AI-подписка через год: почему стоимость замены растёт быстрее цены

2026-07-01

AI-подписка через год: почему стоимость замены растёт быстрее цены

В отчётах Harvey, AI-сервиса для юридических фирм, за 2025 год прозвучала цифра, которая обычно остаётся за периметром AI-дискуссий: net revenue retention 167% при firmwide adoption 90%+ у клиентов уровня Am Law 100. Это не история про рост клиентской базы — это история про одних и тех же клиентов, которые через год платят почти в 1.7 раза больше. Базовая цена подписки у Harvey, по публичному разбору тезиса service-as-software у Foundation Capital и в собственной коммуникации фирмы, при этом не выросла. Выросла стоимость замены: к 12-му месяцу клиент уже не может вынуть Harvey из рабочего контура, не переписав часть процессов.

Это и есть структурный сюжет AI-сервиса по подписке. Цена остаётся прежней — стоимость воссоздания того же эффекта у другого поставщика к 12-му месяцу вырастает в 3–5×. Не потому, что исполнитель что-то поднял, а потому что внутри клиентского контура накопились три слоя, каждый из которых стоит времени и денег при попытке выйти. Эти слои — данные цикла «решение → исход», доменная схема представления вертикали и вшитые операционные регламенты — растут с накопительным эффектом (compounding): каждый следующий месяц добавляет к ним больше, чем предыдущий.

Цена подписки и стоимость её замены — это разные кривые

Классический разговор о подписке исходит из того, что цена и ценность контракта со временем должны сходиться. Поставщик берёт фиксированную сумму, клиент за неё получает фиксированный объём — оба знают условия. В AI-сервисе по подписке это допущение не работает уже на третий месяц. Поставщик берёт ту же сумму, но получает доступ к контуру, в котором накапливаются данные, а вместе с ними — структурная зависимость клиента от настроек агента, доступов и регламентов.

Foundation Capital в тезисе service-as-software формулирует это так: AI-услуга, встроенная в операционный поток клиента, переоценивается не по затратам поставщика и не по стоимости разработки, а по стоимости воссоздания того же эффекта другим контуром. Это сдвигает экономику подписки в сторону, которая редко обсуждается на этапе подписания. Покупатель платит за месяц работы. Через год он платит за месяц работы плюс невыплачиваемый, но реальный депозит — годовую историю контура, который к моменту попытки уйти стоит замены в несколько раз больше своей месячной цены.

Фред Райхельд, автор классической работы Bain об экономике удержания, формулирует общий принцип: «Increasing customer retention rates by 5% increases profits by 25% to 95%». В AI-сервисе по подписке эта арифметика работает агрессивнее, потому что эффект удержания накладывается на эффект накопления данных. Каждый удержанный месяц не только продлевает контракт — он делает следующий месяц структурно более ценным для клиента. SaaStr в разборе четырёх уровней prompt portability показывает обратную сторону: AI-сервисы, у которых ценность сидит только в промпте, теряют клиентов быстрее, чем средний SaaS, потому что промпт мигрирует копированием. Удержание держится на том, что промпт скопировать можно, а накопленный контур — нет.

Данные цикла «решение → исход»: чего нельзя унести

Решение → исход — это пара записей о том, какое решение приняли (агент или человек после подсказки агента) в конкретной операционной ситуации, и какой исход это решение дало в течение часов или дней после. Это не лог транзакций и не история запросов — это связка «выбранное действие → измеренный результат», привязанная к контексту бизнеса.

У такого слоя две особенности, которые отличают его от любых других «корпоративных данных». Во-первых, он возникает только в режиме реальной работы — его нельзя восстановить из CRM, из выгрузки 1С или из переписки. Во-вторых, он становится полезен только после нескольких сотен записей: ниже этого порога паттерн не отличается от шума.

В практике B2B AI-агентов, которые работают как часть операционного контура, а не как ассистенты на стороне, первый порог переключательных издержек начинается около 50–100 записей таких пар — этого хватает, чтобы калибровка модели под клиента уже не была переносимой через простое копирование промпта. Второй порог — около 1000 событий в течение 90 дней — превращает накопленное в структурное преимущество: новый поставщик не сможет получить эквивалентный набор данных быстрее, чем за тот же квартал реальной работы. a16z в разборе vSaaS как новой формы вертикального ПО фиксирует тот же сдвиг с другой стороны: устойчивая защита в AI-native B2B появляется не на уровне модели, а на уровне накопленных операционных данных конкретной вертикали, и она растёт по мере того, как агент работает в реальном потоке.

К 12-му месяцу клиент сидит на полугоде-году таких данных. Их нельзя ни перенести в другой контур одной кнопкой, ни воссоздать. Это первый слой, который никто не оплачивает явно — но при попытке уйти оплачивается дважды: один раз потерей точности нового агента, второй раз временем, нужным на накопление эквивалентной истории заново.

