Setup умирает к 2027. Retainer - не весь. Карта того, что уцелеет в B2B AI

2026-05-01

Setup умирает к 2027. Retainer - не весь. Карта того, что уцелеет в B2B AI

Цифра, после которой остальное - следствия

За 23 месяца стоимость inference уровня GPT-3.5 упала с $20 до $0.07 за миллион токенов. Это сокращение в 280 раз (Stanford HAI AI Index 2025, через анализ Tobias Pfütze). Средняя цена за enterprise-токен упала на 75% за один год - $10 → $2.50 в 2024-2025 (Ramp Velocity). Frontier-модели дешевеют со скоростью 5-10× в год; commoditized capability tiers - 40-900× в год (arXiv 2511.23455, ноябрь 2025). DeepSeek V3 стоит $0.14 за миллион входных токенов против $3.00 у GPT-4o - минус 95% при сопоставимом качестве на большинстве enterprise-задач (Silicon Canals).

Это не пузырь и не маркетинговая флуктуация. Это s-curve, которую видели cloud computing, hard drives, оптоволокно и до них - десятки технологических волн. Исторически такие кривые реверсировались только при регуляторном вмешательстве или физическом лимите ресурсов; в inference ни того, ни другого не просматривается. Studio, продающая «AI-интеграцию под клиента» в 2026 году по тем же правилам, по которым она продавала её в 2024, продаёт продукт, себестоимость которого через 18 месяцев упадёт ещё в несколько раз - а его конкуренция переедет на этаж выше: туда, где сидят управляемые сервисы Bitrix24, RetailCRM и Yandex AI Studio.

Это не «рост рынка». Это commoditization cliff. И механика у него ровно та же, что у web-dev в 2005 году и у SEO в 2015 - только сжатая в три раза по времени.

Не первый раз: web-dev, SEO, digital-marketing

Каждое технологическое поколение проходит одну и ту же кривую: инновация → высокая маржа → инструментарий → платформы → race-to-bottom для commodity-слоя → выживают только специализированные ниши. Цикл одинаковый, разница - в скорости.

Отрасль	Период первичной маржи	Триггер commoditization	Что уцелело на премиальных ценах
Web-dev	1995-2005	Wix, Squarespace, WordPress	Custom enterprise dev, e-commerce платформы, design-first agencies
SEO	2005-2015	Google updates + content tools	Vertical SEO (legal, medical), Answer Engine Optimization
Digital-marketing	2008-2018	Self-serve рекламные платформы	Performance-at-scale, бренд-стратегия
AI services	2023-2027?	LLM commoditization + managed-платформы	Открытый вопрос

Пять-десять лет занимала каждая из предыдущих волн. AI-волна сжата до 24-36 месяцев по трём причинам: commoditization модели идёт не через open-source-копии, а через прямое снижение цен у самого провайдера; execution-платформы (Bitrix24, RetailCRM, Yandex AI Studio) уже встроены в клиентский стек; генеративные инструменты ускоряют сами agencies, и они конкурируют сами с собой.

Boutique Consulting Club описывает это как double squeeze. «AI deflates the value of codifiable tasks, while execution platforms move upstream by bundling light advisory into their products», - пишет Danilo Kreimer, основатель Boutique Consulting Club, в эссе «How Consultants Can Compete (And Win) In An AI-First World». Сверху давит платформенный bundling, снизу - обнуление кодифицируемой работы. Их прогноз по сегментам к 2035 году: low-complexity codifiable работа сохранит 10-20% человеческого присутствия, типичная messy implementation - около 33%, стандартная strategy work - 20-30%, complex transformations - около 50%. Таймлайн - 1-3 года для simple stuff, 3-6 лет для mid-weight knowledge work.

