ЧТО ДЕЛАТЬ ПОСЛЕ БЛОКИРОВКИ

CLAUDE CODE ЗАБЛОКИРОВАЛИ

Полный гайд на 9 страниц: клиенты вместо Claude Code (Cline, Kilo Code, OpenCode, Aider), четыре бюджета с реальными ценами, Mac на M4 Pro / Max / Ultra, серверные V100 и 2080 Ti 22 ГБ, связки через NVLink и выбор моделей. **Цены — с Авито, DNS, Citilink, Ozon, BestMac и G.Race.**

ТЕРМИНАЛ

```
$ claude
> editing src/app.ts
```

403 · account banned

02 **Два пути:** облако или своё железо
03 **Клиенты:** Cline, Kilo Code, OpenCode, Aider
04 **Четыре бюджета:** 50 / 150 / 200 / 300 тыс. ₽
05 **Mac:** M4, M4 Pro, M2/M4 Max, M2/M3 Ultra
06 **V100, 2080 Ti и P40:** дешёвая память
07 **NVLink и мульти-GPU:** когда нужен мост
08 **Модели:** Qwen3.5 и Qwen3-Coder в Q4
09 **Готовые связки** и быстрый старт за 15 минут

ДВА ПУТИ

ОБЛАКО ИЛИ СВОЁ ЖЕЛЕЗО

Облако — это аренда: быстро и дёшево на старте, но сервис может отключиться в любой день. Своё железо — собственность: платите один раз, не зависите от блокировок и не отдаёте код и данные наружу.

БЫСТРЫЙ ПУТЬ

ОБЛАКО

- + 10 минут до первой строки кода
- + максимальное качество моделей, оплата за токены
- + ничего не нужно покупать и обслуживать
- − зависимость от чужого сервиса и правил
- − код и промпты уходят на внешний сервер
- − доступ, тарифы и оплата могут измениться

OpenCode, Cline, Kilo Code или Qwen Code + облачный API.

НЕЗАВИСИМЫЙ ПУТЬ

СВОЁ ЖЕЛЕЗО

- + никто не заблокирует и не отключит
- + код, договоры и данные остаются внутри
- + платите один раз за железо, не за токены
- − локальные модели уступают топовым облачным
- − нужно один раз настроить стек
- − у GPU-сборок — шум, нагрев и электричество

OpenCode, Cline, Kilo Code или Aider + локальная модель через Ollama / vLLM.

Облачные модели: что доступно из России

СЕРВИС	ДОСТУП ИЗ РФ	ОПЛАТА	КАЧЕСТВО В КОДЕ	КУДА УХОДЯТ ДАННЫЕ
GigaChat	прямой	рубли, счёт юрлица	среднее (шаблоны, типовой код)	серверы в РФ
YandexGPT	прямой	рубли, счёт юрлица	среднее (скрипты, документация)	серверы в РФ
Qwen (Alibaba)	через API / агрегаторы	зарубежная карта / агрегатор в ₽	сильное (Qwen3-Coder, Qwen3.5)	за рубежом
DeepSeek	через API / агрегаторы	зарубежная карта / агрегатор в ₽	сильное (агентный код, рефакторинг)	за рубежом

Какой путь выбрать

СИТУАЦИЯ	РЕШЕНИЕ
Нужно вернуться к работе прямо сегодня, проект без секретов	ОБЛАКО
Поддержка кода под NDA, персональные данные, финтех, госсектор	ЛОКАЛЬНО
Команда из 3–10 разработчиков, общий внутренний контур	СЕРВЕР
Чувствительные модули и данные — внутри, сложная архитектура — в облаке	ГИБРИД

ЧЕМ ЗАМЕНИТЬ CLAUDE CODE: КЛИЕНТЫ И IDE

Claude Code — это агентная оболочка над моделью. Сама оболочка заменяется бесплатными открытыми инструментами за 10 минут: они умеют читать проект, править файлы, запускать тесты и работать с локальными моделями через Ollama, LM Studio или vLLM.

