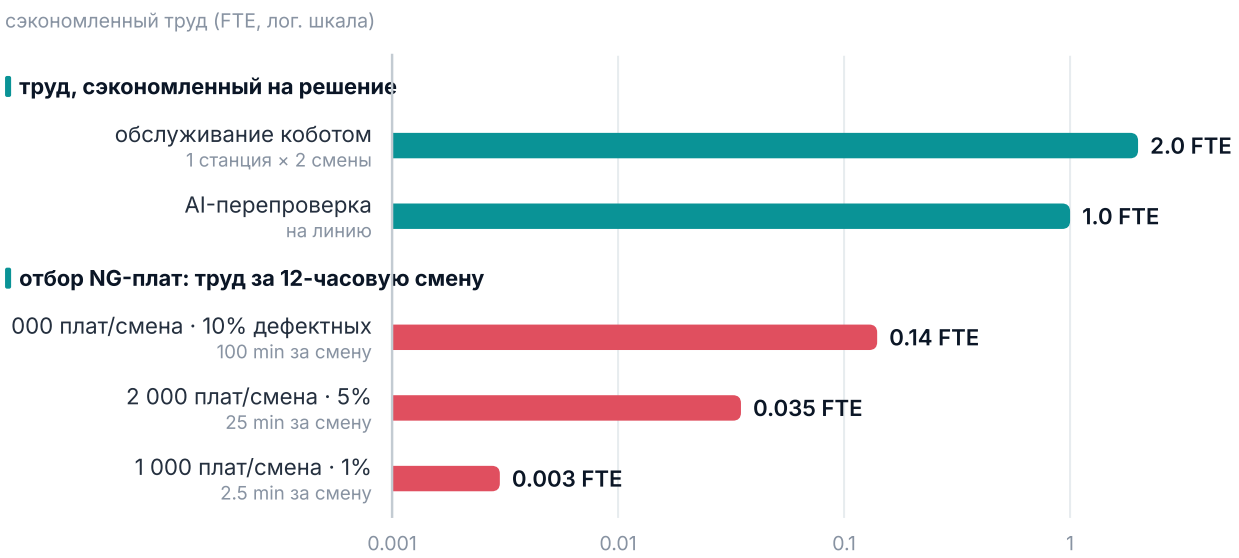
Источник: Модель автора: бокс AI-перепроверки RMB 50k на линию; ячейка обслуживания коботом RMB 200k или $140k, две смены.

Все значения — оценки из указанных выше входных данных. Чувствительности: сократите сэкономленных людей вдвое — и каждая окупаемость примерно удвоится; работайте в одну смену — и (b)–(d) удвоятся; 100% амортизация первого года в США сокращает окупаемость примерно на 20%, тогда как пошлины на китайское, японское или корейское оборудование удлиняют её примерно на 10–40%; местная субсидия 15% в Китае сокращает её примерно на 15%. Я не нашёл независимого опроса порогов окупаемости покупателей; вендоры говорят, что AOI окупается «обычно за 9–18 месяцев» (Averroes), а кейс Siemens заявляет восемь месяцев vendor claims. Я использую один и два года как линии отсечения. Модель опускает выигрыши в качестве (гарантия, возвраты) и затраты на текучесть и обучение, поэтому она консервативна.

Контринтуитивная часть «детектирования и отбора» в том, что труд — в перепроверке, а не в снятии немногих реальных NG-плат с линии. При 15 секундах обработки на NG-плату estimate:

Платы за смену	Уровень дефектов	Время отбора за 12-часовую смену	Люди
1 000	1%	2.5 мин	0.003
2 000	5%	25 мин	0.035
4 000	10%	100 мин	0.14

FIG 20 **Почему отбор плохих плат не окупается** ***При выходе годных выше 90% отбор занимает лишь 0.003–0.14 человека за смену.***



Источник: Оценки автора.

Базовый случай (2 000 плат за смену, 5% брака, две смены) — 0.07 человека, стоимостью около 6 000 RMB в год в Шэньчжэне, меньше примерно 12 000 RMB годовых эксплуатационных расходов манипулятора, поэтому он не окупается нигде; даже худший случай (4 000 плат за смену, 10% брака) занимает около 13 лет в Шэньчжэне и 9 в США estimate. Если беспокоит остановка линии для отбора, стандартный NG-отводчик или буферный конвейер решают это без манипулятора. Люди в отрасли говорят «при выходе годных с первого прохода выше 90% остановите линию и снимите плату вручную», и **это верно для отбора, но не для инспекции**. Поэтому дизайн продукта на линии SMT — «инспекция плюс маркировка или отвод», с манипуляторами для мест, где обрабатывается каждая деталь (загрузка тестовых приспособлений, работа с лотками и магазинами, упаковка), где инспекция становится почти бесплатным дополнением и применим сценарий (c).

Субсидии, налоги и пошлины сдвигают окупаемость на 15–25%, но не порядок

Материковый Китай: поддержка в основном снижает стоимость финансирования; отдельные машины редко квалифицируются напрямую. Национальная субсидия на кредит платит 1.5 пункта процента на срок до двух лет в 14 приоритетных цепочках (Yicai) verified, около 9 000 RMB на машину за 300 тыс. RMB; награда Гуандуна за модернизацию до 20% требует зарегистрированного проекта с не менее 7 млн RMB нового оборудования и исключает Шэньчжэнь (Guangdong rules), а Дунгуань фактически выплатил 4.42% в июле 2026 (Dongguan allocation) single source, primary. Устройство AI-инспекции одобряется только как оборудование «с сопутствующим ПО» внутри зарегистрированного проекта покупателя, поэтому базовый случай предполагает отсутствие субсидии.

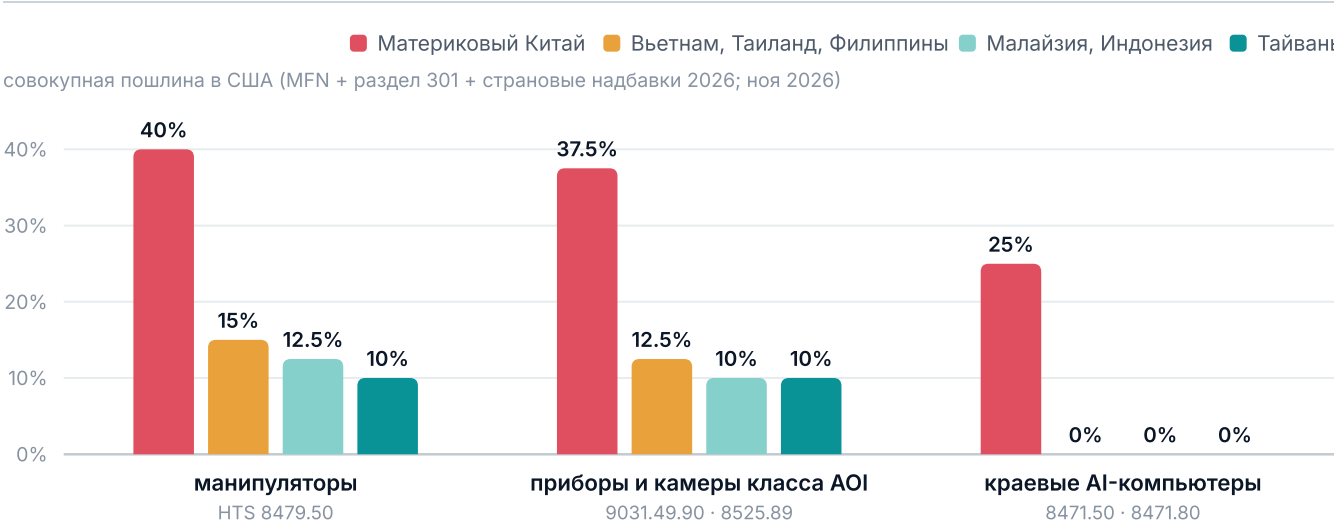
Тайвань: стимулы идут покупателю. По статье 10-1 Закона о промышленных инновациях, изменённой в мае 2025 для включения ИИ, покупатель может зачесть 5% расходов 2025–2029 на новые умные машины или ИИ (включая инспекцию на CNN/DNN) против подоходного налога того года, или 3% в год на три года (Statute, credit rules) verified; гранты для поставщиков закрыты для материкового капитала (MOEA) single source.

Юго-Восточная Азия: только Малайзия и Таиланд вознаграждают покупку автоматизации. 200% капитальная надбавка Малайзии на первые 10 млн RM стоит около 24% capex прибыльному покупателю (MIDA, PwC Malaysia) verified; BOI Таиланда даёт трёхлетнее освобождение от налога до 50–100% инвестиций в модернизацию, с более высоким потолком при участии тайских поставщиков автоматизации (BOI) single source, primary, поэтому вход в Таиланд лучше всего работает с локальным интегратором. Вьетнам, Индонезия и Филиппины вознаграждают строительство заводов, а не покупку автоматизации (Vietnam Law Magazine, PwC Philippines).

США: амортизация сокращает окупаемость; пошлины удлиняют её. Активы, купленные после 19 января 2025, получают постоянную 100% амортизацию первого года, около 21% федерального налогового щита для прибыльного покупателя (IRS Notice 2026-11) single source, primary ; расследование раздела 232 по роботам и промышленному оборудованию завершилось 28 октября 2026 переговорами, а не новыми пошлинами, с переоценкой в течение 180 дней (White House, Mayer Brown) verified . Финальная сборка в Юго-Восточной Азии сокращает надбавку на манипуляторы и приборы типа AOI с примерно 37.5–40% до 10–15% (глава 1).

