Harness стал commodity. Что осталось — и почему это другая компания

2026-04-25

Harness стал commodity. Что осталось — и почему это другая компания

Один квартал, четыре анонса

8 апреля 2026 Anthropic выпустил Managed Agents — managed execution loop, persistent memory через интерфейс `/memories`, sandboxing, multi-agent orchestration, агенты, которые «self-evaluate and iterate until they reach a result». В публичном API beta header — `managed-agents-2026-04-01` (Claude API docs). 22 апреля на Cloud Next Google переименовал Vertex AI в Gemini Enterprise Agent Platform — Agent Builder, Agents Gallery, A2A protocol, $750M партнёрский фонд. В тот же день AWS отгрузил managed agent harness preview в Bedrock AgentCore — Runtime, Gateway, Identity, Memory, Observability как managed primitive (AWS docs). OpenAI в это время добивал AgentKit — Agent Builder с визуальным canvas, Connector Registry, ChatKit, встроенные Evals и Guardrails (OpenAI).

За один квартал четыре крупнейших провайдера независимо схлопнули в managed-продукт тот самый слой, который ещё в марте называли «agentic moat» (businessengineer.ai). На рынке, который год назад продавал «настройку агента» за сотни тысяч рублей, теперь это `npm install` плюс конфиг в облачной консоли. Это не прогноз. Это то, что произошло.

Что сломалось у «AI-агентств» за один квартал

Бизнес-модель студии, продающей «AI-бота под клиента», держится на одной арифметике: harness (memory, tool routing, orchestration, evals, guardrails) — это инженерный труд, который можно перепродать с маржой. Когда Anthropic берёт на себя execution loop, OpenAI — Connector Registry, AWS — Memory и Observability, вся себестоимость, на которой строилась цена, испаряется. Не потому что заказчик стал умнее, а потому что у него теперь есть кнопка «Agent Builder» бесплатно или почти бесплатно.

Разговор в октябре 2025, который в развивающемся community вокруг agent stacks звучал как откровение — «модель — commodity, harness определяет успех агентов» — за полгода устарел. Harness теперь тоже commodity. И это меняет не одну строчку в P&L, а весь жанр: студия, которая продаёт «соберём вам агента», в 2026 — это студия, которая в 2018 продавала «настроим вам Kubernetes». Ещё работает, но потолок виден.

Производное следствие важнее самого факта. На воронке такой студии теперь стоит не «вам нужен бот?», а «вам нужен бот, или ваш CTO уже открыл AgentKit

и собрал его за вечер?». В обоих случаях продаваемая ценность — не в harness. Она где-то ещё. Вопрос только в том, понимает ли студия, где именно, до того как закроется следующий квартал.

Что именно стало commodity

Имеет смысл смотреть не на «агентов» вообще, а на компоненты. Tool routing и function calling — нативная часть API всех четырёх провайдеров; это больше не код, это конфиг. Memory short-term и long-term — managed интерфейс у Anthropic, AgentCore Memory у AWS, managed context у Google; кастомные RAG-пайплайны теряют ROI на глазах. Multi-step orchestration — Agent Builder у OpenAI, AgentCore Runtime у AWS, A2A protocol у Google (TNW). Evals и self-verification — OpenAI Evals с trace grading, AgentCore Observability с session recording и replay. Guardrails и PII detection — open-source у OpenAI, встроенные у Google, AgentCore Identity у AWS.

Стопка компонентов, которую год назад студия описывала клиенту как «наша архитектура», теперь почти полностью совпадает с прайс-листом Bedrock и таблицей фич AgentKit. Это и есть определение commoditization: универсальная форма, которую любой может купить за деньги, а не за время. Anthropic не случайно переименовал Claude Code SDK в Claude Agent SDK — это сигнал, что вся стопка теперь продукт, а не пример кода.

То, что **не** провалилось внутрь платформы, тоже понятно. Институциональное знание — то, как именно конкретная компания принимает решения, какие у неё edge cases, кто и когда эскалирует. Trajectory data — закрытый цикл «вход → решение агента → исход через N дней», который восстанавливается только через работу, не через промпт. Глубокая интеграция в окружение клиента — 1С, amoCRM, отраслевые API, legacy. Эти три слоя в Forbes-эссе Sanjay Srivastava названы non-absorbable не из стилистических соображений, а потому что их нельзя упаковать в SDK, не имея внутри клиента.

Три слоя, которые нельзя купить через npm

Названия здесь важны, потому что от их понимания зависит, чем компания будет заниматься следующие два года.