ИНСТРУМЕНТ	ГДЕ ЖИВЁТ	ЛОКАЛЬНЫЕ МОДЕЛИ	АГЕНТНОСТЬ	ЧЕМ СИЛЁН И КОМУ ПОДХОДИТ
OpenCode	Терминал (CLI)	Да (Ollama, vLLM, любой OpenAI API)	Сильная	**Прямая замена Claude Code в терминале.** Сессии, редактирование файлов, запуск команд, удобен по SSH и на сервере.
Cline	VS Code	Да (Ollama, LM Studio, OpenAI API)	Сильная	**Стандарт в VS Code (60k+ звёзд).** Режим Plan/Act, показывает каждую правку как diff, запускает терминал и браузер, встроенный MCP Marketplace.
Kilo Code	VS Code / JetBrains / CLI	Да (Ollama + 500+ моделей)	Сильная	**Живой наследник Roo Code.** Режимы Architect / Code / Debug, оркестрация субагентов, индексация репозитория, автодополнение.
Aider	Терминал (CLI)	Да (Ollama, vLLM, OpenAI API)	Сильная (git)	**Лучший для git-дисциплины.** Строит карту репозитория, делает чистые атомарные коммиты, экономнее расходует контекст на средних моделях.
Qwen Code	Терминал (CLI)	Да (локальные Qwen и API)	Сильная	Заточен под семейство Qwen3-Coder и его формат tool calling; быстрый старт, если основная модель — Qwen.
Continue	VS Code / JetBrains	Да (Ollama, vLLM)	Средняя	Лучший выбор для быстрых инлайн-правок, автодополнения (FIM) и чата по коду на слабом железе (модели 4B–9B).
Zed / Void	Отдельный редактор	Да (Ollama, OpenAI API)	Средняя / сильная	Zed — очень быстрый нативный редактор со встроенным агентным чатом. Void — открытая альтернатива Cursor без привязки к чужому облаку.
Tabby / Twinny	Self-hosted / VS Code	Да (только локально)	Автодополнение	Tabby поднимается одним контейнером на команду для автодополнения и чата. Twinny — невесомый плагин для локального Ollama.

Важно: Roo Code закрыт в мае 2026. Репозиторий заморожен и больше не развивается. Если вы сидели на Roo или видите его в старых гайдах — переходите на **Kilo Code** (у него есть прямая миграция настроек и режимов) или на базовый **Cline**.

Почему маленькие модели ломают агентов. Cline, Kilo Code и OpenCode на каждом шаге требуют точного вызова инструментов (JSON/XML со схемой). Модели на 7–9B быстро теряют формат и зацикливаются. **Практический минимум для агентной работы — Qwen3-Coder 30B-A3B или Qwen3.5 27B/35B (от 24 ГБ памяти).** Если памяти 8–16 ГБ, ставьте Continue или Twinny для автодополнения и чата, либо Aider в режиме простых правок одного файла.

РЕАЛЬНЫЕ ЦЕНЫ · НОЯБРЬ 2026

ЧЕТЫРЕ БЮДЖЕТА

Цены — типичные по Москве на 3–4 ноября 2026. Б/у — Авито и BestMac, новое — DNS, Citilink, Ozon. Курс ЦБ — 80,97 ₽ за доллар. Перед покупкой проверяйте: цены на 3090 сильно зависят от состояния карты.

БЮДЖЕТ	ЧТО КУПИТЬ	МОДЕЛИ И КЛИЕНТ	СКОРОСТЬ	ЧТО ТЯНЕТ
50 000 ₽	**Мини-ПК N100, 16 ГБ** — 10–20 тыс. или **Mac Mini M1 16 ГБ б/у** — медиана 44,75 тыс.	Qwen3.5 9B (6,6–7,6 ГБ) Qwen3.5 4B (3,3–4 ГБ) Continue / Twinny	4–15 ток./с	Автодополнение, простые скрипты, вопросы по коду. Для полноценных агентов слабоват.
150 000 ₽	**Mac Mini M4 24 ГБ / 512 ГБ** — 132–140 тыс. или **ПК с RTX 4060 Ti 16 ГБ** — 90–110 тыс.	Qwen3.5 9B / 27B Continue, Aider, простые задачи в Cline	8–25 ток./с	Код и небольшие правки. 27B (17–20 ГБ) на 24 ГБ идёт впритык без запаса под большой контекст.
200 000 ₽	**Б/у RTX 3090 24 ГБ + сборка**: карта 120–130 тыс. + тушка 40–50 тыс. Итого 160–180 тыс.	**Qwen3-Coder 30B-A3B** (19 ГБ) Qwen3.5 27B / 35B OpenCode, Cline, Kilo, Aider	30–50 ток./с (30B-A3B)	**Лучший баланс.** Полноценная агентная разработка, диффы по проекту, 30B-A3B (MoE, 3,3B активных).
300 000 ₽	**Mac Studio M2 Max 64 ГБ б/у** — 180–250 тыс. или **AI Max+ 395 128 ГБ** — 299 тыс. или **сервер 2×3090 (48 ГБ)** — от 300 тыс.	**Qwen3-Coder-Next 80B-A3B** (52 ГБ, на 64–128 ГБ) 70B Q4 — на 2×3090	8–20 ток./с	Почти облачный уровень. На 2×3090 (48 ГБ) идёт 70B Q4; для 80B-A3B (52 ГБ) нужны 64–128 ГБ памяти.

