

1 — Что считается computer-use агентом?

- ☐ Он управляет софтом как человек — смотрит на экран, кликает, печатает — без заранее построенного «чёрного хода».
- ☐ Он выполняет цикл: НАБЛЮДЕНИЕ (скриншот) → РАССУЖДЕНИЕ (vision-language модель выбирает одно действие) → ДЕЙСТВИЕ (клик/текст) → снова наблюдение.
- ☐ Если у целевой системы уже есть нормальный интерфейс для программ, ограниченная интеграция быстрее, дешевле, надёжнее. Агент — только когда «чёрного хода» нет.

2 — Какой вопрос вы задаёте? (Track A против Track B)

Track B — строить быстрее : агент управляет ВАШИМИ инструментами (код, терминал). Ошибка = плохой коммит на ревью.
Track A — действовать за юзера : агент управляет продуктом ЗА пользователя. Ошибка = не то действие на аккаунте.
Это два отдельных решения с разным профилем риска. Не смешивайте их.

3 — Пять продуктов 2026 в одном взгляде

Claude Code (B) : пишет/тестирует/коммитит код по словам. Лучше всего: миграции, рефакторинги, починка CI.
OpenAI Operator (A) : управляет браузером; теперь «agent mode» ChatGPT. Отдельный сайт закрыт (июль 2025).
Manus AI (A) : длинные автономные задачи из одного промпта. Кредиты = непредсказуемо; куплен Meta (дек. 2025).
Perplexity Comet (A) : AI-браузер на Chromium; делает рутину. Агент РЯДОМ с приложением, а не внутри него.
Apple Intelligence(A): на устройстве, приватность; Siri + App Intents открывают ОГРАНИЧЕННЫЕ действия. Только Apple.

4 — Проверка реальностью на OSWorld (задайте ожидания)

- ☐ Меряйте агентов против ~72% уровня ЧЕЛОВЕКА на OSWorld (369 реальных десктоп-задач), а не против 100%.
- ☐ Только-веб-задачи гораздо легче (~87% WebVoyager), чем полный десктоп (Operator CUA набрал 38,1% на OSWorld).
- ☐ Надёжность накапливается: пять шагов по 90% = $0,9^5 = \sim 59\%$ от начала до конца. Длинные цепочки падают чаще, чем кажется в демо.

5 — Строить, покупать или пропустить — и три обязательные рамки

- ☐ Track B: используйте кодового агента (Claude Code / Cursor). Держите тесты зелёными, ревью каждого коммита, архитектура — за человеком.
- ☐ Track A: если задача веб-образна и узка — агент готов к продакшену. Если нужен широкий контроль десктопа — подождите или ограничьте.
- ☐ РАМКА 1 — одобрение человека на любое действие, которое тратит деньги, отправляет сообщение или удаляет данные.
- ☐ РАМКА 2 — жёсткий лимит циклов на задачу (чтобы запутавшийся агент не заиклился навсегда).
- ☐ РАМКА 3 — жёсткий лимит расхода (каждый цикл тарифицируется; ~1-2 цента быстро складываются на длинных задачах).