Service-as-software: как агенты переписывают формулу выручки в B2B

2026-05-12

Service-as-software: как агенты переписывают формулу выручки в B2B

В марте 2026 года Жюльен Бек из Sequoia опубликовал статью «Services: The New Software», к которой за два месяца публично вернулись Y Combinator в RFS Summer 2026 и Bessemer в AI Pricing Playbook: на каждый $1 расходов корпоративного бизнеса на программное обеспечение приходится примерно $6 на людей, которые делают сервис, поддерживаемый этим программным обеспечением (Julien Bek, Sequoia Capital - Services: The New Software, 5 March 2026). До 2026 года инструментальный слой ($1) брали поставщики SaaS, сервисный слой ($6) - операторы, агентства, штатники. Vertical AI впервые в истории корпоративного программного обеспечения претендует на оба слоя одной транзакцией - и это та часть тезиса, которую Sequoia формулирует как верхний потолок, а не центральный прогноз.

Это смена правил, по которым продукт извлекает выручку: не лицензия на инструмент, а оплата за результат работы, выполненной агентом на стороне поставщика. Y Combinator в Summer 2026 Request for Startups ставит ту же рамку - буквальная формулировка с открывающей страницы: «AI has stopped being a feature and started being the foundation» (Y Combinator, Requests for Startups, Summer 2026). В английской терминологии это закрепилось как service-as-software (SaaS2, SaS); по-русски - модель, в которой агент продаёт результат, а не инструмент.

Сдвиг, который пока называют не своим именем

Service-as-software чаще всего описывают как «следующее поколение SaaS» - удобная, но обманчивая формулировка. Меняется ценностная единица: SaaS продаёт лицензию на инструмент и выставляет счёт за seat; service-as-software продаёт завершённую работу и выставляет счёт за outcome - закрытое обращение, оформленную декларацию, забронированную встречу.

В классическом SaaS поставщик отвечает за работоспособность инструмента - ответственность за результат лежит на клиенте. В service-as-software поставщик принимает на себя стоимость вычислений, ошибки и интеграцию - в обмен на цену, анкорованную против стоимости труда, а не инструмента.

Bessemer формулирует это короче: «По мере перехода от consumption через workflow к outcome-pricing вы принимаете больший риск по себестоимости в обмен на более плотное совпадение с ценностью» (Bessemer Venture Partners, AI Pricing Playbook). 92% AI-компаний, по данным того же отчёта, уже работают

в смешанных моделях, где база подписки сочетается с usage- или outcome-компонентом.

Та же оптика звучит и в публичных материалах руководства Y Combinator: десятки вертикалей - здравоохранение, юриспруденция, финансы, страхование, compliance, логистика, customer service - пока находятся в single-digit проникновении AI, а расходы на труд в каждой исчисляются $100B+ в год (Y Combinator RFS Summer 2026). Победитель в такой вертикали - не «SaaS-юникорн с $100M ARR», а значимая доля сервисного слоя индустрии, перепрошитая под экономику программного обеспечения.

Что показывают компании, которые уже играют по новым правилам

К маю 2026 года в публичных данных есть четыре опорных кейса, которые показывают, как меняется формула выручки на практике.

Harvey, юридический AI, дошла от нуля до $195M ARR за 36 месяцев (Sacra, Harvey at $195M ARR). Чек - $1 200-$2 500 на одного юриста в месяц, минимум 20 рабочих мест, двенадцатимесячный контракт. Это формально per-seat, но ценовой якорь - не «лицензия на программное обеспечение». Якорь - 5-7% от стоимости рабочего времени associate в крупной юридической фирме. Если AI-инструмент стоит как 5% от человека, которого он частично заменяет, $14 000 за рабочее место в год становятся не дорогими, а очевидными. По данным того же Sacra-профиля, Harvey движется в сторону revenue-share: доля с billable hours, выставленных клиентам через AI (Sacra, Harvey at $195M ARR). Это переход от per-seat к outcome без отказа от первого.

Sierra, AI для customer support, собрала $100M ARR на чистой outcome-модели: оплата только за разрешённое обращение, сохранённую подписку, состоявшийся апсейл (Sierra, Outcome-based Pricing). Стартовый контракт - $150 000 в год, с расширением до $1M+ при многоканальной голосовой интеграции. Sierra доказала, что в узком домене с измеримым исходом можно совсем отказаться от seat. Но даже у Sierra модель смешанная: для рутинной маршрутизации обращений действует per-conversation, для ценных разрешений - per-resolution. Чистый outcome-pricing - миф, к которому индустрия стремится, но в производственном развёртывании всегда живёт гибрид.