Параллельный сигнал из marketing-сектора: Productive.io опросили 180+ агентств в ноябре 2025 - около трети уже получили запросы на «AI discount», ещё половина ждут такие запросы в ближайшие месяцы (Productive.io). Search Engine Land формулирует это короче: «Agencies embraced AI to cut costs, but clients did the same. Now expectations are rising, budgets tightening, and value

under scrutiny» (Search Engine Land, апрель 2026). Себестоимость падает быстрее, чем цена. Это финансовое определение squeeze.

Историческая параллель полезна для калибровки ожиданий. В 2005 году средняя цена корпоративного сайта в Москве была 150-500 тысяч рублей; к 2012 году WordPress + готовая тема + дизайнер-фрилансер закрывали тот же объём задач за 20-40 тысяч. Уцелели не те, кто продолжал продавать «сайт под ключ», а те, кто переехал в e-commerce-платформы для крупных ритейлеров и в продуктовый дизайн. Та же геометрия повторилась в SEO между 2010 и 2018: дешёвая on-page оптимизация ушла в $200-500, а специализированный legal или medical SEO с compliance-знанием закрепился на $10-15K в месяц. В обоих случаях средний слой исчез не из-за исчезновения спроса, а из-за того, что спрос распался на два класса - самообслуживание и vertical expertise.

Что уже произошло в РФ - а не «может произойти»

Главная ошибка в обсуждении commoditization - говорить о ней в будущем времени. В русском B2B нижняя планка retainer-уровня уехала вниз уже сейчас.

Yandex AI Studio — это managed-платформа Яндекса для сборки LLM-агентов поверх YandexGPT и внешних моделей без кода. Она в связке с SpeechKit покрывает четыре кейса, на которых живёт средняя AI-студия в РФ: квалификация лидов, контроль закрытий сделок, реактивация отказников, SLA-мониторинг скорости ответа. Стоимость для команды 10-15 менеджеров - 12 000-30 000 ₽/мес из коробки (vc.ru / Salekit, апрель 2026). Маркетплейс Bitrix24 содержит подборку AI-приложений и агентов от сторонних разработчиков — часть бесплатна, часть стоит 5–15 К ₽/мес как надстройка к основной подписке. RetailCRM штатно предлагает AI-агентов внутри тарифа, без отдельной интеграции.

В США тренд тот же: Salemwise приводит бенчмарк basic AI automation для SMB - $500-5 000 единоразово плюс $49-500 в месяц. То есть в долларовом эквиваленте дешёвый retainer уже сейчас стоит 4 000-40 000 ₽/мес.

Это сильно меняет арифметику разговора с клиентом. До 2025 года студия могла честно сказать: «то, что мы делаем, нельзя купить как продукт». В 2026 году у клиента уже открыта вкладка с маркетплейсом Bitrix24, и он видит там AI-помощника за 7 000 ₽/мес. Чтобы продать ему ту же базовую интеграцию по типичному студийному бенчмарку (см. разбор среднего бюджета внедрения ИИ в бизнес на vc.ru) — нужно объяснять разницу - а разница больше не очевидна, потому что harness теперь товар (этот сдвиг мы разбирали подробнее в заметке «Harness как commodity, operating layer как moat»).

Прогноз на 2027-2028, основанный на текущей траектории цен и платформенной активности:

Категория	Сегодня (2026)	2027-2028
Setup «подключить AI к CRM», ≤3 use cases	рыночный разброс от фрилансера до студии (vc.ru обзор бюджетов)	Commodity-floor сжимается к стоимости managed-платформ; премиальный setup выживает только как архитектурный аудит
Retainer «поддержка AI-агента»	managed-сервисы 12–30 К ₽/мес (Yandex AI Studio + SpeechKit); студийный retainer сверху	Floor сближается с managed-ценой; премиум выживает только за accountability за business outcomes
Setup «AI-операционная система внутри vertical»	на порядок выше commodity-setup, под конкретный операционный ландшафт	Растёт в связье с ростом сложности governance и compliance
Retainer «AI Operations с outcome-SLA»	редкая категория на рынке РФ в 2026	Стандартный формат vertical-верха: база + % от outcomes

Это и есть форма обрыва: рынок раскалывается на два слоя - commodity-низ и vertical-верх. Среднего слоя, в котором сейчас живёт большинство российских AI-студий, через 18-24 месяца не будет.