Главное про цены. RTX 3090 в сборе — это не 150, а **160–180 тысяч рублей**: живая карта стоит 120–130 тыс., ещё 40–50 тыс. уходит на корпус, БП от 850 Вт, плату, процессор и ОЗУ. В бюджет 150 тыс. честно помещаются только Mac Mini M4 24 ГБ или сборка с новой RTX 4060 Ti 16 ГБ (сама карта в DNS — 43–51 тыс.).

Чек-лист при покупке б/у RTX 3090. 30 минут под нагрузкой (FurMark + тест VRAM), температура чипа и особенно **HotSpot / память** (у 3090 чипы памяти стоят и на обратной стороне платы и часто перегреваются), отсутствие артефактов и ошибок памяти, живые вентиляторы без люфта, не оплавленные разъёмы 8-pin. Берите с проверкой или Авито Гарантией.

MAC: КАКОЙ ЧИП БРАТЬ

У Apple Silicon единая память (Unified Memory) доступна и процессору, и графике. Поэтому решают две цифры: **объём памяти** (какая модель влезет) и **пропускная способность** (сколько токенов в секунду вы получите). Цены — Москва: i-store, DNS, re-store, Megabait, BestMac.

МОДЕЛЬ	ЧИП	ПАМЯТЬ	BANDWIDTH	ЦЕНА	ЧТО ЗАПУСКАЕТ И КАК БЫСТРО
Mac Mini M1 16 б/у	M1	16 ГБ	68 ГБ/с	~44 750 ₽	Qwen3.5 9B (10–15 ток./с), 4B. Тихий входной билет.
Mac Mini M4 16/512	M4	16 ГБ	120 ГБ/с	77 400 ₽	Qwen3.5 9B и 4B. 27B не влезет.
Mac Mini M4 24/512	M4	24 ГБ	120 ГБ/с	132–140 тыс.	9B свободно; 27B (17–20 ГБ) впритык (8–12 ток./с), контекст мал.
Mac Mini M4 Pro 24/512	M4 Pro	24 ГБ	273 ГБ/с	157–200 тыс.	Те же 9B/27B и 30B-A3B впритык, но **в 1,8–2,2 раза быстрее M4**.
Mac Mini M4 Pro 48/512	M4 Pro	48 ГБ	273 ГБ/с	210–255 тыс.	**Золотая середина компактного Mac:** 27B, 35B, 30B-A3B с большим контекстом.
Mac Mini M4 Pro 64/512	M4 Pro	64 ГБ	273 ГБ/с	240–320 тыс.	30B-A3B быстро; 70B Q4 и 80B-A3B влезают, но узкая шина (4–7 ток./с).
Mac Studio M2 Max 64 б/у	M2 Max	64 ГБ	400 ГБ/с	180–250 тыс.	**Лучший б/у Mac:** 30B-A3B (25–35 ток./с), 70B Q4 (8–12), 80B-A3B (10–14).
Mac Studio M4 Max 36/512	M4 Max	36 ГБ	410 ГБ/с	256 900 ₽	27B, 35B и 30B-A3B летают; 70B/80B (40–52 ГБ) не влезут.
Mac Studio M4 Max 64/1TB	M4 Max	64 ГБ	546 ГБ/с	536 410 ₽	70B Q4 (12–16 ток./с), 80B-A3B (12–15), 30B-A3B (35–45).
Mac Studio M2 Ultra 128 б/у	M2 Ultra	128 ГБ	800 ГБ/с	450–555 тыс.	70B и 80B-A3B с огромным контекстом, Qwen3.5 122B Q4 (8–10 ток./с).
Mac Studio M3 Ultra 96/1TB	M3 Ultra	96 ГБ	819 ГБ/с	679 900 ₽	70B/80B, 122B Q4 (10–12 ток./с). Максимальная скорость среди Mac.