Intercom Fin - самая чистая иллюстрация принципа: $0.99 за разрешённое обращение, никаких seat-fee (Intercom Pricing). Outcome определён однозначно: Fin закрыл тикет без передачи человеку. Bessemer использует Fin как эталонный пример совпадения ценообразования с ценностью.

Glean показывает противоположный паттерн - расширение поверх классического SaaS, а не замена ему. База $45-50 на пользователя в месяц, надстройка Work AI за $15, отдельный SKU за governance, новый consumption-billing для агентных нагрузок поверх per-seat. Результат: $100M → $200M ARR за 9 месяцев, оценка $7.2B (Futurum Group on Glean). Прикладное следствие: устойчивый

SaaS может не ломать существующую модель, а добавлять новые ценовые слои сверху — без перехода на чистый outcome-pricing.

Между четырьмя кейсами есть общее, и оно фиксируется по их же публичным материалам (Sacra Harvey, Sierra outcome-pricing, Intercom pricing, Futurum on Glean). Никто из них не продаёт «AI-инструмент». Harvey продаёт замещение стоимости юриста, Sierra - закрытое обращение, Intercom - разрешённый тикет, Glean - слой корпоративного знания. Каждая компания смогла объяснить клиенту единицу ценности, против которой выставляется счёт, и привязать её к показателям, которые клиент уже считает внутри своей отчётности.

Почему формула «5-7% от стоимости труда» меняет калькуляцию?

В классической экономике корпоративного программного обеспечения цена продукта анкорилась против цены другого продукта. Salesforce стоил против Siebel, потом против HubSpot. Потолок задавался не «сколько ценности», а «сколько готов платить клиент относительно альтернативного программного обеспечения». Это давало 80-90% валовой маржи, потому что предельная стоимость дополнительной лицензии стремилась к нулю.

В service-as-software цена анкорится против стоимости труда, который продукт замещает. Это поднимает потолок: годовой бюджет на двадцать associate в крупной юридической фирме при базе $250-400 тыс. в год на человека - это $5-8M, и $300 000 за инструмент, который реально снимает 15-20% их операционной нагрузки, выглядит дёшево. Но это же опускает валовую маржу: AI-first компании работают на 50-60%, а не на 80-90% (Bessemer AI Pricing Playbook). Себестоимость вычислений вернулась в баланс.

Следствие, которое часто пропускают: в SaaS underpricing вреден, но в service-as-software он смертельно опасен. Поставщик берёт на себя стоимость inference, ошибки, дообучения и поддержки рабочего процесса. Один power-user, который генерирует 10x ожидаемого объёма outcome, способен в один квартал увести юнит-экономику в отрицательную зону. Replit в 2025 году публично прошёл через колебания валовой маржи от 36% до минус 14% за два месяца (Sacra, Replit profile); GitHub Copilot при базовой цене $10 терял до $20 на среднем пользователе и до $80 на heavy-user, что привело к переходу на usage-based pricing в 2026 (DevOps.com, GitHub resets Copilot pricing). Каждый кейс - не сбой стартапа, а структурное следствие новой модели.

Net revenue retention 130%+, который по бенчмаркам Bessemer относится к top decile в классическом SaaS, в выживших service-as-software компаниях становится не идеалом, а условием существования. Harvey демонстрирует 290% YoY роста по данным Sacra - траектория, типичная для service-as-software: от per-seat к revenue-share, от ассистента к workflow-платформе по мере углубления в клиентскую операцию.

Где эта рамка ломается и кто проигрывает

Консенсус Sequoia / YC / Bessemer выглядит ровным, но у него есть три уязвимые точки, которые часто опускают в венчурных текстах.

$1:$6 - это верхний потолок, а не центральный прогноз. Sequoia осторожно подчеркивает это в своём же материале (Bek, Services: The New Software): реальный capturable share зависит от того, насколько глубоко поставщик входит в операцию клиента. В большинстве вертикалей сервисный слой не сжимается до нуля - AI ассистирует человеку, а не замещает его полностью. Реалистичный диапазон захвата для удачной AI-первой компании в 2026 году - не «6× software-рынок», а порядка 1.2-2× software-бюджета вертикали. Это всё равно огромный рынок, но венчурная презентация с множителем «×6» в слайде TAM обычно разбивается о первый квартал продаж.