Что не схлопывается под прессом коммодитизации?

Защитная геометрия. Из четырёх независимых источников за последний год - BVP AI Pricing Playbook, Pfütze, Euclid Ventures о moat в vertical AI, Boutique Consulting Club - сходится один консенсус: после commoditization модели и harness остаются три источника differentiation, и ни один из них не покупается у поставщика.

Контекст и качество данных. То, что компания накопила внутри себя за годы: исторические решения, паттерны клиентов, regulatory context, граничные случаи. BigDATAwire 2026 enterprise predictions формулируют это как «context capitalism»: differentiation смещается от доступа к модели к точности понимания operating environment. AI-студия не владеет этими данными, но может построить контур, который их захватывает и кодифицирует - и тогда контур становится живым активом клиента, а не отчуждаемым артефактом.

Архитектура рабочего потока. Где сидят governance gates, как outputs верифицируются до того, как они влияют на бизнес, какие триггеры эскалации, какие правила ценообразования, какие исключения допустимы. StackAI 2026 enterprise adoption называет это «repeatability - ability to deliver one governed workflow and replicate it 20 times» и определяет как defining strategic asset. Это

не задача, которую решает Agent Builder. Это задача, которую решает человек, неделями сидящий внутри операции конкретной компании.

Trust и accountability. Кто отвечает, если AI ошибся. Кто согласует с регулятором. Кто поддерживает культурный change. Boutique Consulting Club пишет об этом без эвфемизмов: «what survives are the messy human bits - the politics, trust-building, the delicate art of herding cats».

Деталь, которую в дискуссиях о коммодитизации пропускают: средний слой схлопывается всегда, но *оба полюса* становятся жёстче. Commodity-низ дешевеет быстрее, чем показывает публичный рынок; vertical-верх дорожает быстрее, чем это видно снаружи. Та же динамика наблюдалась в SEO: к 2018 году дешёвая on-page оптимизация стоила $200-500, а специализированный legal SEO с compliance-знанием - $15 000+ в месяц. Разрыв вырос, не сжался.

Почему capability convergence обнуляет «выбор стека» как продукт

Ещё один сигнал, который меняет правила воронки. На SWE-bench Verified - наиболее credible real-world coding benchmark в 2026 году - топ-5 моделей разделены 1.3 percentage points, что Smartscope называет «effectively a tie». Capability convergence на frontier-уровне означает, что выбор модели перестал быть стратегическим решением. Это просто tier-выбор: дороже-точнее или дешевле-почти-как.

Для агентств, которые строили часть своего authority на «мы знаем, какую модель брать под какую задачу», это плохая новость. Эта экспертиза обнуляется. Но домен и контекст - нет. Pfütze формулирует это так: «scaffold, context, and agentic framework determine outcomes more than the intelligence of the model». В 2024 году это было observation. В 2026 - это рабочая гипотеза для всей monetization-стратегии в B2B AI.

Что это значит на практике для studio в 2026

Логичный шаг - не переписывать pricing-страницу, а переписать определение того, что продаётся.

Первое: setup перестаёт быть «технической интеграцией» и становится «архитектурным аудитом и workflow design». Тот же объём работы, но с другой формой ценности. Аудит - это не «подключим API», а «опишем, какие сущности первичны в вашей операции, где governance-gates, какие данные траекторий мы будем собирать с первого дня, как будет выглядеть слой представления через шесть месяцев». Это работа, которую Agent Builder не делает и не сделает - потому что Agent Builder не сидит внутри клиентской операции.