Про Mac Mini на M4 Pro. У M4 Pro шина памяти 273 ГБ/с против 120 ГБ/с у базового M4. Но версия **M4 Pro 24 ГБ** упирается в тот же потолок объёма, что и обычный M4 24 ГБ (macOS отдаёт под модель ~18–20 ГБ). Если берёте M4 Pro именно под агентный кодинг — берите **48 ГБ** (210–255 тыс.): туда свободно встаёт Qwen3-Coder 30B-A3B и остаётся запас под контекст репозитория.

А если на 64 ГБ? В бюджете 200–250 тыс. **б/у Mac Studio M2 Max 64 ГБ** (шина 400 ГБ/с, мощное охлаждение) выгоднее и быстрее на больших моделях, чем новый Mac Mini M4 Pro 64 ГБ (273 ГБ/с за 240–320 тыс.). Ноутбуки **MacBook Pro M4 Max** (36 ГБ — от 352 990 ₽, 64 ГБ — ~450–550 тыс.) берите только ради мобильности: под долгой генерацией они греются и сбрасывают частоты.

V100 И 2080 TI: ДЕШЁВАЯ ПАМЯТЬ

Списанные серверные ускорители Tesla V100 и модифицированные RTX 2080 Ti дают 16–32 ГБ быстрой памяти в разы дешевле RTX 3090. Это рабочий путь для энтузиастов и домашних лабораторий, но со своими инженерными компромиссами.

КАРТА	ПАМЯТЬ	ШИНА	ТОЧНАЯ ЦЕНА В РФ	ЧТО ЗАПУСКАЕТ	ГЛАВНЫЙ КОМПРОМИСС
Tesla V100 16 ГБ SXM2	16 ГБ HBM2	900 ГБ/с	14 990–18 000 ₽ (Авито)	Qwen3.5 9B Q4	**Не PCIe.** Нужна серверная плата с SXM2 или дорогой переходник.
Tesla V100 16 ГБ PCIe	16 ГБ HBM2	900 ГБ/с	26 411–29 345 ₽ (G.Race) 27 352 ₽ (Авито)	Qwen3.5 9B Q4 (очень быстро); 27B не влезет	Пассивный радиатор без кулера, нет видеовыходов, старая Volta.
Tesla V100 32 ГБ SXM3	32 ГБ HBM2	900 ГБ/с	45 000–60 000 ₽ (Авито)	27B, 35B, 30B-A3B Q4	Серверный форм-фактор SXM3 (питание 48 В), в обычный ПК не встанет.
Tesla V100 32 ГБ PCIe	32 ГБ HBM2	900 ГБ/с	64 900 ₽ (Авито) 77 741 ₽ (Маркет) новая со склада: 250 тыс.	27B, 35B, 30B-A3B Q4 с запасом контекста	250 Вт, пассивное охлаждение (нужна «улитка»), нет нативного BF16 / INT4.
RTX 2080 Ti 11 ГБ	11 ГБ GDDR6	616 ГБ/с	20 000–37 999 ₽ (Авито)	Qwen3.5 9B Q4 (быстро)	11 ГБ мало для агентных 27B–30B моделей; 250 Вт.
RTX 2080 Ti 22 ГБ (мод)	22 ГБ GDDR6	616 ГБ/с	40 000–55 000 ₽ в наличии комплект 2 шт. + мост: 97,5–100 тыс.	27B и 30B-A3B Q4 целиком в памяти	Перепаянная память из Китая, пропатченный BIOS, без заводской гарантии.
Tesla P40 24 ГБ	24 ГБ GDDR5	346 ГБ/с	18 000–27 352 ₽ (Авито)	27B Q4, но медленно	Архитектура Pascal (2016), нет Tensor Cores и нормального FP16.