Риск, переданный поставщику, не нравится крупным клиентам. В регулируемой отрасли CFO неохотно подписывает контракт, в котором внешняя система отвечает за исход операции. В страховании, медицине, финансах compliance-команды требуют человеческого надзора как условие закупки - это возвращает гибрид к per-seaty и снижает долю outcome-компонента в итоговом чеке. Регулятор страхового рынка ЕС в обзоре генеративного AI 2025 года прямо фиксирует доминирование human-in-the-loop как нормы в страховой индустрии (EIOPA Generative AI EU Market Survey - разбор RPC Legal). Крупный юридический, финансовый и медицинский enterprise подписывает service-as-software медленнее, чем ожидает венчурная форвард-проекция.

Три условия удачи редко выполняются одновременно. Модель работает, когда клиент уже мерит исход в своём дашборде, стоимость замещаемой операции достаточно высока, чтобы окупить себестоимость inference, и поставщик финансово выдерживает первоначальный период отрицательной маржи на heavy-userax. В большинстве сделок хотя бы одно из условий отсутствует - и «outcome-pricing» в контракте превращается в слайд о будущем, а счёты выставляются по usage с минимальным консьюмпшеном.

Проигрывают в этой модели три категории. Универсальные «AI-агенты» без вертикальной экспертизы зажаты между управляемыми платформами гиперскейлеров и узкими отраслевыми продуктами. Классический per-seat SaaS, который не успел нарастить consumption-billing поверх своей подписки, проигрывает в renewals к концу 2026 - 92% AI-компаний уже работают в гибридах, по данным того же Bessemer AI Pricing Playbook. Консалтинговые контракты по «AI-трансформации», которые заканчиваются отчётом, а не выходят в продукт внутри операции клиента, не замыкают петлю «решение → исход» и не накапливают защиты поверх лицензионного слоя — подробнее этот механизм разобран в «Trajectory Data: the Decision→Outcome Loop Moat».

Где компаундирующая защита - и почему универсальная обвязка её не даёт

Это и есть та граница, которая отделяет vertical AI с устойчивой выручкой от стартапов, делающих «универсального AI-агента» и упирающихся в потолок на $1-3M ARR. Вертикальные AI-платформы воспроизводят паттерн, который раньше дал Veeva в фармацевтике, Procore в строительстве, ServiceTitan в сервисном бизнесе. Veeva в продуктовой развёртке Vault AI встраивает агентов прямо в отраслевые модули (Veeva Vault AI) - это сигнал, что отраслевой SaaS не намерен отдавать сервисный слой горизонтальным игрокам. NFX в анализе data network effects формулирует условие устойчивости: данные должны быть автоматически захвачены при использовании продукта, давать видимое клиенту улучшение, иметь высокий порог входа для конкурента и медленную асимптоту насыщения (NFX, The Truth About Data Network Effects). Большинство B2B-компаний этот тест не проходят: они собирают данные, но улучшение продукта происходит вручную, а не непрерывно.

Закрытая петля «решение → исход» в конкретной операционной среде - это то, что нельзя восстановить ретроспективно из логов. Через 12 месяцев работы внутри одного клиента у поставщика накапливается доменная модель данных, набор воспроизводимых решающих правил и история фактических исходов под каждым решением - три слоя, которые в академической литературе и индустрии обозначаются как domain model, business logic и outcome history. Это и есть compounding switching cost: горизонтальная обвязка не может воспроизвести их без аналогичного срока работы внутри той же операции.

Универсальный AI-агент, который ставится из конфига за один спринт, в эту защиту не попадает. Он сжимается между управляемыми платформами от Anthropic, OpenAI, AWS и узкими отраслевыми продуктами с собственным системным учётом, регуляторным контекстом и многолетними данными. Средний слой исчезает - и у большинства горизонтальных агентных продуктов нет ответа на этот сжимающийся спред.

Где это работает, где нет, и что важно проверить до контракта

У service-as-software есть условия применимости, и три из них определяют, работает ли модель в конкретной вертикали.

Первое условие — измеримый исход. Регулятор страхового рынка ЕС в обзоре генеративного AI 2025 года (EIOPA — разбор RPC Legal) и McKinsey в State of AI 2025 (Global Survey on AI) фиксируют одно и то же: в регулируемых сервисах человеческий надзор остаётся доминирующим, а галлюцинации — главным цитируемым риском. Compression цены подтверждена в узких доменах с однозначным правильным ответом (поддержка, бронирования, рутинные документы) и не подтверждена там, где исход размыт или требует регуляторного одобрения.