Representation layer. Это то, как бизнес моделируется в данных. Какие сущности первичны — лид, актив, контрагент, событие. Какие у них статусы. Что считается дублем. Что считается просрочкой. У большинства компаний это знание не лежит в одном месте: оно размазано между CRM, табличкой в Google Sheets, головой менеджера и чатом в WhatsApp. Когда компания строит AI-контур, она впервые вынужденно фиксирует representation как явный, структурированный объект — vertical schema. И этот объект, в отличие от модели, не commodity: его нельзя скачать с Hugging Face, потому что он точен ровно настолько, насколько точно описывает конкретную операционную реальность. Gartner прогнозирует, что через 2026 год 60% AI-проектов будут свёрнуты из-за недостаточного качества данных — именно потому что representation не

формализован. Жанр такой работы ближе к тому, что Foundation Capital в эссе про Service-as-Software называет «кодификацией экспертизы», чем к традиционной разработке.

Trajectory data. Закрытый цикл. Что агент сказал → что человек сделал → что произошло через час, день, неделю. Бенчмарки 2025–2026 (BEAM из UAlberta, MemoryAgentBench из UCSD, LongMemEval) показали, что 1M-context window не равно 1M-token памяти: модели деградируют именно на задачах разрешения противоречий, упорядочивания событий и обновления знания во времени. Это значит, что temporal trajectory — отдельная архитектурная задача, и open-source memory-стек (Graphiti, Letta, Cognee) её решает только частично: он даёт хранилище, но не данные. Данные появляются от того, что агент работает в реальной операции, а не на синтетике. Их нельзя восстановить задним числом без месяцев работы в том же контексте — это и есть структура switching cost.

Institutional SOP. Operating playbook компании, превращённый в исполняемую политику. Не «инструкция в Notion», а граф решений: триггеры follow-up, правила ценообразования, пороги эскалации, кто имеет право подписать скидку, какие исключения допустимы, какие нет. До AI это знание жило в людях. Когда оно становится частью контура агента, происходит смена носителя: SOP теперь не «как мы тут работаем», а артефакт, который компания владеет, версионирует и аудирует. Vendep в эссе про vertical workflows формулирует это резко: forget the data moat, the workflow is your fortress — и это не маркетинг, это описание того, что именно остаётся, когда модель и harness уходят вниз по стеку.

Эти три слоя не покупаются через `npm install`. Их строят. Медленно, по одному клиенту, по одной вертикали. И их главное свойство — они компаундируются: каждое следующее внедрение делает следующий точнее.

AI-native organization как сборка этих слоёв

Эти три слоя — не дополнительные фичи поверх агента. Они — форма самой компании, которая их использует. Именно про это говорит Andrej Karpathy, когда называет LLM «kernel of a new operating system» в своём докладе 2023 года «Intro to Large Language Models» — где «LLM OS» выделен как отдельная глава. Именно про это Jack Dorsey написал в акционерном письме Block в феврале 2026: «Intelligence tools have changed what it means to build and run a company» — и срезал 4 000 позиций (40% штата), обосновав это напрямую как операционную перестройку под AI. И это же — основная мысль Foundation Capital про переход от software-as-a-service к service-as-software: AI продаёт не инструмент, а готовую работу.

Пока этот паттерн собирают по частям, но не как целое. Описать его можно тремя сменами акцента.

Cron > project. В классической компании работа упакована в проекты — сущности с началом, концом, бюджетом и project manager. В AI-native компании

ключевая единица — повторяющийся цикл. Агент, который каждое утро в 9:00 проверяет состояние воронки и эскалирует то, что зависло. Агент, который раз в час пробегает по тикетам поддержки и помечает то, что требует человека. Cron-задачи — не вспомогательные скрипты, а основная операционная ткань. Проекты остаются для исключений; рутина живёт в расписании.

Memory > document. Документ оптимизирован под человека: его читают глазами, забывают, переписывают по поводу. У агента есть структурированная память — vertical schema плюс temporal knowledge graph — которая обновляется continuously, держит valid_at и invalid_at, разрешает противоречия не «последний прав», а с историей. Продукт-документация, инструкции для саппорта, описание процессов — всё, что в обычной компании живёт в Notion и устаревает за квартал, в AI-native компании живёт как структурированная память, которая обновляется в момент выполнения работы.

Role > seat. В классической компании единица найма — место. Junior support, middle sales, senior PM. В AI-native компании единица — роль с явным контрактом: что роль делает, какие у неё входы и выходы, какие escalation paths, какая память. Один человек может занимать несколько ролей. Часть ролей делают агенты. Часть — гибрид. Roadmap ролей похож не на org chart, а на список сервисов в кубернетесе: версии, зависимости, observability.

Это не теория. Klarna, по собственному заявлению, заместил порядка 40% штата AI-системами. Salesforce сократил 4000 позиций под тем же предлогом. Sam Altman публично сделал ставку на появление в течение года первой компании-юникорна с одним человеком в штате. Это можно считать пиаром, но направление одно. И направление состоит не в том, что «AI помогает работать», а в том, что операционная единица другая: cron, memory, role.