Регион	Основные инструменты	Влияние на окупаемость покупателя estimate
Материковый Китай	Субсидия на кредит 1.5 пункта до двух лет; провинциальные и городские награды за модернизацию (Гуандун требует ≥7 млн RMB зарегистрированных проектов; Дунгуань фактически выплатил 4.42%); меры Шэньчжэня 2026–2029 (ставки не публичны)	Около 0–5% для отдельной машины; 4–15% внутри крупного проекта
Тайвань	Зачёт §10-1, 5% или 3% × 3 года	Примерно на 5% короче, если покупатель платит достаточно налога
Малайзия	200% капитальная надбавка за автоматизацию (первые 10 млн RM)	Примерно на 15–25% короче для прибыльных покупателей
Таиланд	Трёхлетнее освобождение BOI до 50–100% инвестиций	Примерно на 15–25% короче для прибыльных покупателей вне налоговых каникул
Вьетнам, Индонезия, Филиппины	Налоговые льготы за строительство заводов; дополнительные 20% амортизации оборудования на Филиппинах	Почти ничего для одиночных покупок автоматизации
США	100% амортизация первого года; ≈37.5–40% совокупной пошлины на оборудование материкового производства	Амортизация ≈на 20% короче; пошлины ≈на 10–40% длиннее

FIG 21 **Пошлины США на роботов, AOI и краевые компьютеры** ***Манипуляторы из материкового Китая платят около 40%; из Юго-Восточной Азии или Тайваня — 10–15%.***



Источник: Гармонизированный тарифный список USITC, глава 99; Federal Register; USTR (ноябрь 2026).

По регионам: в материковом Китае бокс AI-перепроверки окупается за месяцы, а ячейка кобота за 1.3–1.5 года, тогда как новый AOI, купленный только для сокращения труда, занимает 3–4 года, а интегрированное детектирование-и-отбор 4.5–5.3, оба за пределом двухлетнего экрана. Тайвань чуть лучше Шэньчжэня (0.53

года для перепроверки, 1.1 для ячейки кобота), и нанимать там труднее. В Малайзии и Таиланде AI-бокс занимает 1–1.5 года, а ячейка кобота 2–3, примерно на 15–25% быстрее с локальными налоговыми стимулами для прибыльных покупателей. В Индонезии, Вьетнаме и на Филиппинах только очень дешёвый AI-бокс окупается в пределах 2–3 лет; спрос там больше исходит из требований качества заказчиков и нехватки труда. В США каждый вариант, кроме чистого манипулятора для отбора, окупается за 0.2–3.4 года, с пошлинами, переоценкой раздела 232 и сертификацией NRTL как главными переменными.

05

Рынок и конкуренция

Рынок, конкуренция и как платят покупатели

Размер рынка: опубликованные цифры различаются в разы; полезен только подсчёт снизу вверх

Каждую цифру рынка надо читать с её охватом. Рынок машинного зрения Китая составил 18.15 млрд RMB в 2024 (Zhiyan) verified, с электроникой 3C как крупнейшим покупателем; глобальные оценки машинного зрения различаются примерно втрое, а AOI — примерно в 6.5 раза (MarketsandMarkets, Vision Systems Design/Interact Analysis, Zhiyan). Снизу вверх из выручек Koh Young, TRI, Jutze, Sinic-Tek и ViTrox, глобальное оборудование инспекции SMT — порядка $1–2 млрд estimate.

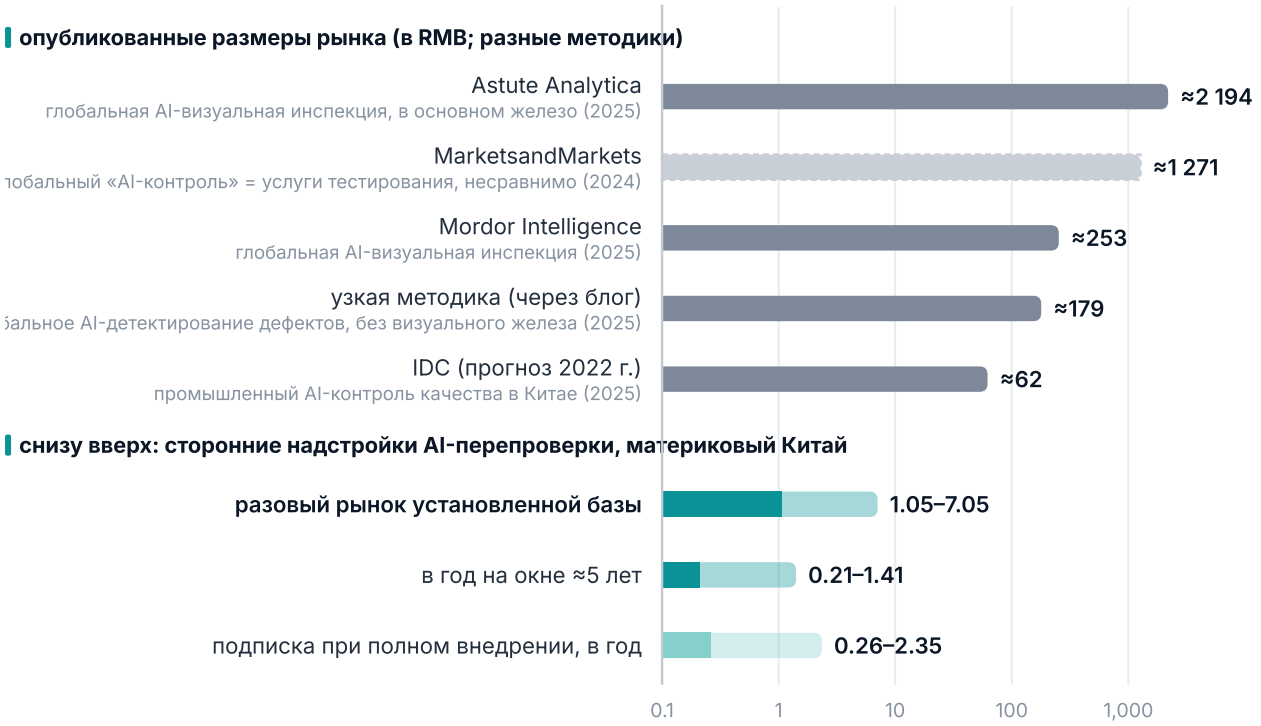
Отдельного рынка «ПО для визуальной инспекции на ИИ» не существует. Опубликованные цифры различаются более чем в десять раз и все включают оборудование или услуги:

Охват	Издатель	Цифра
Глобальное «AI-детектирование промышленных дефектов» (узко, без визуального оборудования)	Неназванная фирма, через блог iFactory	$2.66 млрд в 2025, CAGR 8.6% до 2035, электроника 34%
Глобальная «AI-визуальная инспекция»	Mordor Intelligence	$3.76 млрд в 2025, $12.8 млрд в 2031, CAGR 22.67%
Глобальная «AI-визуальная инспекция» (в основном оборудование)	Astute Analytica	≈$32.66 млрд в 2025
Глобальный «AI-контроль» (услуги тестирования и сертификации; несравнимо)	MarketsandMarkets	$18.92 млрд в 2024
Китайский «промышленный AI-контроль качества»	IDC, прогноз 2022 года	$958 млн в 2025 (≈6.2 млрд RMB), 91.5% в связи, электронике и двух других отраслях

Источники: (iFactory, GII/Mordor, GII/Astute, MarketsandMarkets, IDC) **каждый single source, в основном пересказы платных отчётов**.

FIG 22 **Размеры рынка различаются более чем в десять раз*Для решений полезны только узкая методика и подсчёт снизу вверх.***

RMB 100M (лог. шкала; USD по курсу 6.7188)



Источник: Mordor Intelligence, Astute Analytica, MarketsandMarkets, IDC (2022), iFactory; оценка автора снизу вверх.

Снизу вверх есть только один публичный подсчёт линий SMT: New Venture Research оценила около 73 231 линии в мире в 2020, 47% из них в Китае (19% в 2000), около 70% высокого класса и 30% низкого (EMSNow, NVR SMT2020 summary) single source, 2020 ; NVR теперь принадлежит IPC, и новые данные платные (NVR). Рост на 3–5% в год с долей Китая 40–47% даёт около 87 000–98 000 линий в мире в 2026 и 35 000–46 000 в материковом Китае (база 40 000) estimate .

Вход	Низ	Баз	Высок	Основа
Линии SMT материкового Китая (2026)	35 000	40 000	46 000	База NVR 2020 × рост × доля
Доля без AI-перепроверки	75%	80%	85%	Окно модернизации, ниже
Доля, достижимая третьей стороной (не внутренняя, не привязана к OEM-ИИ, может платить)	40%	50%	60%	Допущение: ≈30% линий низкого класса плюс часть смешанных высококлассных, исключая внутренние сборки в стиле DBG
Обслуживаемые линии	**10 500**	**16 000**	**23 500**	Произведение
Разовая цена на линию	10 тыс. RMB	20 тыс. RMB	30 тыс. RMB	Лестница цен и потолок (глава 4)
Разовый обслуживаемый рынок (установленная база)	**105 млн RMB**	**320 млн RMB**	**705 млн RMB**	Произведение
В год на окне ≈5 лет	21 млн RMB	64 млн RMB	141 млн RMB	Окно сжимается на 8–12% в год
Подписка (2.5 тыс. / 5 тыс. / 10 тыс. RMB на линию в год, полное внедрение)	26 млн RMB в год	80 млн RMB в год	235 млн RMB в год	Якорь годовой платы в стиле Baidu

FIG 23 **Снизу вверх: 40 000 линий → 16 000 обслуживаемых → ≈RMB 320M разово** ***Ниша сторонней AI-перепроверки в материковом Китае реальна, но мала.***



новичок за 3 года берёт 2–5% обслуживаемых линий: ≈320–800 линий, ≈RMB 2–5M в год
как подписка (RMB 5k на линию в год): RMB 1.6–4M рекуррентно, лишь 0.3–2.3% прогноза IDC по промышленному AI-контролю Китая

Источник: NVR (база 2020); оценки автора.