Подробно про Tesla V100. Главное достоинство V100 — память HBM2 со скоростью 900 ГБ/с (на уровне RTX 3090) и объём 32 ГБ в старшей версии, куда свободно помещается 30B-A3B с большим контекстом. Но есть три подводных камня: **1)** архитектура Volta (Compute Capability 7.0) не поддерживает современные инструкции BF16 и FlashAttention-2 так, как Ampere, поэтому на длинном контексте (prefill) она заметно медленнее 3090; **2)** PCIe-карта идёт без вентиляторов — на неё обязательно печатают переходник под серверную турбину (шумная); **3)** дешёвые лоты за 15–18 тыс. — это почти всегда SXM2, которые нельзя воткнуть в обычную материнскую плату.

Подробно про RTX 2080 Ti 22 ГБ. С завода у 2080 Ti 11 ГБ памяти. Китайские мастерские распаивают на плату чипы удвоенной плотности и прошивают модифицированный BIOS — получается 22 ГБ за 40–55 тыс. ₽ (втрое дешевле 3090). Шина 616 ГБ/с и архитектура Turing (CC 7.5) дают хорошую скорость на Qwen3-Coder 30B-A3B. Риск — качество пайки и б/у чипов памяти: брать только с проверкой в MATS/MODS и получасовым прогоном под полной загрузкой VRAM.

NVLINK: КОГДА НУЖЕН МОСТ МЕЖДУ КАРТАМИ

Когда модель не помещается в одну видеокарту (например, 70B Q4 весит ~40 ГБ), её делят между двумя GPU. Карты можно связать двумя способами: просто вставить в PCIe-слоты или соединить сверху жёстким мостом **NVLink**.

> **Главное заблуждение.** NVLink на бытовых картах **не превращает две карты по 24 ГБ в одну карту на 48 ГБ** автоматически. Память всё равно остаётся раздельной, и движок (vLLM, llama.cpp, ExLlamaV2) должен явно резать модель. NVLink ускоряет именно пересылку данных между картами на каждом слое.

Два режима работы двух видеокарт

PIPELINE PARALLEL (ПО СЛОЯМ)

БЕЗ NVLINK (ПО PCIE)

- Первая карта считает слои 1–40, передаёт вектор второй, та считает слои 41–80.
- Обмен между картами минимальный — хватает обычного PCIe 3.0 / 4.0.
- Работает в **Ollama и llama.cpp** из коробки.
- **Минус:** карты работают по очереди, поэтому скорость генерации почти как у одной карты.

TENSOR PARALLEL (ВМЕСТЕ)

С МОСТОМ NVLINK

- Обе карты считают **каждый слой одновременно**, складывая свою пропускную способность памяти.
- Карты синхронизируются десятки раз за один токен — узкая шина PCIe здесь тормозит.
- Работает в **vLLM и ExLlamaV2** (`--tensor-parallel-size 2`).
- **Плюс:** на 2× RTX 3090 мост NVLink даёт **до +40–50% скорости** относительно PCIe.

Поддержка NVLink и цены на мосты в России

СВЯЗКА КАРТ	СУММАРНО VRAM	СКОРОСТЬ NVLINK	ЦЕНА МОСТА / КОМПЛЕКТА НА АВИТО	ВЕРДИКТ
2 × RTX 3090 (24 ГБ)	48 ГБ	112,5 ГБ/с	2-слота: **24 000 ₽** 3-слота: **38 990–40 000 ₽** б/у проскакивают от 3 500–15 000 ₽	Последняя GeForce с NVLink. Имеет смысл под vLLM и 70B Q4. На 4×3090 мост объединяет только пары (+~10%).
2 × RTX 2080 Ti 22 ГБ	44 ГБ	100 ГБ/с	Мост Turing отдельно: **~7 000 ₽** Готовый комплект 2×22 ГБ + мост: **97 500–100 000 ₽**	Самые дешёвые 44 ГБ с быстрым интерконнектом. Тянет 70B Q4 впритык, но высокий риск по железу.
2 × Tesla V100 32 ГБ PCIe	64 ГБ	100–200 ГБ/с (1–2 моста)	Редкие серверные мосты Quadro/Tesla GV100: **10 000–25 000 ₽**	Даёт 64 ГБ (влезает 80B-A3B Q4), но сложная сборка с внешним обдувом. В SXM2/SXM3 NVLink 300 ГБ/с уже в плате.
RTX 4090 / 4060 Ti / 3060	—	**Нет NVLink**	Разъём физически убран NVIDIA	Работают в паре только через PCIe (на платах с режимом x8/x8 и P2P).
Mac (M-серия) / AI Max+ 395	до 128 ГБ	**Не нужен**	Единое поле памяти на одном чипе	Никаких мостов, БП на 1,5 кВт и потерь на синхронизацию между картами.