Второе: retainer перестаёт быть «поддержкой агента» и становится «AI Operations с outcome-метриками». Здесь критичен сдвиг в формулировке SLA. Не «время ответа на тикеты» и не «процент uptime агента», а измеримый бизнес-исход - время до контакта с лидом, конверсия на этапе квалификации,

доля реактивированных отказников, потерянные заявки в неделю. То, что попадает в финансовую отчётность клиента, а не в дашборд интегратора. И, как часть того же retainer, гибридный component - % от outcome поверх базы. Это убивает позиционирование «мы такие же, как Yandex AI Studio, только дороже» и заменяет его на «мы берём деньги за результат, а не за инфраструктуру».

Третье - и это структурно важнее первых двух: vertical, а не horizontal. Studio, которая знает один-два рынка глубоко, не попадает в squeeze, потому что её защита - не код, а domain context. Владельцы небольших операционных бизнесов в РФ не покупают «AI». Они покупают решения конкретных операционных болей: потерянные заявки, медленная диспетчеризация, потеря клиента после первого касания, ручная сверка по складу. Между «AI-агент» и «потерянная заявка» лежит пропасть, которую закрывает не модель, а человек, понимающий конкретный бизнес.

Boutique Consulting Club приводит ту же мысль в формате «four escape routes»: либо ты уходишь в depth (узкая вертикаль), либо в accountability (берёшь outcome-SLA), либо в trust (становишься частью команды клиента, а не подрядчиком), либо в proprietary IP (строишь продукт, не сервис). Все четыре маршрута имеют общее свойство - они не масштабируются за счёт SDK. Они масштабируются только за счёт человеческого контакта с конкретным operating environment.

Что мы будем наблюдать

Несколько сигналов, которые покажут, в какую сторону рынок движется быстрее, чем ожидается.

Первый - публичные кейсы в маркетплейсах Bitrix24, RetailCRM, AmoCRM. Когда managed-AI начнёт показывать измеримые outcome-кейсы (не «внедрили AI», а «сократили время до первого контакта на 40%»), нижняя планка retainer-сегмента сдвинется ещё ниже - потому что у клиента появится бенчмарк, против которого он будет считать стоимость работы студии.

Второй - появление outcome-based pricing у заметных российских AI-студий. BVP AI Pricing Playbook фиксирует этот сдвиг в US с 2024–2025: outcome-based контракты выросли с ~5% до ~15% выборки за 18 месяцев. Рынок РФ почти целиком живёт в time-and-materials или setup+retainer; первый, кто публично закроет крупный контракт по схеме «процент от exposed business outcome», задаст фрейм, в который остальные будут вынуждены войти.

Третий - поведение Bitrix24, RetailCRM и Yandex AI Studio в части vertical-templates. Если они начнут выпускать готовые шаблоны под конкретные ниши (стройка, медицина, ритейл), это сожмёт окно для horizontal-агентств ещё на 6-9 месяцев. Если останутся на универсальных конструкторах - окно открыто чуть дольше.

Четвёртый - публичные данные по AI-discount-запросам. Productive.io уже зафиксировали тренд в США и Европе. Аналогичный замер в РФ ещё не сделан,

но появится в течение 2026 года, и его значение будет важнее, чем большинство инвесторских прогнозов.

Пятый - динамика зарплат внутри AI-студий. Если в 2026-2027 средняя ставка middle-инженера, занимающегося интеграцией LLM, начнёт расти медленнее общего IT-индекса в РФ, это будет первым внутренним сигналом, что бюджеты на horizontal-интеграцию ужимаются на стороне клиента. Та же метрика срабатывала в SEO в 2014-2015 годах за 9-12 месяцев до того, как сжатие маржи стало очевидным в публичных финансовых отчётах агентств.