Схема представления вертикали: чего нельзя пересобрать за квартал

Второй слой не про данные, а про структуру, в которой они хранятся и обрабатываются. У каждой вертикали — будь то аренда оборудования, оптовая дистрибуция или сетевой ритейл — есть свой набор сущностей и связей: что считается операцией, что событием, что объектом, как они соединяются. Это и есть доменная схема — внутренняя модель данных, в которой агент представляет реальность клиента.

В первые недели работы эта схема всегда выглядит как «общая» — то, что поставщик принёс с собой как стартовый шаблон. Через два-три месяца она перестаёт быть общей: в неё вшиваются особенности конкретной отрасли, конкретного формата работы, конкретных партнёров клиента. К полугоду схема знает, как у этого клиента считается смена, что такое «закрытый расчёт», какие три типа возвратов он различает, чем «спорный» отличается от «отменённого». Через год эта схема — отдельный продукт, который клиент не покупал, но получил.

McKinsey в последнем State of AI фиксирует, что компании, которые сообщают о EBIT-эффекте от AI, чаще остальных ссылаются на переработку структуры данных как на ключевой шаг. Этот шаг не делается одной миграцией — он делается итеративно, по мере того как агент работает с реальными ситуациями и поднимает на поверхность неучтённые сущности. У AI-сервиса по подписке это происходит автоматически: каждая новая исключительная ситуация добавляет в схему один узел или одну связь, которых в начале года там не было.

Стоимость воссоздания этого слоя у другого поставщика — это не цена разработки. Это срок: 4–6 месяцев работы в реальном контуре, чтобы накопить ту же глубину представления. На этот срок клиент в момент замены остаётся с менее точным агентом, который ошибается там, где предыдущий уже не ошибался. Это и есть скрытая часть цены, которую никто не пишет в коммерческом предложении.

Вшитые операционные регламенты: то, что нужно перепродавать заново

Третий слой — самый малозаметный и самый дорогой. Когда AI-агент в течение года работает внутри контура клиента, вокруг него постепенно складывается набор регламентов: кто что отвечает, в каких случаях агент решает сам, в каких эскалирует, какие шаги делает автоматически, какие требуют подтверждения. Часть этих регламентов записывается в документы. Большая часть — нет: они становятся неявным знанием команды, тем самым «у нас так принято».

HBR в разборе того, как B2B-продажи перестают быть линейными фиксирует общий факт: значительная часть операционных знаний в зрелом процессе хранится не в документации, а в практике команд. В AI-сервисе по подписке это знание перераспределяется: часть его уходит в настройки агента и в инструк-

ции, часть остаётся у команды, но обе части начинают работать как одно целое. Через год клиент уже не помнит, какой триггер срабатывает в каком случае — потому что не нужно помнить. Контур работает.

В материале Anthropic о паттернах построения агентов (Building effective agents) Anthropic описывает этот сдвиг прямо: «The most successful implementations use simple, composable patterns rather than complex frameworks» — то есть ценность накапливается не в архитектуре, а в том, как простые блоки вживляются в конкретный процесс. Это и есть переход от инструмента к роли. Инструмент можно заменить — нужно найти аналог и настроить. Роль заменить нельзя — нужно перепродать её внутри организации, заново согласовать границы ответственности, заново обучить команду тому, кому что отдавать и в какой форме. Foundation Capital в более жёстком тезисе о том, как поставщики моделей съедают рынок снизу, отмечает: единственная защита от вытеснения — то, что не вынимается из клиента вместе с подпиской, а остаётся внутри контура как часть процесса. Регламенты — это и есть та часть.

Три слоя рядом: что растёт и сколько стоит воссоздать

Три слоя переключательных издержек отличаются не только природой, но и скоростью накопления и ценой выхода. Одна таблица показывает, почему суммарная стоимость замены к 12-му месяцу вырастает в 3–5×.

Слой	Что накапливается	Порог ценности	Стоимость воссоздания у другого поставщика
Данные «решение → исход»	Пары «действие → результат» из реальной работы	50–100 записей — первый порог; ~1000 за 90 дней — структурный	Квартал реальной работы (нельзя ускорить)
Доменная схема вертикали	Сущности и связи конкретной отрасли и клиента	~3 месяца — схема перестаёт быть общей	4–6 месяцев менее точного агента
Вшитые регламенты	Неявные правила эскалации и автономии агента	Складываются к концу первого года	Перепродажа роли внутри организации заново

Этот разбор продолжает линию предыдущей заметки Подписка против проекта: три класса экономики B2B-агентов: там разделялись классы экономики, здесь — механика того, почему один из этих классов удерживает клиента структурно.