Инженерные грабли с NVLink на 3090. 1) Шаг слотов жёсткий: мост на 3 слота (60 мм) подойдёт только если между слотами PCIe на материнской плате ровно 60 мм, а карты одинаковой высоты (лучше брать одинаковые модели, в идеале 2-слотовые турбины). **2) Охлаждение:** верхняя 3090 задыхается от горячего бэкплейта нижней — нужен корпус с мощным боковым обдувом или серверные турбины. **3) Питание:** пики двух 3090 требуют честного блока питания на **1200–1600 Вт** и ограничения лимита мощности (`nvidia-smi -pl 280` почти не режет скорость LLM, но сильно снижает нагрев).

КАКИЕ МОДЕЛИ СТАВИТЬ

Все модели ниже открытые и ставятся одной командой в Ollama или vLLM. Размеры указаны для 4-битного квантования (Q4_K_M / AWQ); память должна вмещать и саму модель, и KV-кэш контекста репозитория.

МОДЕЛЬ В OLLAMA	ВЕС Q4	НУЖНО ПАМЯТИ	СКОРОСТЬ (ОРИЕНТИР)	НАЗНАЧЕНИЕ
qwen3.5:4b	3,3–4 ГБ	8 ГБ	8–15 ток./с на N100	Слабые мини-ПК и ноутбуки: автодополнение в Continue и простые справки.
qwen3.5:9b	6,6–7,6 ГБ	12–16 ГБ	10–15 на M1; 35–45 на 2080 Ti; 45–60 на 3090	Универсальная мультимодальная модель (текст + картинки). Отличный быстрый помощник, но для сложных агентов мала.
qwen3.5:27b	17–20 ГБ	24 ГБ	15–22 на 3090; 8–12 на M4 24 ГБ; 18–25 на M4 Pro	Плотная модель для рассуждений и кода. На 24 ГБ оставляет скромный запас под контекст.
qwen3.5:35b	22 ГБ	24–32 ГБ	12–18 на 3090; 15–20 на V100 32 ГБ / M4 Pro 48 ГБ	Потолок для одной 24 ГБ карты (контекст 8–16K). На 32–48 ГБ раскрывается полностью.
qwen3-coder:30b (30B-A3B MoE)	19 ГБ	24 ГБ	**30–50 ток./с на 3090;** 25–35 на M2 Max / 2080 Ti 22 ГБ	**Главный выбор для одной 3090, V100 32 ГБ или M4 Pro 48 ГБ.** 30B всего, 3,3B активны на токен, специально обучена под агентный кодинг.
Llama / Qwen 70B Q4	~40–42 ГБ	48 ГБ	12–18 на 2×3090; 18–25 на 2×3090 с NVLink (vLLM)	Оптимальный потолок для связки **2 × RTX 3090 (48 ГБ)** или Mac Mini M4 Pro 64 ГБ.
qwen3-coder-next (80B-A3B MoE)	52 ГБ	64–96 ГБ	10–15 на Mac Studio 64 ГБ; 15–22 на 2×V100 32 ГБ; **на 2×3090 (48 ГБ) в Q4 не влезает**	**Лучший локальный кодер.** 80B всего, 3B активны, 256K контекст, обучена на 800K исполняемых задач. Для Mac Studio 64 ГБ+, AI Max 128 ГБ или 4×3090.
qwen3.5:122b	81 ГБ	128 ГБ	8–12 на M2/M3 Ultra; 5–8 на AI Max+ 395 128 ГБ	Тяжёлый локальный максимум для систем со 128 ГБ единой памяти.

Почему MoE-модели (30B-A3B и 80B-A3B) выиграли локальный рынок. В классической модели каждый токен проходит через все миллиарды параметров — поэтому 32B или 70B генерируют текст медленно. В архитектуре Mixture-of-Experts (MoE) в памяти лежит большая база знаний (30B или 80B), но для каждого слова включаются только нужные «эксперты» на **3–3,3 млрд параметров**. В итоге вы получаете ум большой модели при скорости генерации в 2–3 раза выше.