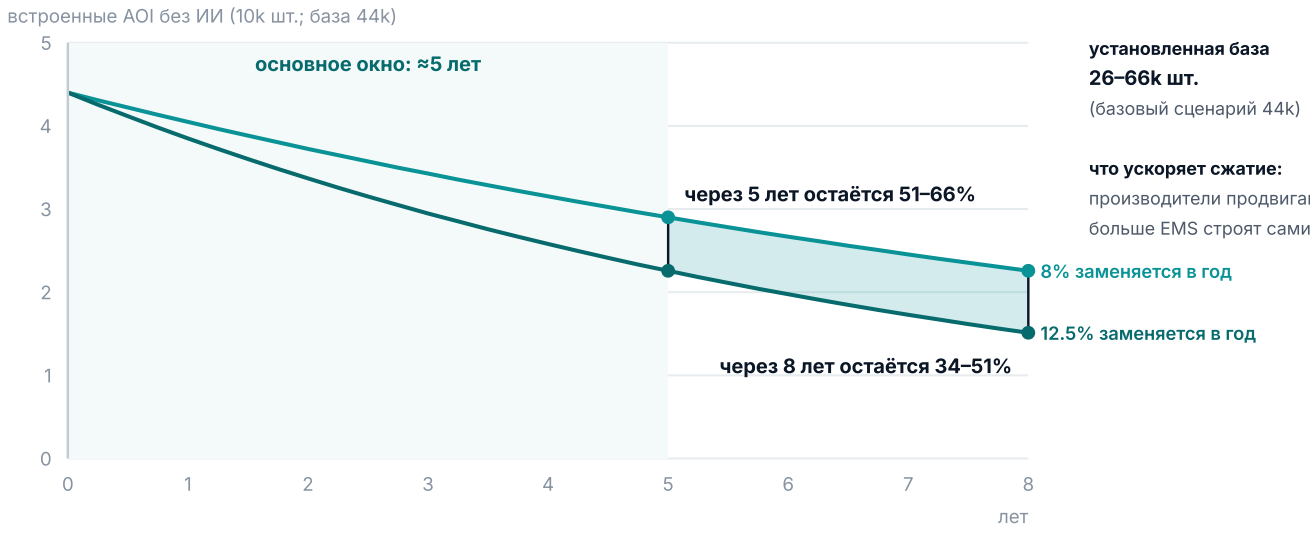
Всё это оценки; ни доля, ни цена публично не измерялись. Ниша стоит лишь 0.3–2.3% в год от прогноза IDC для всего промышленного AI-контроля качества Китая в 2025 (около 6.2 млрд RMB). Новичок, взявший 2–5% базовых обслуживаемых линий за три года, имел бы около 320–800 линий: 6.4–16 млн RMB суммарно при 20 000 RMB на линию, около 2–5 млн RMB в год, или 1.6–4 млн RMB в год рекуррентной выручки при 5 000 RMB на линию estimate. Этого достаточно для бизнес-единицы внутри компании, но не для венчурной отдельной компании. Каждая 1 000 установленных линий на Тайване, в Юго-Восточной Азии или США стоит примерно 3–15 млн RMB разово по материковым ценам; на Тайване и в США цены выше, но покупатели там чаще используют OEM 3D AOI с опциями ИИ estimate.

Окно модернизации — около пяти лет

Ни один производитель AOI не публикует срок службы машин; EMS-компании амортизируют оборудование за 3–10 лет (DBG 2025 annual report), а старые AOI теряют обновления ПО и запчасти после окончания срока службы (I-Connect007) single source. Я использую 8–12 лет (база 10).

При 1.0–1.65 встроенных AOI на линию в материковом Китае около 35 000–78 000 встроенных AOI в эксплуатации (база около 55 000), согласуется с известными поставками (Circuits Assembly, Sinic-Tek 2025 annual report). Ни один источник не измеряет, сколько работает с AI-перепроверкой; предполагая, что отечественные машины предлагают её примерно с 2023 и 40–60% недавних поставок используют её, **около 26 000–66 000 встроенных AOI в материковом Китае (база около 44 000) не работают с ИИ** estimate. Если каждая новая машина теперь поставляется с ИИ, этот пул сжимается на 8–12.5% в год: через пять лет остаётся 51–66%, а через восемь — 34–51%, быстрее, если производители AOI продвигают AI-апгрейды новым машинам. **Большая часть возможности приходится на ближайшие пять лет или около того**, с ядром в средних EMS со старыми отечественными 2D AOI, смешанными брендами и без пути AI-апгрейда.

FIG 24 **Окно модернизации — около пяти лет** ***Встроенные AOI без ИИ сжимаются на 8–12.5% в год по мере замены или апгрейда.***



Источник: Оценки автора; установленная база и доли не проверялись независимо.

Конкуренты уже сильны: клин перепроверки окружён

Клин AI-перепроверки сужается с трёх сторон. Лидеры AOI и зрения встраивают ИИ в новые платформы: Jutze «полностью приняла алгоритмы ИИ» (Jutze investor relations, June 2026), выручка OPT от промышленного ИИ выросла на 96.77% в 2025 (OPT 2025 annual report) single source, primary, а ViTrox заявляет до 95% точности автономной перепроверки (ViTrox) vendor claim. Крупные производители строят свои (iThome). А венчурные AI-компании продают в масштабе: клиенты периферийных сенсоров SmartMore выросли с 65 до 251 за два года (36Kr) single source, Hikvision продаёт свою модель Guanlan для инспекции (Xinhua), а интеграция «глаз, мозг и рука» уже предлагается SmartMore, Hikrobot и Mech-Mind (36Kr, Securities Star).

FIG 25 **Клин перепроверки окружён с трёх сторон** ***Пространство для третьих сторон — в средних EMS и в нишах инспекции, слабо покрытых AOI.***



Источник: Годовые отчёты Jutze, Sinic-Tek и DBG; публичные материалы Hikvision и SmartMore; сводка автора.

Пространство, оставшееся новичку, узкое: средние EMS со старыми 2D/3D AOI без AI-апгрейда от производителя; линии, смешивающие несколько брендов AOI (один вендор модернизации уже поддерживает Omron, VI, Koh Young и Mirtec (Sohu on Hengtianyi)); ниши, которые SMT AOI покрывает слабо, такие как сквозные и ручные паяные соединения, конформные покрытия, кабели, разъёмы и FPC, финальный внешний вид и винты; и малые заводы, которые не могут позволить себе AOI за 230–280 тыс. RMB.

Ближайший листингованный аналог даёт предупреждение. Weiyi Zhizao, продающая оборудование AI-инспекции и роботов, увидела падение выручки 3C на 79.9% год к году в январе–мае 2026, признала «усиливающуюся конкуренцию и ценовое давление», отложила листинг в Гонконге 25 октября 2026 и раскрыла 15 проектов, прекращённых до завершения, 12 из них — проекты воплощённых роботов (Weiyi Zhizao prospectus, postponement notice) single source, primary. AInnovation завершила 2025 с 29.8% дебиторки под обесценение (AInnovation 2025 annual report). Клин инспекции 3C переполнен, а ценовая сила на линию слаба.