Что это значит для основателя и руководителя

Для основателя AI-сервиса арифметика этих трёх слоёв означает, что ценообразование «за месяц работы» к концу года занижает реальную ценность контракта. Если подписка выставлена на «месяц работы», к 12-му месяцу клиент платит за месяц, но фактически выкупает доступ к контуру, который к этому моменту стоит замены 3–5 месячных платежей. Это создаёт асимметрию, на которой и строится NRR: либо клиент остаётся и платит больше за то же, что у него уже есть, либо уходит и платит ту же или большую сумму на восстановление эквивалента у другого поставщика.

Прагматический тест для основателя: на каждый месяц работы посчитать, сколько часов команды клиента ушло на использование агента — и сколько часов другого поставщика потребуется, чтобы воспроизвести эту же интеграцию с нуля. Если второе число превышает первое в 3 раза и больше — у подписки есть структурная защита, и её цену можно индексировать без сопротивления.

Если нет — поставщик строит услугу, а не повторяемый сервис, и через год её удержать не получится. На тех же данных у SaaStr это формулируется короче: при низком уровне prompt portability подписка живёт, при высоком — рано или поздно мигрирует к более дешёвому поставщику.

Для руководителя в роли покупателя AI-сервиса арифметика обратная. Через 6 месяцев цена выхода из контракта перестаёт быть равна цене входа в новый. Это значит, что выбор поставщика к этому моменту нужно делать не на основе сравнения месячного прайс-листа, а на основе оценки того, как быстро и насколько точно текущий контур накапливает три слоя выше. Бенчмарки рынка показывают, что компании с NRR выше 120% в B2B AI — это всегда компании с глубокой интеграцией в операционный поток клиента, а не с широкой и поверхностной воронкой. Это не маркетинговый сигнал — это сигнал того, что покупатели этих сервисов через год не могут уйти без потерь.

Конкретный шаг: на 90-й и 180-й день работы с подпиской запросить у поставщика три показателя по своему контуру — накопленный объём пар «решение → исход», изменения в схеме данных за период, список регламентов, которые были вшиты в работу агента. Если на эти вопросы нет ответа в цифрах и списках — поставщик доставляет не повторяемый сервис, а услугу, и через год отказ от неё ничего не стоит. Если ответ есть и накопление видно — это и есть то, за что стоит платить дальше, не сравнивая месячную цену с альтернативами, потому что альтернатив на этом сроке уже нет.

Как понять, в чью сторону смещается асимметрия?

Сигналы того, что эта механика разворачивается в сторону покупателя или поставщика, появятся в двух местах. Первое — публичные NRR-цифры AI-нативных B2B-компаний. Если медианный NRR в категории остаётся выше 120% на горизонте 18–24 месяцев, это означает, что три слоя выше работают как защита и поставщики научились их защищать. Второе — рост числа интеграционных партнёров и стандартов экспорта данных. Чем больше появляется готовых каналов вынимания истории «решение → исход» из одного контура в другой, тем быстрее размывается первая стенка выхода. Пока второго не видно — асимметрия в пользу поставщиков с накопленным контуром.

Главное

- Цена AI-сервиса по подписке может не меняться 12 месяцев, но стоимость воссоздания того же эффекта у другого поставщика к этому моменту вырастает в 3–5×.
- Три слоя переключательных издержек растут с накопительным эффектом: данные цикла «решение → исход», доменная схема представления вертикали и вшитые операционные регламенты.
- Bain показывает, что рост удержания на 5% даёт +25–95% к прибыли; в AI-сервисе по подписке этот эффект усиливается, потому что удержание и накопление данных работают в одну сторону.

- Для основателя: цена «за месяц работы» к концу года занижает реальную ценность контракта; NRR 120%+ возможен только при глубокой интеграции, а не при широкой воронке.
- Для руководителя: на 90-й и 180-й день нужно мерить накопление трёх слоёв в цифрах; если их нет — подписка ничего не защищает, если есть — заменить её через год будет дороже, чем продлить.

FAQ

Чем «стоимость замены» отличается от цены подписки? Цена подписки — это сколько клиент платит в месяц; она может не меняться весь год. Стоимость замены — это сколько стоит воссоздать тот же эффект у другого поставщика; к 12-му месяцу она вырастает до 3–5 месячных платежей. Это две разные кривые.

Почему нельзя просто скопировать промпт и переехать к дешёвому поставщику? Промпт скопировать можно, но он даёт лишь верхний слой. Накопленные данные «решение → исход», доменная схема и вшитые регламенты не мигрируют копированием — их нужно накапливать заново кварталами реальной работы.

Как понять, что поставщик строит реальный контур, а не разовую услугу? На 90-й и 180-й день запросить три показателя: объём пар «решение → исход», изменения в схеме данных и список вшитых регламентов. Если ответа в цифрах нет — это услуга, от которой через год можно отказаться без потерь.

Почему NRR выше 120% — это сигнал глубины, а не маркетинга? Потому что высокий net revenue retention в B2B AI достигается не широкой воронкой, а глубокой интеграцией в операционный поток: клиенты таких сервисов через год не могут уйти без потерь, поэтому остаются и расширяются.