Что остаётся, когда volatile уходит вниз

Главный практический вывод не в том, что AI-студии умрут к 2027 году — параллель с web-dev 2010-х показывает, что большинство выживает при падении среднего чека в 7–10 раз. Главный вывод в том, что между 2026 и 2028 рынок проходит через перераспределение выручки и пропорция «horizontal vs vertical» инвертируется. Универсальная интеграция AI к CRM — основная масса бизнеса в 2024 — к 2027 станет commodity внутри Yandex AI Studio, Bitrix24 AI и RetailCRM, и маржа уйдёт в платформы. Vertical AI Operations с outcome-SLA внутри одной отрасли, выглядевшая узкой нишей в 2024, к 2027 станет основной формой устойчивой маржи в секторе.

Studios, которые рассчитывали «делать всё для всех», в этой рамке теряют экономику. Выживают те, кто готов сидеть внутри одной операционной реальности достаточно долго, чтобы накопить контекст, который не помещается в API. Про bootstrapped-экономику в такой модели я писал отдельно - «Три клиента вместо раунда». Узкий рынок, который большинство сейчас называет ограничением, через 24 месяца окажется единственным местом, где осталась маржа.

Средний слой рынка исчезает не потому, что исчез спрос на AI в B2B, а потому, что спрос распался на commodity-низ и vertical-верх. Студия, которая перестаёт продавать «AI-интеграцию под ключ» и начинает продавать измеримые outcome в одной вертикали, попадает в верхний полюс до того, как этот полюс закроется.

Главное

- **280× за 23 месяца.** Стоимость inference уровня GPT-3.5 упала с $20 до $0.07 за миллион токенов. Это s-curve, из которой нет реверса.
- **Нижняя планка retainer уже уехала.** Yandex AI Studio, Bitrix24, RetailCRM закрывают базовые кейсы за 12–30 К ₽/мес — у клиента появился сравнительный бенчмарк, против которого он считает стоимость работы студии.
- **Рынок раскалывается на два полюса.** Commodity-низ дешевеет быстрее, вертикальный верх дорожает. Средний слой — «AI-интеграция под ключ» — исчезает.

- **Что уцелеет.** Vertical-верх с outcome-SLA, контекст и качество данных клиента, архитектура верифицируемых workflow, accountability за бизнес-исходы. Выбор модели и setup-интеграция — не уцелеет.

FAQ

Правда ли, что setup-фаза AI-проектов исчезнет к 2027? Исчезнет не setup как таковой, а setup в нынешней форме «подключить LLM к CRM под ключ» за 150–350 К ₽. Под давлением managed-платформ эта работа переводится в commodity-флор 30–80 К ₽. Премиум-setup сохранится только как «архитектурный аудит и workflow-design» внутри конкретной вертикали.

Чем retainer в vertical-верхе отличается от «поддержки агента»? Формой SLA. Не «uptime бота» и «время ответа на тикеты», а измеримый бизнес-исход: время до первого контакта с лидом, конверсия на квалификации, доля реактивированных отказников, потерянные заявки в неделю. Outcome-based часть в схеме вознаграждения — обязательный элемент защиты от сравнения с managed-платформами.

Можно ли выжить как horizontal AI-студия? Можно, но на порядок меньшей выручке и на рынке клиентов, которые сравнивают вас с Bitrix24 AI и Yandex AI Studio. Историческая параллель — web-dev-агентства 2010-х, которые выжили как horizontal: их средний чек упал в 7–10 раз.

Применима ли эта логика к РФ в условиях 152-ФЗ и суверенизации стека? Да. Yandex AI Studio, GigaChat и RetailCRM решают вопрос локализации и комплаенса внутри managed-платформы — то есть суверенизация сама по себе ускоряет commoditization, потому что выбивает у horizontal-интегратора один из сильных аргументов.

Что делать сейчас, если вы — средневековая horizontal-студия? Выбрать одну вертикаль из тех, где уже есть 2–3 проекта. Закрыть в ней 1–2 операционных боли до измеримого outcome. Переформулировать retainer в outcome-SLA. И в течение 12 месяцев накопить domain-context, который не помещается в API.