Как собрать своего ИИ-агента в VS Code

Пошагово для тех, кто никогда этого не делал · ВВК IT · ноябрь 2026

> **Что получите:** помощника на своём компьютере, который сам открывает файлы в папке, ведёт таблицы, разбирает документы и готовит черновики писем. Вы пишете задачу обычными словами и проверяете результат.
> **Сколько стоит:** 0 ₽. Все программы бесплатные, интернет нужен только для скачивания.
> **Сколько времени:** 1–2 часа в первый раз, большая часть — ожидание загрузки.
> **Куда уходят данные:** никуда. Всё работает на вашем компьютере, поэтому подходит даже для договоров и цен.

Сначала пять слов, без которых непонятно

ИИ-агент	ИИ, который не только отвечает в чате, но и сам работает с файлами: открывает, создаёт, исправляет. Перед каждым действием спрашивает разрешение.
VS Code	Бесплатная программа от Microsoft. Похожа на проводник с редактором: слева папки и файлы, справа — их содержимое. В неё и «поселим» агента.
Расширение	Дополнение к VS Code, как приложение для телефона. Агент ставится как расширение.
Терминал	Окно, куда вставляют команды текстом. Понадобится два раза: скопировать команду и нажать Enter.
Модель	«Мозг» агента. Мы скачаем бесплатную модель Qwen, она понимает русский.

Проверьте компьютер

Нужно **от 16 ГБ оперативной памяти** и около 10 ГБ свободного места на диске.

- **Windows:** нажмите Ctrl+Shift+Esc → вкладка «Производительность» → «Память». Цифра справа вверху — объём памяти.
- **Mac:** меню с яблоком → «Об этом Mac» → строка «Память».

Если памяти 8 ГБ, агент запустится, но будет медленным и часто ошибаться. Лучше начать с варианта из раздела «Дальше».

Установка: 5 шагов

> Если VS Code у вас на русском, названия меню указаны в скобках. Если окно узкое и строки меню наверху нет — она спрятана под значком ≡ в левом верхнем углу.

Шаг 1. Установите VS Code

Откройте code.visualstudio.com, нажмите большую синюю кнопку Download и установите как обычную программу. На все вопросы установщика можно отвечать «Далее».

Что увидите: окно с надписью Welcome. Его можно закрыть.

Шаг 2. Установите Ollama — программу, которая запускает модель

Откройте ollama.com/download, выберите свою систему, скачайте и установите.

Что увидите: окно Ollama с полем для сообщений и значок ламы рядом с часами (Windows) или в верхней строке экрана (Mac). На Windows значок часто спрятан: нажмите маленькую стрелку ^ слева от часов.

Сразу сделайте одну важную настройку. В окне Ollama откройте настройки (Settings) и найдите ползунок **Context length** — сколько текста модель читает за раз. Поставьте **32k**. Агенту Cline нужно много места под служебную инструкцию, и со значением по умолчанию (4k) он будет отвечать невпопад или молчать. Если компьютер после этого сильно тормозит — поставьте 16k.

Окно Ollama потом можно закрыть: программа продолжит работать у часов.

Шаг 3. Скачайте модель

В VS Code откройте меню **Terminal → New Terminal** (Терминал → Создать терминал). Внизу появится тёмная панель с мигающим курсором. Вставьте в неё команду и нажмите Enter:

```
ollama pull qwen3:8b
```

Что увидите: полоску загрузки примерно на 5 ГБ. В конце появится слово success. Это займёт от 10 минут до часа.

Если ошибка «ollama не является командой» или «command not found»: закройте VS Code полностью и откройте снова. Не помогло — перезагрузите компьютер.

Шаг 4. Поставьте агента Cline

Нажмите **Ctrl+Shift+X** (на Mac — Cmd+Shift+X). Слева откроется магазин расширений. В строке поиска наберите **Cline** и нажмите Install у расширения с самым большим числом установок.

Что увидите: на узкой панели значков у левого края VS Code появится новый значок Cline. Нажмите на него — откроется окно агента.

Шаг 5. Подключите модель к агенту

При первом запуске Cline предложит войти в аккаунт. Входить не нужно: выберите вариант со своим ключом (Use your own API key) или откройте настройки по значку шестерёнки. Дальше:

- **API Provider** — выберите **Ollama**;
- **Base URL** — оставьте как есть (`http://localhost:11434`, это адрес Ollama на вашем же компьютере);
- **Model** — выберите или впишите `qwen3:8b`.