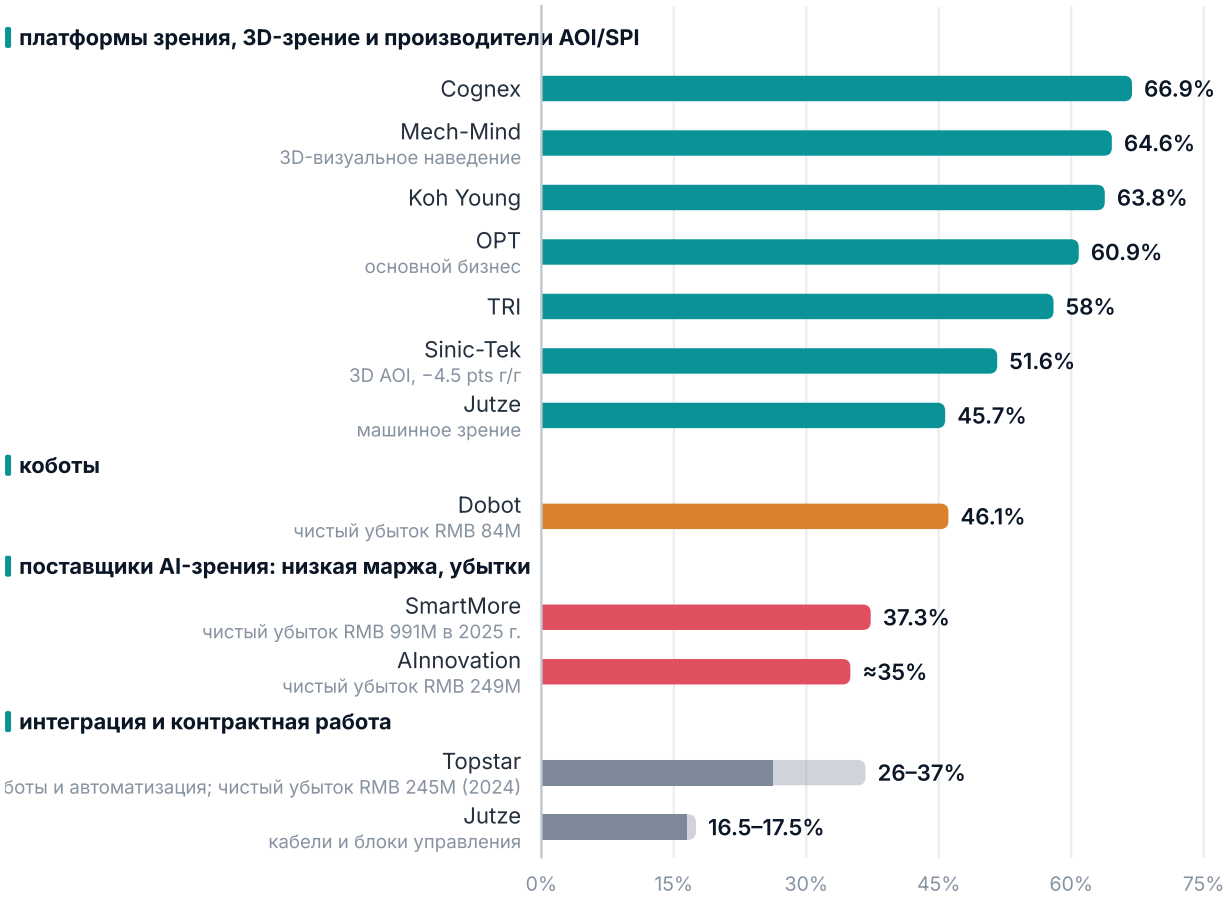
Маржа — в продуктном железе

Позиция в цепочке	Типичные компании и валовая маржа 2025	Прибыль и поступления
Платформы зрения	Cognex 66.9%	Прибыльна; выводит ≈$22 млн низкомаржинальной выручки
Оборудование AOI/SPI	Koh Young 63.8%; TRI 58%; основной бизнес OPT 60.85%; Sinic-Tek 3D AOI 51.62% (−4.53 пункта г/г); Jutze машинное зрение 45.74%	Прибыльно; дни дебиторки 139 (Jutze) до 247–274 (OPT)
Наведение по 3D-зрению	Mech-Mind 64.6%	—
Вендоры AI-зрения	SmartMore 37.3%; AInnovation ≈35%	Убыточны: чистый убыток SmartMore 991 млн RMB в 2025, AInnovation 249 млн RMB
Коботы	Dobot 46.1%	Чистый убыток Dobot 84 млн RMB; выручка UR с $326 млн (2022) до ≈$232 млн (2025, прогноз)
Системная интеграция	Роботы и автоматизация Topstar 26.2–36.7%	Чистый убыток 245 млн RMB (2024); дни дебиторки 180–240
Контрактное оборудование	Кабели и блоки управления Jutze 16.5–17.5%	—

Источники: (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 **Маржа — в продуктовом железе** ***Платформы зрения и производители оборудования зарабатывают 46–67% валовой маржи; AI-вендоры 35–37% и в убытке.***

валовая маржа 2025 г.



Источник: Годовые отчёты или проспекты 2025 г.

Новичок сохраняет маржу, только превращая инспекцию в продукт, оборудование или ПО; настройка проект за проектом толкает маржу к примерно 35% с постоянными убытками. Механика тоже работает против малых компаний: выручка признаётся только после приёмки на месте, авансы малы (Jutze, OPT), пилоты обычно бесплатны (A3/Mindtrace), а владельцы брендов часто задают оборудование инспекции на линиях своих контрактных производителей.

Как платят покупатели: 30% при подписании, 60% после приёмки, 10% удерживаются год

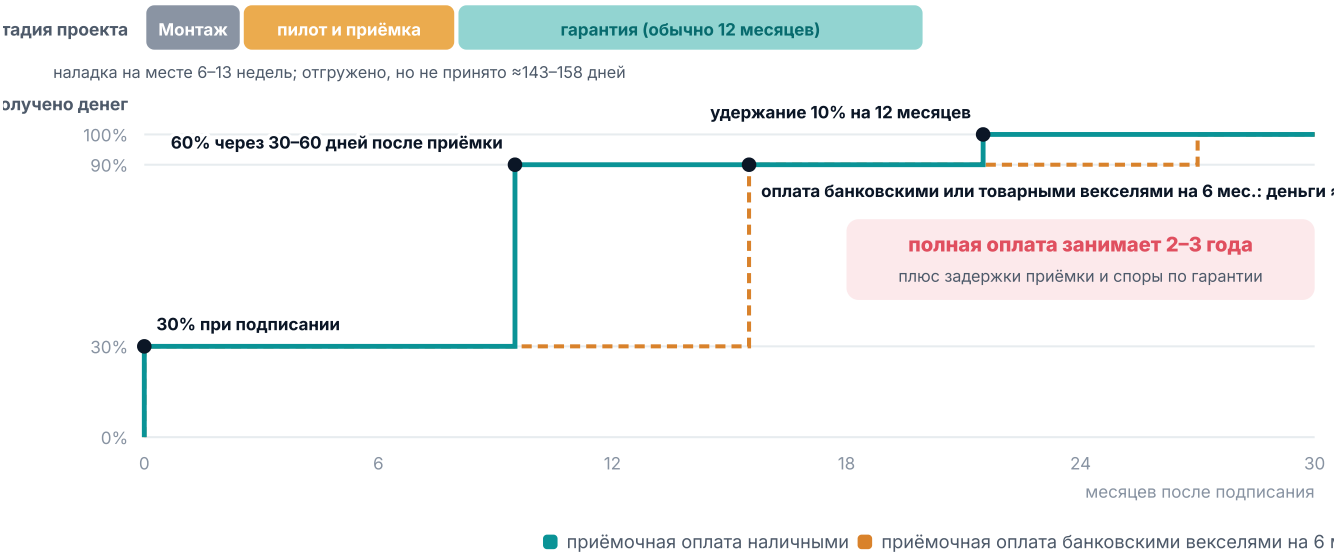
Условия оплаты, которые материковые заводы электроники используют для оборудования инспекции, прописаны в 13 именованных контрактах покупателей, раскрытых в ответе Sinic-Tek на запрос о листинге (Sinic-Tek inquiry reply 8-1-1) single source: один первичный документ, несколько независимых покупателей, 2020–2022 :

Покупатель (год)	Подписание	Доставка	Приёмка	Удержание или остаток
Подразделение Hikvision (2021)	30% в течение 15 дней	—	65% в течение 30 дней после финальной приёмки	5% гарантийное удержание, выплачивается через 12 месяцев после приёмки; часть — банковскими акцептами на 6+ месяцев
Nanchang Huaqin (2020–2021)	30% в течение 10 рабочих дней	—	60% через 60 дней после приёмки	10% после годовой гарантии

Покупатель (год)	Подписание	Доставка	Приёмка	Удержание или остаток
Guangdong Yinuo (2021, 3D AOI)	30% в течение 10 рабочих дней	—	60% в течение 30 дней	10% после трёх месяцев безотказной работы
Xiamen Hongxin, Jingmen Hongyi (2022)	30% в течение 10 рабочих дней	—	60% в течение 30 рабочих дней после финальной приёмки, банковские векселя на 6 месяцев	10% после годовой гарантии
AKM Panyu, Suzhou (2021–2022)	30% в течение 7–10 рабочих дней	—	60% через месяц после приёмки и выставления счёта, банковские векселя на 180 дней	10% через 360 дней
BYD Precision, Shenzhen, Huizhou (2020)	—	30% по прибытии	70% после приёмки и выставления счёта	Всё товарными векселями на 6 месяцев
Shenzhen H&T (2020)	—	—	90% после приёмки и выставления счёта	10% после шести месяцев безотказной работы

Другие первичные документы показывают ту же структуру: крупные проекты Mech-Mind платят 30% при подписании, 30% до доставки, 30% после приёмки и 10% гарантийного удержания на 12–15 месяцев (Mech-Mind prospectus); GKG говорит, что удержание в отрасли оборудования обычно 5–10% контракта, выплачиваемое через год после приёмки (GKG inquiry reply) single sources, primary. Норма, таким образом, около 30% при подписании (20–30%, ноль на некоторых заводах, таких как BYD), 60–65% после приёмки (выплачивается через 30–60 дней, часто шестимесячными банковскими или товарными векселями, что задерживает пригодные деньги ещё на полгода) и 5–10% удержания на 3–12 месяцев (12 наиболее распространено) verified по нескольким контрактам.

FIG 27 **Как платят покупатели *30% при подписании, 60% после приёмки, 10% удерживаются год: полная оплата занимает 2–3 года.***



Источник: Ответ Sinic-Tek на запрос, Mech-Mind, GKG, JT и другие отчёты; схема по обычным условиям, сроки оценены.

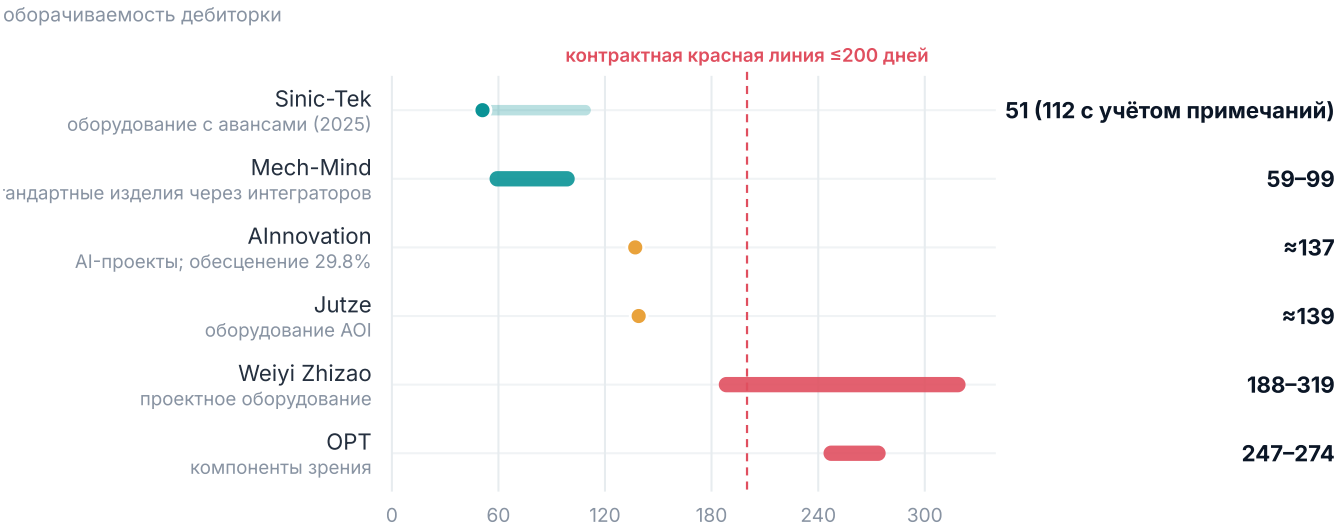
Приёмка и испытания медленны и оплачиваются продавцом. Отгруженные, но не принятые товары Sinic-Tek оборачивались около 143–158 дней, потому что крупные EMS и бренды требуют интеграционного тестирования всей линии и продлённого испытательного прогона (Sinic-Tek inquiry reply 8-1-2); AI-проекты Weiyi зани-

мают 4–8 месяцев от оценки до развёртывания (Weiyi Zhizao prospectus) single sources, primary. Данных о конверсии пилотов AI-перепроверки нет; Sinic-Tek конвертировала лишь около 18–19% своих демо-машин напрямую в продажи. Я планирую один из трёх–пяти для бесплатных пилотов и около 80% для платных пилотов после шорт-листа estimate.