Второе условие — клиент уже считает этот исход. Bessemer в AI Pricing Playbook прямо фиксирует: outcome-pricing работает там, где выбранный value metric уже входит в публичную отчётность клиента (Bessemer AI Pricing Playbook). CFO, который мерит «разрешённые тикеты» или «выигранные тендеры» до того, как поставщик пришёл, — это правильный value metric. Если единицу измерения нужно изобретать вместе с продуктом, цикл сделки удлиняется, а отказ становится дешёвым.

Третье условие — достаточная маржа в операции, чтобы было что замещать. Bek в том же Sequoia-эссе осторожно отмечает: дисраптить работу за $50 в час сложно, замещать работу за $500 в час — оправдано (Bek, Services: The New Software). На нижнем краю операционной экономики переход в outcome-pricing не окупает себестоимость вычислений.

Эти три условия складываются в практический фильтр: модель работает только при пересечении измеримого исхода, его уже существующего учёта у клиента и достаточной маржи операции. Если хотя бы одно условие отсутствует, разумнее оставаться в гибриде с большой долей фиксированной подписки и небольшой usage-надбавкой, а не запускать чистый outcome-pricing.

Для руководителя, который оценивает внедрение: чем меряет себя поставщик в публичных кейсах? Если только «сэкономленные часы» - это эффект первого года в горизонтальной обвязке. Если в SLA фигурирует измеримый бизнес-результат, привязанный к финансовой отчётности клиента, - это другая категория контракта, и она требует другой due diligence.

С инженерной точки зрения главный технический риск — отсутствие фиксации decision traces с первого дня. Связка «решение → исход» — архитектурное решение, которое нельзя добавить позже; без неё через 12 месяцев у компании будут логи, но не будет компаундирующей защиты.

Главное

- **$1:$6 - это не метафора.** Sequoia формализовала: software-слой и services-слой исторически делили выручку 1 к 6; service-as-software забирает оба одной транзакцией. Это смена правил захвата, а не следующее поколение SaaS.
- **Цена анкорится против труда, не против программного обеспечения.** Это поднимает потолок ($150K-$1M+ годовых контрактов на одного клиента) и опускает валовую маржу до 50-60%. Underpricing в этой модели смертелен.
- **Чистый outcome-pricing - миф.** Harvey, Sierra, Intercom Fin, Glean - все работают в гибриде: база + usage или outcome поверх. 92% AI-компаний уже не на чистом per-seat.
- **Средний слой исчезает между двумя полюсами.** Сверху давят отраслевые платформы (Veeva, Procore, ServiceTitan) с собственным AI; снизу - управляемые среды исполнения от гиперскейлеров. Универсальный AI-агент сжимается между ними за 18 месяцев.

- **Компаундирующая защита — в закрытой петле решение→исход.** Доменная модель, воспроизводимые решающие правила и история фактических исходов накапливаются только изнутри операционной среды клиента. Горизонтальный конкурент воспроизводит их либо годами работы внутри той же операции, либо никак.

FAQ

Чем service-as-software отличается от обычного SaaS с AI-фичами? Единицей выставления счёта. SaaS берёт деньги за доступ к инструменту; service-as-software - за выполненный исход. Это переносит операционный риск на поставщика и анкорит цену против стоимости человеческого труда, а не против стоимости альтернативного программного обеспечения.

Можно ли запустить service-as-software без отказа от подписки? Да, и это безопаснее. Glean показал паттерн расширения: per-seat остаётся ядром, поверх него добавляются SKU за AI-возможности, потом - consumption-billing за агентные нагрузки, потом - отдельные модули за governance. Это даёт NDR 130%+ без слома существующих контрактов.

Где outcome-pricing не работает? Там, где исход неоднозначен, регулятор требует человеческого надзора или маржа исходной операции ниже $50 в час. Compression цены подтверждена в support, бронированиях, рутинных документах; не подтверждена в стратегических решениях, регулируемой медицине, сложных трансформациях.

Что должен накопить продукт за первые 12 месяцев, чтобы не быть заменимым? Три базовых слоя из обычной software-инженерии, но привязанные к конкретному клиенту: domain model (какими сущностями оперирует бизнес и как они связаны), business logic (исполнимые правила решений, эскалаций, ценообразования) и outcome history (история фактических исходов под каждым решением с контекстом момента). Порознь берётся любым SaaS-продуктом, вместе они воспроизводимы только внутри той же операции.