



AI ДЛЯ БИЗНЕСА · РАЗБОР

Решает не модель, а слой над агентом

Один и тот же агент на одной модели может работать чисто или нарушать правила, которых вы не видите. Разницу делает тонкая прослойка между «предложил» и «сделал». Разбор свежего исследования и практический контур для бизнеса.

Что показало исследование.

В июне на arXiv вышла работа Organizational Control Layer (2606.04306). Агента-продавца проверили в состязательном торге на 50 эпизодах: без слоя управления и с ним. Вот что важно прочитать правильно.

Дошёл до сделки

94% → 96%

агент и без слоя закрывал почти все сделки

С нарушением правил

88% → почти 0%

столько закрытых сделок ломали бюджет или условия

Видите ловушку. Базовый агент выглядел успешным: он закрывал 94% сделок. Но 88% из них он закрывал с нарушением, вышел за бюджет, согласился на плохие условия, нарушил правило. Чистыми были только 12%.

Что сделал контрольный слой

Сделок осталось столько же, но нарушения он убрал почти в ноль. Слой не сделал агента умнее. Он сделал его честным. Это один тестовый торг на 50 сделках, держите как иллюстрацию механики, не как KPI для вашего бизнеса.

Что это за прослойка.

Сейчас у агента два шага слиты в один: он придумал действие и тут же его исполнил. Идея работы в том, чтобы их разделить.

Агент только предлагает. Исполнять или нет, решает отдельный слой над ним. На этой границе слой сверяет действие с правилами. Если оно нарушает политику или выглядит рискованным, он не исполняет его сам, а передаёт выше: человеку или старшему контуру. Модель не трогают, слой ставится поверх любой.

Проверяли не на синтетике. Взяли среду, где два агента торгуются за сделку с реальными деньгами и один давит на невыгодные условия. То есть состязательная история, близкая к переговорам, закупкам, оплате.

Честный нюанс: за надёжность платят гибкостью

Жёсткий контур убирает почти все нарушения, но в сильно зажатых условиях агент иногда упускает выгодную, но рискованную сделку. Это осознанный размен: предсказуемость против максимума выгоды. Когда на кону деньги и клиенты, предсказуемость обычно дороже.

Почему все смотрят не туда.

Рынок второй год спорит про модели. Какая умнее, чья дешевле, кто кого обогнал по бенчмаркам. А цифры показывают другое.

Цифра 12% не случайна. Она всплывает в трёх независимых источниках за пару недель:

Три источника, одно узкое место

Исследование OCL: только **12%** сделок были и закрыты, и по правилам.
Отраслевой опрос (май): **12%** мультиагентных систем доходит до прода.
Разбор Сбера: в **93%** случаев инженеры подтверждают действия агента не вчитываясь.

Узкое место везде одно. Не способность модели. Контур управления вокруг неё.

Контур из 4 шагов.

Минимум, который стоит собрать прежде, чем дать агенту доступ к деньгам, клиентам или публикации.

1 Подтверждение на необратимое. Отправка денег, удаление, письмо клиенту, публикация: агент это не делает сам, он останавливается и спрашивает человека. Черновики и расчёты пусть делает свободно.

2 Лимит на запуск. Жёсткий потолок по токенам или шагам. По данным разбора Сбера, мультиагентная схема расходует примерно в 15 раз больше обычного чата. Без потолка один зациклившийся агент выест бюджет за ночь.

3 Эскалация при сомнении. Правило «не уверен, спроси». Если агент не понял задачу или данные противоречат, он не угадывает, а передаёт человеку. Один вопрос дешевле уверенной ошибки.

4 Лог всех шагов. Каждое действие записано. Иначе вы не чините систему, а гадаете, где она свернула не туда.

В Claude Code и Cowork кирпичи уже готовы.

Если вы работаете в этих инструментах, контур не придётся писать с нуля.

Cowork

Подтверждение на необратимое (отправка, удаление, публикация, оплата) встроено по умолчанию. Агент видит только ту папку, которую вы открыли.

Claude Code

Слой собирается из разрешений на инструменты и хуков, которые перехватывают действие до исполнения. Плюс есть готовый плагин проверки безопасности.

Вторая сторона контура.

Контур это не только «что агенту разрешено», но и «кто может дотянуться до самого инструмента». На этой неделе для конструктора агентов без кода Flowise выложили рабочий эксплойт с оценкой опасности **9.9 из 10**: один импорт чужого сценария, и сервер чужой, ключи открыты. Дыры теперь не в модели. В обвязке вокруг неё.

Что с этим делать.

Перестаньте выбирать вендора как главное решение. Модель сегодня скорее сменная деталь: открытые веса почти догнали топ, цены падают. Ценность ушла в слой, который превращает умную, но непредсказуемую модель в сотрудника, которому можно доверить деньги и клиентов.

Соберите контур из четырёх шагов. Начните с подтверждения на необратимое и лимита расхода, остальное добавите по ходу.

Главная мысль

Опасность не в том, что агент не работает. Он работает и закрывает сделки. Опасно, что значительная часть его действий нарушает правила, которых вы не видите. Это чинит слой, не модель.

Источник: Organizational Control Layer, arXiv:2606.04306 (3 июня 2026), код github.com/SHITIANYU-hue/ama_i_ocl. Все цифры из одной работы, на 50 эпизодах синтетического торго, независимо пока не перепроверены. Держите их как иллюстрацию механики, не как обещание результата для вашего бизнеса. Опрос про 12% в проде и разбор Сбера про 93% автоподтверждений приведены как параллели, не как часть того же исследования.



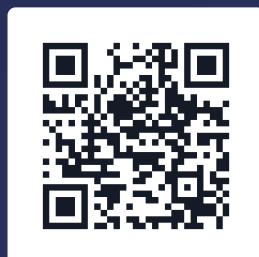
Разбираю AI для бизнеса без хайпа и без воды

Беру свежие исследования, релизы и реальные кейсы, проверяю цифры и перевожу на язык владельца бизнеса. Что брать в работу, что обойти стороной, сколько это стоит.

- Разборы с цифрами, а не маркетинг
- Практики Claude Code и Cowork, которые можно повторить
- Контур управления для агентов: как доверить AI деньги и клиентов

@gorilla_under_hood

Telegram · t.me/gorilla_under_hood



наведите камеру