Компания	Бизнес-модель	Дни оборота дебиторки
Sinic-Tek (2025)	Оборудование с авансами	≈51 (≈112 с учётом векселей)
Mech-Mind	Стандартные продукты через интеграторов	59–99
Jutze	Оборудование AOI	≈139
AInnovation (2025)	AI-проекты	≈137 на конец года, с обесценением 29.8%
Weiyi Zhizao	Проектное оборудование	188–319
OPT	Компоненты зрения	247–274

Источники: годовые отчёты и проспекты выше; дебиторка (включая контрактные активы) ÷ выручка × 365 primary + estimate.

FIG 28 **Оборачиваемость дебиторской задолженности** ***Стандартные изделия 51–99 дней; проектное оборудование 188–319 дней.***



Источник: Отчётность компаний; дебиторка (вкл. контрактные активы) ÷ выручка × 365.

Лицензии AI-перепроверки, продаваемые через производителей AOI или интеграторов, должны собираться как у Mech-Mind за 60–100 дней; прямые проектные продажи — как у Weiyi за 190–320 дней. Собранная воедино как денежная хронология: бесплатный пилот один–три месяца, квалификация у крупного ODM до примерно двенадцати месяцев, сборка, доставка и приёмка три–семь месяцев; 30% приходят с заказом, 60% примерно через четыре–девять месяцев после него и последние 10% примерно через 16–21 месяц. **От первого пилота у крупного EMS до полной оплаты может пройти два–три года**; средние EMS без формальных тендеров могут сократить квалификацию до недель estimate. Критерии приёмки в раскрытых контрактах покрывают только интеграцию всей линии, пропускную способность (один тайваньский проект требовал 100 единиц в час) и месяцы безотказной работы; ни один не задаёт уровни пропусков или ложных срабатываний (Anda inquiry reply) single source, primary. Письменные критерии приёмки AI-перепроверки надо согласовать до пилота.

Четыре паттерна неудач

Документированные неудачи делятся на четыре паттерна: отгрузка до надёжности продукта (Rethink (The Robot Report)); слой ПО поверх оборудования лидера, который экосистема лидера позже заменила (READY Robotics (The Robot Report)); нехватка денег на дне цикла (Dextrous Robotics (Robots & Startups)); и модельные компании, заканчивающие acqui-hire (Covariant (GeekWire)). Уроки: продуктционируйте узко, а не настраивайте проект за проектом; держите оборотный капитал для выручки на основе приёмки и долгих поступлений; предоставляйте локальный сервис для заводов, работающих 24 часа; и не масштабируйтесь до доказательства продукта.

Тайминг: инспекция ничего не выигрывает от ожидания; оборудование манипуляторов дешевеет

Аргументы за ожидание слабы. Оборудование не подешевело в 2026: производители роботов подняли цены на 5–15%, некоторые продукты на 25–50% (Sina Finance) single source , а наборы NVIDIA выросли в третий раз (36Kr). Производственные VLA не придут в пределах окна (IFR, Instrumental). Тем временем каждая замена AOI убирает цель модернизации, а каждое развёртывание OEM-ИИ запирает данные и процесс линии; возможности моделей коммодитизируются, а дефицитны доступ к заводам и размеченные заводские данные. У локальной поддержки тоже есть окна: правила Шэньчжэня 2026 «пионерский город ИИ», с до 2 млн RMB на демонстрационные проекты применения ИИ и ваучерами на модели, покрывающими 30% покупок, действуют до **30 июня 2027** (Yicai, May 2026) условия single source .

Что всё ещё быстро дешевеет — гуманоиды и ловкие руки (руки упали с примерно 50 000 RMB до менее 10 000 за два года (Sina Finance)) и начальные коботы по мере смещения продуктовой линейки вниз; это аргумент за отсрочку воплощённой части, а не части инспекции. **Итог: инспекция — сейчас, как узкий продукт; робоячейки — сейчас, через партнёров и интеграцию; воплощённый ИИ — позже, сегодня лишь одна R&D-станция.**

Воплощённый ИИ

06 Воплощённый ИИ: сначала он меняет то, как учат роботов, позже — как они двигаются

Что меняется — восприятие и программирование; движение остаётся детерминированным

Каждый лидер робототехники теперь продвигает «физический ИИ», но то, что отгружено, делится на три слоя: обученное восприятие без обучения (AI Masterless Vision от FANUC, отгружается в январе 2027 (FANUC); нейронный выбор захвата KUKA (KUKA)); большие модели, пишущие программы и планирующие задачи (iiQWorks.Copilot от KUKA, который предупреждает, что сгенерированный ИИ код следует проверять в цифровом двойнике (KUKA); Yaskawa с Gemini Robotics-ER, где «генеративный ИИ решает, что делать», а робот решает, «как двигаться» (Yaskawa)); и платформы вычислений и данных (AI Accelerator от UR с Jetson Orin (Universal Robots)) vendor claims. Ни один лидер не публикует производственные уровни успеха для своих AI-функций, а японские лидеры партнёрятся с Google, SoftBank, Microsoft и другими, а не строят фронтир-VLA (FANUC news index): слой обученных политик придёт от модельных компаний, а сертифицированные манипуляторы и контроллеры — от производителей манипуляторов.

FIG 29 **Стек воплощённого ИИ** ***Большие модели планируют офлайн, малые политики работают в ячейке, движение и безопасность остаются детерминированными.***

Слой	что делает и насколько далеко заходят доказательства	временной масштаб · кто владеет
Планирование L4	**фронтир-модели (офлайн)** писать или менять программы робота офлайн, переналадка, объяснение сбоев переписывает программы ABB RAPID на 77–99.7%, выдаёт после симуляции контур управления лишь 0.2–0.4 Hz: вне реального времени	**секунды–минуты** модельные гиганты
навыки / политика L3	**ACT, диффузионные политики, малые VLA** структурированные станции с оснасткой 90–99.4% краевые чипы: Jetson Thor, D-Robotics RDK, Ascend, Rockchip	**чанки действий 100–300 ms** чипмейкеры · производители
управление движе L2	**контроллер робота** траектории, серво, силовое управление, проверка столкновений детерминированный код; обученные политики лишь задают цели	**миллисекунды (реальному времени нужно >83 Hz)** производители роботов
Безопасность L1	**безопасный ПЛК, сканеры, ограничения силы и скорости** сертифицировано по ISO 10218-1/-2:2025 и местным нормам никакого ML в функциях безопасности: с 20 янв 2027 г. требуется оценка нотифицированного органа ЕС	**жёсткое реальное время** сертифицированный робот +

шанс малых фирм — в интеграции и сенсорном железе: встраивать обученное восприятие и малые политики в сертифицированные робо

Источник: Сводка автора (главы 3 и 6); ISO 10218:2025; регламент ЕС по машинам.

Насколько далеко до неструктурированной, гибкой, мелкосерийной работы: структурированные станции за 1–3 года, гибкая сборка — 5+

Реальные заводские цифры начали появляться, но они вендорские, узкие и со структурированных станций. Восемь гуманоидов AgiBot на линии финального тестирования планшетов Longcheer выполнили 64 828 задач

с 99.99% за более чем 64 часа (AgiBot, Yicai Global) vendor and customer, но это обслуживание, которое может делать обычный манипулятор со зрением. UBTech говорит, что её Walker S2 работает на 30–50% эффективности человека (Irish Times/FT) single source. 11 месяцев Figure 02 на BMW Spartanburg дали цифры объёма (1 250+ часов, 90 000+ деталей), тогда как её цикл, успех и число вмешательств — цели, а не результаты (Figure, BMW Group) verified, только числители. Nucleus говорит, что около 40% её заводского видео телеуправляется (Interesting Engineering) vendor claim, а расследование Reuters указывало около 20% выработки рабочего для некоторых китайских гуманоидов (Notebookcheck) single source. Ни одно развёртывание не публикует вместе уровень успеха, время цикла и частоту вмешательств, и ни одно не аудировано, а именно эти три и нужны для расчёта окупаемости.