Сохраните настройки (кнопка Done или Save).

Что увидите: окно чата Cline с полем для ввода задачи.

Если модели нет в списке: проверьте, что значок Ollama горит рядом с часами, и что шаг 3 закончился словом success.

Входить в аккаунт Cline не нужно. В его настройках можно отключить отправку анонимной статистики использования (Telemetry) — тогда наружу не уходит ничего.

Названия кнопок в новых версиях Cline могут немного отличаться, но логика та же: провайдер Ollama, модель qwen3:8b.

Первая задача: проверьте, что всё работает

1. Создайте рабочую папку. На рабочем столе создайте папку, например «Агент-закупка». В VS Code: **File → Open Folder** (Файл → Открыть папку) → выберите её. Если спросят «Доверяете ли вы авторам?» — нажмите Yes.

2. Напишите агенту в окне Cline:

```
Создай в этой папке файл suppliers.csv с колонками: Компания, Товар, Цена, Валюта, Срок поставки. Разделяй колонки точкой с запятой. Добавь три строки-примера.
```

Подождите. На обычном ноутбуке первый ответ может занять несколько минут — модель загружается в память. Ждите до 10 минут, прежде чем решать, что всё зависло. Дальше ответы пойдут быстрее.

3. Подтвердите действие. Агент покажет, какой файл хочет создать, и кнопки Approve (разрешить) и Reject (отклонить). Нажмите Approve.

Что увидите: слева в списке файлов появится suppliers.csv. Кликните по нему — он откроется прямо в VS Code.

Как открыть в Excel без кракозябр: откройте пустой Excel → вкладка «Данные» → «Из текста/CSV» → выберите файл → в поле «Источник файла» выберите «65001: Юникод (UTF-8)», разделитель — точка с запятой → «Загрузить».

Агент только описывает план, но ничего не делает? Внизу окна Cline есть переключатель Plan / Act. В режиме Plan агент лишь планирует — переключите на Act и повторите задачу.

Получилось — агент работает. Дальше дайте ему инструкцию, чтобы он помнил вашу задачу.

Не сработало? План Б

> **1. Проверьте связь с моделью.** Напишите агенту: `Ответь одним словом: работаешь?`
>
> **Ответа нет или ошибка** — агент не видит модель. Проверьте значок Ollama у часов и повторите шаг 5.
>
> **Отвечает пусто, невпопад или пересказывает чужую инструкцию** — почти всегда виновата настройка Context length. Проверьте, что в настройках Ollama стоит 32k (шаг 2), перезапустите Ollama и повторите задачу.
>
> **Ответил, но файл создать не смог** — связь есть, просто модель слабовата для работы с файлами. Повторите задачу ещё раз, короче. Если у вас 32 ГБ памяти и больше, скачайте модель покрупнее командой `ollama pull qwen3:14b` и выберите её в шаге 5. Если памяти меньше — смотрите раздел «Дальше».

Инструкция для агента: чтобы не объяснять каждый раз

Инструкция — это обычный текстовый файл. Агент читает его перед каждой задачей, как сотрудник — должностную инструкцию.

Как создать файл: сначала откройте список файлов — значок «Проводник» (два листа) в самом верху левой панели или Ctrl+Shift+E. Наведите мышь на название папки и нажмите значок листа с плюсом (New File). Впишите имя `.clinerules/agent.md` ровно так, с точкой и косой чертой, и нажмите Enter. VS Code сам создаст папку и файл внутри неё. Вставьте в файл шаблон, замените текст в квадратных скобках и сохраните: Ctrl+S (на Mac — Cmd+S).

```
# Агент: [название, например «Закупка»]

## Роль
Ты — мой помощник по задаче: [например, вести базу поставщиков и готовить сравнение цен].
Я — [имя, должность, компания]. Работаешь проактивно: сам предлагаешь следующий шаг.

## Файлы в папке
- suppliers.csv — база поставщиков, ты её ведёшь
- prices/ — прайсы, которые я присылаю (если папки нет — создай)
- notes.md — твоя память: итоги каждого дня (если файла нет — создай)

## Правила
1. Новые данные сам вносишь в базу, дубли объединяешь.
2. Пустые поля не выдумываешь — пишешь «нет данных».
3. Цены всегда с валютой и количеством.
4. Письма готовишь на [китайском и английском] с переводом на русский. Отправляю я сам.
5. В конце дня дописываешь в notes.md: что сделано, что ждём, что завтра.

## Стиль
По-русски, коротко, сначала вывод. Непроверенное помечаешь «не подтверждено».
```

Память агента — файл notes.md. В конце дня напишите агенту: «Запиши итоги в notes.md». Завтра он прочитает файл и продолжит с того же места.