Связка трёх слоёв — representation, trajectory, SOP — собирает этот другой тип компании в нечто, что можно строить. Не как метафору, а как архитектуру. Harness стал commodity → выживает не тот, кто продаёт агентов, а тот, кто строит operating layer → AI-native organization есть тот operating layer. Эта связка объясняет, почему рынок «AI-агентств» в 2026 раздваивается: на тех, кто всё ещё продаёт harness (и закрывается), и тех, кто продаёт operating system для конкретной вертикали (и масштабируется).

Что это значит для трёх типов читателей

Фаундер AI-проекта. Если ваш текущий продукт — это «обёртка над API провайдера, упрощающая сборку агента», у вас 6–12 месяцев. Это не угроза, это просто календарь: managed harness уже идёт в GA у трёх из четырёх крупных провайдеров. Действие — не «придумать что-то ещё», а посмотреть, какой из трёх слоёв (representation, trajectory, SOP) у вашего продукта и ваших клиентов уже накапливается без вашего участия, и переинвестировать туда. Если ни один не накапливается — это сигнал, что вы строите generic harness, и пора пересобирать гипотезу. Конкретно: если вы не можете внятно описать, какой

data flywheel запускается у вашего клиента в первые 30 дней работы продукта, — у вас нет product-market fit, у вас есть демо.

Руководитель компании, думающий про AI. Главная ошибка 2026 года — покупать «AI-бота» как изолированный SaaS, отдельно от своих процессов. Это эквивалент покупки CRM в 2010-х в надежде, что она «улучшит продажи»: сама по себе не улучшит. Покупать стоит operating layer — поставщика, который интегрируется в ваш контур и помогает вам кодифицировать SOP, накопить trajectory data и зафиксировать representation. Тестовый вопрос на discovery-встрече: «через 6 месяцев работы с вами что у меня есть, чего не было?». Если ответ — «у вас работает бот», это harness-продавец. Если ответ — «у вас есть структурированный operating layer, который вы можете аудировать, версионировать и при необходимости перенести на другого провайдера», — это поставщик другого жанра. Заодно проверьте контракт: фиксирует ли он провайдера. Если да — это не moat, это lock-in, и при следующем шаге провайдеров вверх по стеку вы будете платить за это снова.

Инженер, выбирающий, где работать. Большая часть инженерных команд в AI-агентствах сейчас занята тем, что превращается в ETL-работу нового поколения: интеграция managed-компонентов друг с другом плюс прослойки, которые не выживут до 2027. Это не плохая работа — но она не компаундируется. Команды, где компаундируется опыт, — это те, кто работает в одном вертикале достаточно глубоко, чтобы накапливать representation и trajectory. Признак: команда не описывает свою работу в терминах «мы интегрируем GPT/Claude», а в терминах «мы строим operating layer для X». Если на собеседовании вам показывают архитектуру вокруг harness, а не вокруг domain — это команда, которая через год будет делать миграцию своего же стека на managed harness и думать, чем заняться. Если вокруг domain — у вас есть шанс проработать там пять лет и выйти с компетенцией, которой не будет ни у кого, кроме людей с этим же опытом.

На какие сигналы смотреть

Тренд имеет несколько проверяемых траекторий, по которым видно, ускоряется он или нет. Первая — появление managed agent harness у не-западных провайдеров. GigaChat и YandexGPT в РФ, Qwen и Doubao в Китае. Если в течение Q3–Q4 2026 в их продуктах появляется визуальный agent builder и managed memory — это сигнал, что harness-commoditization вышла на глобальный уровень и больше не зависит от geo. Вторая — появление vertical template marketplaces от провайдеров: «готовые SOP-presets под ритейл, страхование, логистику». Если такие маркетплейсы открываются, low-end сегмент vertical AI становится неустойчивым, и операционный layer как сервис должен сместиться вверх по mid-market. Третья, обратная — появление публичных кейсов компаний, которые ушли с managed harness обратно в собственный стек. Если такие кейсы будут — значит, либо managed-продукт упёрся в потолок применимости, либо exit-cost оказался выше, чем казалось при подключении. Это интересный сигнал для всех, кто сейчас выбирает между «строить» и «купить».

В любом сценарии главный вопрос остаётся одним и тем же. Не «какого агента построить», а «какой operating layer вокруг агентов мы собираем за следующие 24 месяца, и кому мы будем им владеть к 2028». Это вопрос о форме компании, не о технологии. Технология уже общая.

Главное

- Agent harness (память, orchestration, evals, guardrails) стал commodity: Anthropic, OpenAI, Google, AWS выпустили managed-продукты за один квартал.
- Защитимыми остаются три слоя: representation layer (вертикальная схема данных), trajectory data (закрытый цикл вход→решение→исход), institutional SOP (политика компании как исполняемый граф решений).
- AI-native organization — это не метафора, а архитектура: cron > project, memory > document, role > seat.
- Рынок AI-агентств раздваивается: те, кто продаёт harness — закрываются; те, кто строит operating layer для вертикали — масштабируются.