Развёртывание	Уровень успеха	Такт	Вмешательства или телеуправление	Аптайм или интервал отказов	Объём прогона
Figure 02 @ BMW Spartanburg	Только цель (>99% за смену)	Только цель (84 с)	Только цель (0 за смену)	Не раскрыто (предплечье — главная точка отказа)	1 250 ч
Agility Digit @ GXO	Не раскрыто	Не раскрыто	Не раскрыто	Не раскрыто	100 000 контейнеров
Apptronik Apollo @ Mercedes-Benz	Не раскрыто	Не раскрыто	Не раскрыто	Не раскрыто	Не раскрыто
Agile ONE @ Audi	Не раскрыто	Не раскрыто	Не раскрыто	Не раскрыто	Не раскрыто
Nucleus (неназванный завод)	Не раскрыто	Не раскрыто	≈40% телеуправление (метод не раскрыт)	Не раскрыто	≈2 ч видео
AgiBot G2 @ Longcheer	99.99% из 64 828 задач	18–20 с на шаг	Не раскрыто	Один сбой связи, ≈2 мин на восстановление	64+ ч
Гибридная ячейка Neuromeka	322/324	На 12.8% медленнее человека	Не раскрыто	Не раскрыто	324 цикла
Physical Intelligence π0.6 на площадке заказчика	—	—	96.4% автономности за смену	Не раскрыто	Одна смена

Источники: эта глава и глава 3 в основном vendor claims.

FIG 30 **Что на самом деле раскрывают внутризаводские развёртывания роботов** ***Большинство раскрывает только объём; никто не раскрывает вместе процент успеха, такт и частоту вмешательств.***

	процент успеха	такт	вмешательства / телеуправление	доступность / сбои	серийный прогон
Figure 02 @ BMW	только цель >99%/смена	только цель 84 s	только цель 0/смена	н/д	1 250 h
Agility Digit @ GXO	н/д	н/д	н/д	н/д	100 000 контейнеров
Apptronik Apollo @ Mercedes	н/д	н/д	н/д	н/д	н/д
Agile ONE @ Audi	н/д	н/д	н/д	н/д	н/д
Nucleus (неназванный завод)	н/д	н/д	≈40% телеуправление*	н/д	≈2 h видео
AgiBot G2 @ Longcheer	99.99%	18–20 s на операцию	н/д	один сбой, 2 min	64+ h
гибридная ячейка Neuromeka	322/324	на 12.8% медленнее	н/д	н/д	324 цикла
π0.6 @ площадка заказчика	—	—	96.4% автономности	н/д	1 смена

* методика не раскрыта

раскрыто (в основном заявления вендоров) только цели н/д

Источник: Публичные материалы BMW, Figure, Agility, Mercedes-Benz, Audi, AgiBot, Neurometka, Physical Intelligence и др.

Телеуправление значит для экономики больше, чем уровень успеха. Чистое замещение людей на робота за смену ≈ (цикл человека ÷ цикл робота) × аптайм × успех, минус время вмешательств и телеуправления estimate, моя формула. Структурированная станция в стиле Neurometka (на 12.8% медленнее, 99.4% успех, 95% аптайм, 0.5 вмешательства по 2 минуты в час) замещает около 0.82 человека; гуманоид на 40% скорости человека, 90% аптайме и двух 3-минутных вмешательствах в час — около 0.26; при 20% Reuters — около 0.13; а если, как у Nucleus, один оператор телеуправляет им 40% времени — около 0.05, почти ничего. При примерно 600 тыс. RMB за гуманоида и 85 тыс. RMB за человека в год в материковом Китае последние три окупаются десятилетиями; только структурированная станция манипулятора имеет реалистичную окупаемость (глава 4) estimate.

FIG 31 **Чистое замещение труда на робота** ***Структурированная ячейка манипулятора замещает около 0.82 работника; гуманоид под телеуправлением 40% времени — около 0.05.***



Источник: Оценки автора по формулам.

Вне структурированных станций универсальные модели всё ещё набирают 40–90% на новых задачах: Gemini Robotics 2 достигает 74.2% на лабораторном pick-and-place, 78.9% на комплектовании инструментов и 44% на завязывании мешка (eesel/DeepMind); на RoboChallenge Table30, 30 задачах на реальном роботе, π0.5 набирает 43.7%, а лучшая открытая модель 50.3% (RoboChallenge, QbitAI).

Внешние прогнозы расходятся, но ни один не поддерживает широкое применение в течение двух лет. IFR говорит, что универсальный физический ИИ «ещё не индустриального уровня» и большинство гуманоидов «всё ещё прототипы или пилоты» (IFR); Interact Analysis оценивает, что лишь около 10% гуманоидов, произведённых в 2025, пошли в реальное использование, с коммерческим поворотным моментом «после 2032» (automation.com); UBS считает «EV-момент» гуманоидов маловероятным до 2030 (Yicai Global); китайские инсайдеры оценивают ещё два–три года до заводского масштаба (Gasgoo) single sources. Моё суждение estimate: на хорошо обеспеченных заводах электроники и авто с вендором на месте **структурированные станции** (обслуживание, сортировка OK/NG, закреплённая вставка) достигнут ≥99% на близкой к человеческой скорости за примерно 1–3 года; **комплектование смешанных деталей** с универсальными VLA — за примерно 3–5; **свободно прокладываемые жгуты и многошаговая гибкая сборка** — за более чем 5. Окупаемость гуманоидов сегодня не работает: Walker S2 при примерно 600 тыс. RMB (10jqka) single source, на 30–50% эффективности человека в две смены, равен примерно 0.6–1.0 рабочих, простая окупаемость 7–12 лет при материковых затратах на труд до интеграции, обслуживания и простоев estimate.

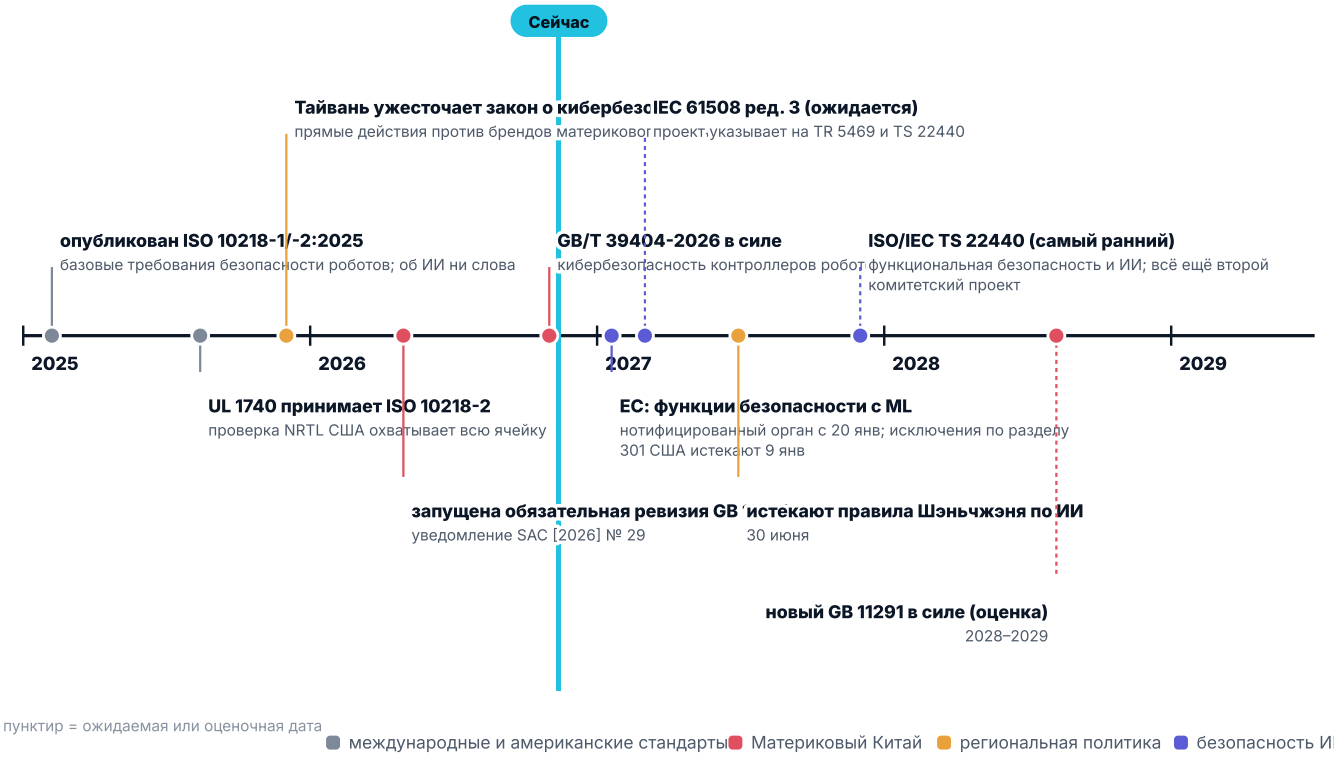
Где это застряло

Звено	Доказательства в 2026
Данные	Отрасль оценивает, что нужно не менее 10 млн часов качественных реальных данных, а к началу 2026 собрано лишь около 500 000 (Gasgoo); Reuters сообщает о около 500 000 часов против около 100 млн нужных (Notebookcheck); новичку-телеоператору нужно около 300 попыток на пригодное движение; заводы не делятся данными процесса; политики, обученные в симуляции, резко падают на реальных роботах
Надёжность	Структурированные станции 98–99.99% (vendor claims, отказ не определён); универсальные модели 40–90% на новых задачах; незнакомые разъёмы 29–67%
Скорость	Гуманоиды на 30–50% скорости человека (UBTech), некоторые около 20% (Reuters); фронтир-агенты около 280 с на квазистатическую задачу на реальном роботе; контуры управления LLM на 0.2–0.4 Гц против необходимых около 83 Гц (Anthropic)
Стоимость	Walker S2 в среднем около 600 тыс. RMB; окупаемость 7–12 лет estimate
Сертификация безопасности	ISO 10218-1/-2:2025 делает функциональную безопасность более явной и вбирает требования совместной работы, но ничего не говорит об ИИ (The Robot Report); ISO/IEC TS 22440 (функциональная безопасность и ИИ) всё ещё во втором комитетском проекте в ноябре 2026 (ISS, ISO Open Data); руководство IEC 61508-3 советует против ИИ выше SIL 1 (Analog Devices); по регламенту ЕС по машинам функции безопасности с машинным обучением требуют оценки нотифицированным органом с 20 января 2027 (EUR-Lex)
Руки и инструменты	Ловкие руки дают около 30 Н (≈3 кг) на палец против около 10 кг у человеческой руки (Gasgoo); смена инструмента всё ещё ручная
Интеграция	Ошибочные показания датчиков, несовпадения протоколов и связи MES на месте; частоты вмешательств почти никогда не раскрываются