Не забывайте про контекст. Агент (Cline, Kilo, OpenCode) отправляет в модель системный промпт, дерево файлов и куски кода — это сразу 15–40 тысяч токенов. Если модель заняла 23 ГБ из 24 ГБ VRAM, под контекст места не останется, и часть слоёв уедет в системную ОЗУ с падением скорости в 5 раз. Оставляйте минимум 4–5 ГБ свободной видеопамяти под KV-кэш (в Ollama включайте квантование кэша `OLLAMA_KV_CACHE_TYPE=q8_0`).

ИТОГ И ЗАПУСК

ГОТОВЫЕ СВЯЗКИ И БЫСТРЫЙ СТАРТ

ВАША ЗАДАЧА	ЖЕЛЕЗО (НОЯБРЬ 2026)	КЛИЕНТ + МОДЕЛЬ
Дешёвый вход / автодополнение	Мини-ПК N100 (10–20 тыс.) или Mac Mini M1 16 ГБ б/у (~45 тыс.)	**Continue / Twinny** + `qwen3.5:4b` / `qwen3.5:9b`
Тихий рабочий Mac на стол	**Mac Mini M4 Pro 48 ГБ** (210–255 тыс.) или **Mac Studio M2 Max 64 ГБ б/у** (180–250 тыс.)	**OpenCode / Cline / Kilo** + `qwen3-coder:30b` (на 64 ГБ — `qwen3-coder-next`)
Максимум скорости за 200 тыс.	**ПК с б/у RTX 3090 24 ГБ** (160–180 тыс. в сборе)	**OpenCode / Cline / Aider** + `qwen3-coder:30b` (30–50 ток./с)
Дешёвый GPU-эксперимент	**RTX 2080 Ti 22 ГБ мод** (40–55 тыс.) или **V100 32 ГБ PCIe** (65–78 тыс.)	**OpenCode / Aider** + `qwen3-coder:30b` (после жёстких стресс-тестов карты)
Сервер на команду / большие модели	**AI Max+ 395 128 ГБ** (299 тыс.), **2×3090 + NVLink** (320–360 тыс.) или **M2 Ultra 128 ГБ б/у** (450–555 тыс.)	**vLLM / Ollama** в локальной сети + `qwen3-coder-next` (80B) или 70B Q4
Бизнес и продакшен-разработка	**Гибрид:** локальный узел 3090 / Mac Studio + облачный API	Локально — NDA, данные и рутина; облако — сложная архитектура

Запуск за 15 минут

1. Поставьте **Ollama** с официального сайта и включите сжатие KV-кэша в окружении: `OLLAMA_KV_CACHE_TYPE=q8_0`, `OLLAMA_FLASH_ATTENTION=1`.
2. Скачайте кодер-модель под ваш объём памяти: `ollama pull qwen3-coder:30b` (для 24–48 ГБ) или `ollama pull qwen3.5:9b` (для 12–16 ГБ).
3. Для работы в терминале запустите **OpenCode** напрямую через Ollama: `ollama launch opencode --model qwen3-coder:30b` (либо укажите локальный endpoint `http://localhost:11434/v1`).
4. Для работы в VS Code поставьте расширение **Cline** или **Kilo Code**, выберите провайдер *Ollama*, укажите модель и увеличьте контекст до 32K–64K токенов.
5. Если у вас **2× RTX 3090 с мостом NVLink**, вместо Ollama поднимайте модель через **vLLM** с флагом `--tensor-parallel-size 2` — так обе карты будут считать каждый токен параллельно.

Нужна помощь с подбором или внедрением? Мы в **айфи.студио** собираем и настраиваем локальные и гибридные ИИ-контуры под задачи разработки и бизнеса: от выбора железа до дообучения моделей под ваши процессы.
Сайт: **aify.studio** · Telegram: **@aify_studio** · Прямой контакт: **@aify_tech** · Почта: **sales@aify.studio**

Срез цен: Москва и РФ, 3–4 ноября 2026 (Авито, DNS, Citilink, Ozon, BestMac, G.Race, i-store, re-store, Megabait). Курс ЦБ РФ — 80,97 ₽/$. Цены на б/у рынке указаны по живым объявлениям и меняются от состояния конкретного экземпляра. Скорость генерации зависит от длины контекста и настроек движка.