Новая задача — новая папка со своей инструкцией. Так постепенно появляется команда: агент закупки, агент продаж, агент контента.

Если что-то пошло не так

Что случилось	Что сделать
Агент очень долго думает	Первый ответ может идти несколько минут, ждите до 10 минут. Дальше будет быстрее. Закройте лишние программы и вкладки браузера.
Отвечает невпопад или бросает задачу	Дробите задачу: одно действие за раз. Маленькая модель хуже справляется с длинными поручениями.
Пишет, что не видит файл	Проверьте, что в VS Code открыта нужная папка (File → Open Folder) и файл лежит внутри неё.
Ошибка подключения к Ollama	Запустите Ollama вручную из меню «Пуск» или «Программы» и повторите задачу.

Пять правил безопасной работы

1. **Читайте, что агент хочет сделать, прежде чем нажать Approve.** Особенно в первые недели.
2. **Проверяйте цифры.** ИИ ошибается в ценах, телефонах и расчётах. Всё, что идёт в КП или договор, сверяйте с источником.
3. **Отправляете вы, а не агент.** Письма он готовит, отправку оставьте за собой.
4. **Одна папка — одна задача.** Смешаете закупку и бухгалтерию — агент начнёт путаться.
5. **Начните с одной рутины.** Когда она заработает, добавляйте следующую.

Дальше: когда захочется умнее и быстрее

Бесплатная модель на ноутбуке — хороший старт, но она слабее облачных. Есть два пути вверх.

Путь 1. Облачная модель

В том же Cline в шаге 5 вместо Ollama выбирают облачного провайдера: Qwen, Claude, Gemini или GPT. Инструкция и папки остаются те же. У каждого есть и свой агент:

Нейросеть	Свой агент	Файл с инструкцией
Qwen (Alibaba)	Qwen Code	QWEN.md
Claude (Anthropic)	Claude Code	CLAUDE.md
Gemini (Google)	расширение Gemini Code Assist или Gemini CLI для терминала	GEMINI.md
GPT (OpenAI)	Codex: расширение для VS Code или программа для терминала	AGENTS.md

Что важно знать до оплаты:

- Облачные модели платные: нужна подписка или API-ключ — пароль доступа, за который списывают деньги по мере работы.
- Claude, Gemini и ChatGPT официально не работают в России, а оплатить Alibaba Cloud напрямую сложно. Обычно ключ берут у российских посредников с оплатой в рублях — для компании выбирайте того, кто даёт договор и закрывающие документы. Или оформляют аккаунт на компанию либо сотрудника в стране, где сервис доступен.
- Данные о стране и плательщике указывайте честно: аккаунт с недостоверными данными могут заблокировать вместе с оплаченным балансом.
- Облачному агенту не отдавайте договоры, цены и данные клиентов — для них остаётся локальный.

Путь 2. Свой сервер для всей компании

Если агент нужен нескольким сотрудникам, с доступом к документам, почте или CRM, а данные нельзя выпускать наружу, модель ставят на отдельный сервер с видеокартой (GPU). Тогда локальный агент работает на уровне облачных, а файлы не покидают компанию. Какой сервер нужен, зависит от задачи и числа пользователей — это лучше рассчитать заранее.

> Как эти агенты работают у нас каждый день, рассказываю в Telegram-канале «Прораб ИИ»: t.me/invest_pro_biz.
>
> Если нужен агент для всей компании на своём сервере — напишите мне в Telegram: **@ronining**. Подберём модель и железо под вашу задачу.
>
> Дмитрий Баев, BBK IT · bbk-it.pro

Источники: Cline и Ollama в VS Code · Ollama: длина контекста · Правила Cline · Быстрый старт Qwen Code · Claude Code в VS Code