Прямое следствие для безопасности: стандарты ни запрещают обученные политики, ни признают их. ISO 10218-1/-2:2025 ничего не говорит об ИИ; ISO/IEC TS 22440 (функциональная безопасность и ИИ) всё ещё в комитетском проекте и вряд ли выйдет раньше конца 2027 или 2028 estimate; ISO/IEC TR 5469:2024 — только руководство, рекомендующее не-ИИ фолбэки, надзорные ограничения и избыточное голосование; IEC 61508 редакция 3 ожидается в начале 2027 (ISO Open Data, BSI, 61508 Association) verified. Ни один опубликованный стандарт пока не предлагает принятого пути сертификации машинно-обучаемого восприятия как функции безопасности. Региональные правила тоже движутся: Китай пересматривает GB 11291 (в силе около 2028–2029) (revision plan 20262543); UL 1740 ввёл ISO 10218-2 в проверку NRTL США с августа 2025; ЕС требует оценки нотифицированным органом функций безопасности с машинным обучением с 20 января 2027; и ни один гуманоид не сертифицирован для безогражденной заводской работы, пока ISO 25785-1 ещё в проекте (i-SCOOP). Защитимая архитектура: машинное обучение принимает только решения о процессе («какая деталь», «хороша ли эта плата»), а сертифицированный детерминированный слой безопасности (обычно PL d / Category 3) и безопасный ПЛК решают скорость, силу и остановки. Относитесь к любому утверждению «ка-

мерный ИИ для безогражденной безопасности» как к недоказанному, пока сертификат PL d или SIL 2 не назовёт компонент машинного обучения.

FIG 32 **Хронология стандартов и регулирования** ***Путь сертификации ИИ внутри функций безопасности не устоится до 2027–2029 гг.***



Источник: ISO, IEC, UL, EUR-Lex, SAC и национальные базы регулирования.

Какой слой — возможность: модели принадлежат гигантам, периферийные боксы — чипмейкерам, малым фирмам — интеграция и сенсорное оборудование

Слой моделей требует капитала и данных и принадлежит гигантам (AgiBot обучила GO-2 на «десятках тысяч часов» (iFeng Tech); сильнейшие модели закрыты или только в облаке). Слой периферийных боксов принадлежит чипмейкерам: Jetson Thor — стандарт де-факто, а в материковом Китае конкурируют платы RDK от D-Robotics (D-Robotics) и сопроцессоры Rockchip (CSDN), в основном без опубликованных результатов VLA.

У малых фирм два проёма. Один — сенсорное оборудование, которое напрямую покупает надёжность: касание поднимает успех вставки ContactDP с 58.7% до 97.8%, а датчики силы/момента и камеры на запястье делают похожее; наборы для сбора данных тоже продаются (10jqka), хотя руки, датчики и манипуляторы все падают в цене. Другой — слой интеграции между производителями манипуляторов и модельными компаниями, владеющий данными задач каждой станции, конвертом безопасности и связью MES, который производители манипуляторов сегодня не продают как продукт. Конкуренты делятся на четыре группы: лидеры манипуляторов (FANUC, Yaskawa, KUKA, ABB, UR, JAKA, Dobot и другие), фирмы «зрение плюс манипулятор» (MechMind, Hikrobot, SmartMore), компании воплощённого ИИ (AgiBot, UBTech, Unitree, Galbot, Figure, Physical Intelligence, Google DeepMind, NVIDIA) и новички из автомобильных чипов.

Приёмка покупателем: нет числа без знаменателя

Поскольку ни один вендор не публикует частоты вмешательств, интервалы отказов или долгосрочные подсчёты успехов, они принадлежат условиям приёмки контракта. Статистически 3 000 последовательных производственных циклов с нулём отказов доказывают уровень отказов не более 0.1% при 95% уверенности; 10

отказов в 3 000 доказывают лишь ≤0.57%; доказательство ≤1%, ≤0.5%, ≤0.1% и ≤0.01% с нулём отказов занимает 299, 598, 2 995 и 29 956 циклов estimate, точный биномиальный. По этому стандарту 322/324 Neuromeka имеет двусторонний 95% интервал 97.8–99.9% и не доказывает ≥99%; примерно 6 отказов AgiBot в 64 828 задачах подразумевают 99.98–100%, но только если «отказ» определён так, как его определяет покупатель estimate. Это требования приёмки, которые я бы предъявил любой ячейке манипулятора или воплощённой ячейке; пороги опираются на собственные цели Figure у BMW и на кейсы Neuromeka и AgiBot, и это суждение, а не стандарт estimate:

Пункт	Требовать	Предлагаемый порог прохождения
Определения	Подписанные определения цикла, успеха, отказа (частичный успех, повторный захват, падение, пропуск качества), вмешательства (касание, удалённое телеуправление, сброс, e-stop) и того, как вычисляется автономность	Определения идут до любой метрики; нет процента без формулы
Вмешательства и телеуправление	На робото-час, по типу и длительности; среднее время между вмешательствами; минуты телеуправления как доля времени работы; операторов на робота	Структурированные станции ≤0.5 вмешательств на робото-час (≥2 ч между ними); нулевое телеуправление при приёмке
Долгосрочный успех	≥3 000 последовательных производственных циклов за ≥3 смены, ≥2 номера деталей и нормальное освещение и вариации подачи; отказы классифицированы; точная нижняя граница	Нижняя граница ≥99.5% для встроенных станций, останавливающих линию, ≥99.0% офлайн; пропуски качества сообщаются отдельно
Время цикла	P50/P90/P99 на деталь включая восстановление, против ≥30 засечённых человеческих циклов на той же станции	≤1.15 × человеческий P50 или в пределах такта линии
Аптайм и отказы	Часы работы, журнал простоев, отказы по компонентам, время замены деталей	≥95% аптайм за ≥2 недели; MTTR ≤20 мин с запчастями на месте
Восстановление	Скриптовая инъекция отказов: падения, окклюзия, смена освещения, остановка конвейера, потеря MES или сети, потеря питания, e-stop	≥90% автономного восстановления, возврат в производство за 2 мин
Переналадка	Часы, демонстрации, навык оператора и стоимость сбора данных на новый номер детали или приспособление	Структурированные станции: ≤1 смена и ≤100 демо на новый номер детали
Безопасность	Манипулятор и ячейка по ISO 10218-1/-2:2025, оценка риска ячейки, локальные правила (GB 11291 в материковом Китае; CNS 14490 плюс письменная оценка на Тайване; ANSI/A3 R15.06-2025 и UL 1740 в США)	Ни одна функция безопасности не реализована машинным обучением; сертифицированный слой безопасности (обычно PL d / Cat 3) ограничивает скорость и силу; никаких безогражденных шагающих роботов до публикации и оценки ISO 25785-1
Вычисления и офлайн-поведение	Оборудование инференса, задержка P99, поведение без сети, версионирование моделей, повторная проверка после каждого обновления	Такт-критический инференс не зависит от облака; полная повторная приёмка после любой смены модели, квантования или движка
Доказательства	Немонтированное видео и сырые журналы за всё окно приёмки; засвидетельствовано командой качества заказчика или третьей стороной	Никаких нарезок или резюме только от вендора

Пункт	Требовать	Предлагаемый порог прохождения
Коммерческие условия	Вехи оплаты, привязанные к метрикам приёмки; штрафы за доступность при оплате за робото-час; правила владения данными и удалённого доступа	Финальный платёж и удержание выпускаются только после приёмки 3 000 циклов и ≥4 недель стабильной работы
Экономика	Чистое замещение людей = отношение скоростей × аптайм × успех − время вмешательств − доля телеуправления	Отклонять любое предложение, чьё чистое замещение не поддерживает заявленную окупаемость

Источники: (Figure, Neurometka, Yicai Global, i-SCOOP); пороги — мои предложения.

Открытые вопросы

07 На что публичные источники не дают ответа и как это решает пилот

Вопросы, на которые публичные источники не могут ответить, делятся на пять групп: за сколько на самом деле продаётся AI-перепроверка и сколько готовы платить покупатели; как производители AOI бандлят ИИ и апгрейдят старые машины; насколько стары установленные AOI и сколько уже работает с ИИ; сколько людей и часов на самом деле занимает перепроверка на линию; и как конвертируются пилоты и как платят покупатели. Поверх этого — как новейшие модели, квантованные сборки и манипуляторы на самом деле работают в условиях конкретного завода. Каждый можно превратить в тест с чётким порогом. Пороги ниже — правила решений, выведенные мной из доказательств этого отчёта estimate; пересчитайте их, когда появятся первые заводские данные.

Открытый вопрос	Как его решить	Как выглядит хороший результат
Цена и готовность платить за AI-перепроверку	Стандартный запрос котировки (например четыре линии со смешанными старыми 2D и новыми 3D AOI, две смены, около 2 000 плат за смену, 12 классов дефектов, только локально, интеграция MES или IPC-CFX), отправленный нескольким сторонним вендорам; опрос ценовой чувствительности из четырёх вопросов для 20–30 заводов EMS	Большинство котировок в пределах 40–120 тыс. RMB разово или 20–40 тыс. в год за четыре линии; как минимум половина заводов принимает 10–30 тыс. RMB на линию с окупаемостью в пределах 12 месяцев
AI-бандлинг и апгрейды производителей AOI для старых машин	Короткие структурированные интервью с примерно дюжиной производителей AOI и дистрибьюторов	Является ли AI-перепроверка стандартной, опциональной или отсутствует на моделях 2025–26; можно ли апгрейдить машины до 2020 года и по какой цене
Возраст установленных AOI и доля работающих с ИИ	Обследование на месте около 30 заводов EMS, стратифицированных по размеру и региону, со списками оборудования под NDA	Медианный возраст AOI ≥5 лет и ≥60% без ИИ означает, что пул модернизации реален
Труд и часы перепроверки	Интервью и 2–4 недели журналов срабатываний AOI, перепроверки, ремонта и смен на двух или более заводах, с выборкой работ	≥1 выделенный перепроверяющий на линию за смену и ≥3 часа перепроверки на линию за смену, с как минимум 0.5 FTE на линию, которые завод готов перераспределить
Производительность самой AI-перепроверки	Золотой набор плюс теневой режим на реальной линии (глава 3): соглашение о данных на неделе 0, базовый уровень и разметка на неделях 1–2, теневой прогон на неделях 2–4, опциональная ассистируемая работа на неделях 4–8	Ложные срабатывания или часы перепроверки сокращены на ≥50–60%; границы пропусков по классам не хуже человеческого базового уровня; ноль пропусков на критических классах; задержка p95 ≤½ такта; автоматический отказ открыто к людям
Качество меток учителя у фронтир-моделей	Прогнать двух или трёх кандидатов на том же золотом наборе, по три формулировки промпта, одно хорошее эталонное изображение, одиночные кропы компонентов ≥42–48 px	Автоматическое принятие только для классов с нулём пропусков на ≥299 дефектах и согласием на человеческом уровне; иначе только предразметка и сортировка

Открытый вопрос	Как его решить	Как выглядит хороший результат
Квантованные сборки	Меняйте только точность; закрепите хеши весов, шаблон, предобработку, декодирование, движок и размер пакета; отключите или сбросьте префиксный кэш; проверьте, что эталонная сборка воспроизводит результаты вендора; сравните пары на золотом наборе и срезе дрейфа	Нет новых пропусков на критических классах; ≤1 новый пропуск на 1 000 дефектов; снижение очистки ≤2 балла; ≥99.5% идентичных решений за три прогона; иначе откат к FP8/INT8
Оборудование у линии, чистота и ESD	≥2 единицы каждого бокса-кандидата под стресс-тестом при удвоенной ожидаемой частоте срабатываний; 72-часовой прогрев при реальном потоке воздуха шкафа; проверка ESD координатором заказчика; тест ISO 14644-14 для заказчиков ISO 7 и чище	p95 ≤½ такта; нет сбоев или троттлинга за 72 часа; подпись ESD; отчёт по частицам на нужном классе
Условия оплаты, правила пилота и конверсия	Пункты оплаты из недавних контрактов на оборудование инспекции у обследованных заводов; интервью о правилах пилота, критериях приёмки и времени квалификации	Около 30/60/10 с деньгами в пределах 60 дней после приёмки; пилоты ≤8–12 недель с названным владельцем подписи; ≥1 из 3 испытаний конвертируется; рамочное соглашение на ≥5–10 линий в первый год
Данные, ИТ и ответственность	Интервью и проверка контракта с юристом	Локальный бокс может подключаться и получать доступ к данным; ответственность ограничена дефектами ПО
Региональное соответствие перед выходом за рубеж	Тайвань: письменный статус заказчика по Закону об управлении кибербезопасностью, самооценка E187; США: таможенное решение о классификации ячейки, котировки NRTL, проверка ITAR; Юго-Восточная Азия: локальная привязка BOI и право на малайзийскую надбавку, материнская компания заказчика; материковый Китай: охват CCC для бандлированных ПК и источников питания	Заказчик не является контролируемым органом, а окупаемость остаётся ≤2 лет с учётом пошлины и сертификации
Приёмка робота или воплощённой ячейки	Таблица приёмки из 12 пунктов в главе 6: подписанные определения, ≥3 000 последовательных производственных циклов за ≥3 смены и ≥2 номера деталей, скриптовая инъекция отказов, ≥30 засечённых человеческих циклов, засвидетельствовано	Нижняя граница ≥99.5% встроенно (≥99.0% офлайн); цикл ≤1.15 × человеческий; ≤0.5 вмешательств в час и нулевое телеуправление; ≥95% аптайм; MTTR ≤20 мин; ≥90% восстановления
Политики ACT и VLA на китайских периферийных чипах	Четыре недели: портирование и микробенчмарк (≥1 000 инференсов, распределение задержки, мощность, 60-минутный прогрев); замкнутая симуляция с измеренной впрыснутой задержкой; A/B на реальном роботе против GPU; выносливость 8–24 часа (≥3 000 циклов)	Чанки ACT p95 ≤100 мс, VLA ≤300 мс; успех на реальном роботе в пределах 3 баллов от GPU и цикл ≤1.2 × человеческий; иначе держите политику на Jetson или GPU
Цены коботов, AMR и интеграции	Котировки по моделям от нескольких дистрибьюторов коботов и полные котировки ячеек от двух–трёх интеграторов для загрузки тестовых приспособлений (одна станция, две смены)	Полная ячейка ≤200 тыс. RMB и окупаемость ≤1.5 лет за две смены в Шэньчжэне
Подсчёты линий SMT на Тайване, в Юго-Восточной Азии и США	Данные IPC (ранее New Venture Research) или установленная база локальных дистрибьюторов оборудования	Достаточно обслуживаемых линий на регион, чтобы оправдать локальный прикладной инжиниринг

Из всех них четыре решают самые чувствительные цифры этого отчёта: цену, которую готовы платить покупатели, труд, который реально высвобождает AI-перепроверка, действительно ли AI-перепроверка пропускает не больше людей, и как быстро возвращаются деньги. Их стоит решить до любого значимого обязательства — заводу, рассматривающему покупку, или поставщику, рассматривающему продукт.

Последовательность пилота, работающая для обеих сторон

Для завода и для вендора, которому нужна заслуживающая доверия ссылка, порядок одинаков:

1. **Недели 0–2: согласуйте данные и определения.** Подпишите соглашение о данных, определите классы дефектов и что считается пропуском, и запишите базовый уровень: ложные срабатывания на плату, секунды перепроверки на вызов, пропуски, найденные ниже по потоку.
2. **Недели 1–4: постройте золотой набор** из архива AOI и сохранённых NG-плат, размеченный вслепую и арбитрированный (глава 3).
3. **Недели 2–4 (или дольше): теневой режим.** Модель записывает свои решения, пока люди решают как прежде; сравнивайте по классам.
4. **Недели 4–8: ассистируемая работа на случайных сменах,** с моделью, очищающей только классы, прошедшие ворота, и всем остальным к людям.
5. **Приёмка, письменно, до масштабирования:** результаты по классам, задержка, поведение отказа открыто и кто владеет моделями и метками. Масштабируйте линия за линией и перепроверяйте на каждом новом продукте, версии модели, изменении квантования или движка.

Ограничения этого отчёта

Это кабинетное исследование. Я не проводил первичных интервью с заводами для этого издания; многие данные линий приходят от вендоров и не были независимо аудированы. Несколько правительственных сайтов, баз тендеров и платформ найма оказались недоступны, поэтому ставки субсидий на модернизацию Шэньчжэня 2026, ориентиры зарплат в дельте Янцзы, транзакционные цены AI-перепроверки, перепроверяющие на линию и уровни конверсии пилотов — это оценки или аналогии; каждый появляется выше как открытый вопрос. Курсы валют — центральный курс Народного банка Китая от 9 ноября 2026 ($1 = 6.7188 RMB) и, для Тайваня, фиксинг SMBC того же дня ($1 = 32.12 NT$); вьетнамский донг использует данные мая 2026. Все пороги — правила решений, выведенные из процитированных доказательств, а не стандарты.

Об авторе

Tan Shuai

Tan Shuai — технолог и основатель с более чем двадцатью годами работы в области чипов, аппаратного и системного программного обеспечения и 20 выданными патентами. Его работа провела его через Кремниевую долину, Нью-Йорк, Чикаго, Тайвань, Большой залив (Шэньчжэнь и Гонконг), Вьетнам и Малайзию, с долгим фокусом на передовом производстве и цепочках поставок. Сегодня он работает исследователем и разработчиком в области ИИ, сосредоточившись на том, как ИИ на самом деле применяется в производственном цеху.

Он приветствует диалог с людьми, работающими над ИИ в производстве, — строите ли вы команду, ищете партнёра или сверяете заметки: i@tanshuai.com · tanshuai.com

Примечания и оговорка

Этот отчёт — независимое кабинетное исследование на основе публичных документов, доступных по 11 ноября 2026 года. Он предоставляется только для информации и не является инвестиционной, юридической, налоговой или инженерной консультацией. Названия продуктов, цены, релизы моделей, экспортные классификации, пошлины и нормативы быстро меняются; проверяйте всё, на что полагаетесь, по первичному источнику. Цифры, помеченные vendor claim или estimate, следует читать именно так. Названия компаний и продуктов принадлежат их владельцам и используются только для идентификации; ни одна компания не рецензировала и не спонсировала этот отчёт.

© 2026 Tan Shuai. Вы можете свободно распространять этот отчёт без изменений с указанием авторства.

ПРОИЗВОДСТВО ЭЛЕКТРОНИКИ · ИИ-ИНСПЕКЦИЯ × РОБОТОТЕХНИКА

Визуальный контроль на основе ИИ и роботы-манипуляторы на заводах электроники: оценка 2026 года

Каждая ключевая цифра снабжена источником и меткой уверенности: verified, single source, vendor claim или estimate. То, на что публичные источники не дают ответа, собрано в главе 7, с указанием того, как пилот закрывает каждый вопрос.

Tan Shuai
tanshuai.comi@tanshuai.com

ПУБЛИЧНАЯ ВЕРСИЯ · ДАННЫЕ НА 11 НОЯБРЯ 2026

