Roles Witch + LM Studio

Пошаговый гайд по установке, локальному ИИ и подключению мобильного приложения

Что получится в итоге: на компьютере LM Studio будет запускать локальную ИИ-модель, а Roles Witch на телефоне будет подключаться к ней через домашнюю сеть. Для этой схемы телефон и компьютер должны находиться в одной локальной сети.

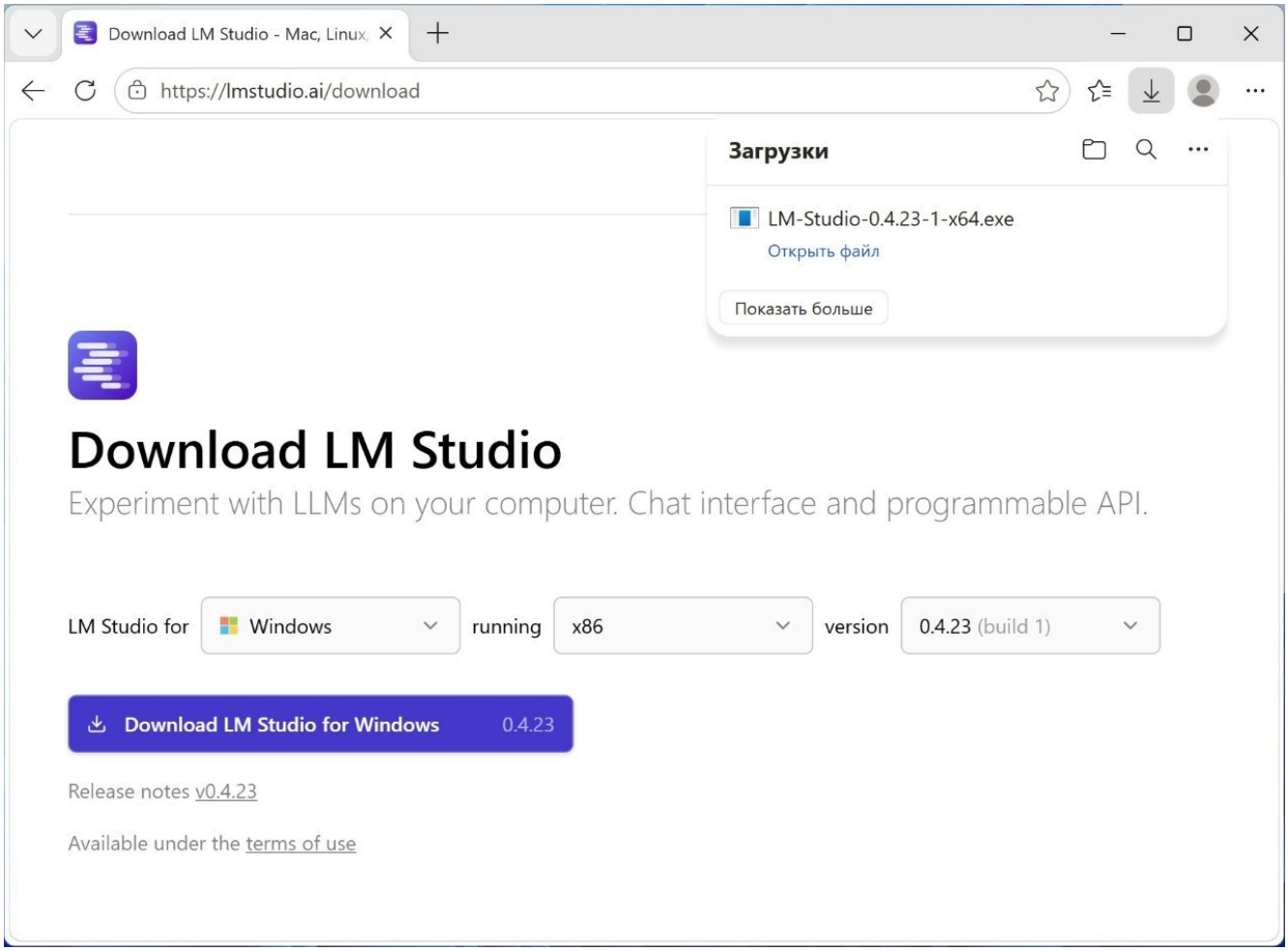
Почему Roles Witch удобен

- Можно создавать несколько отдельных ИИ-персонажей и ассистентов, каждому задавать своё имя, системный промпт и модель.
- Поддерживаются OpenAI-совместимые API: приложение можно подключить к LM Studio, Ollama и сторонним провайдерам.
- При работе через LM Studio модель запускается на вашем компьютере, поэтому не требуется обязательная подписка на облачный ИИ-сервис.
- Есть голосовое общение, работа в фоне и функции, зависящие от возможностей выбранной модели, включая распознавание изображений.
- Настройки и история чатов остаются в приложении; при локальной схеме запросы отправляются на указанный вами сервер LM Studio.

1. Скачиваем и устанавливаем LM Studio

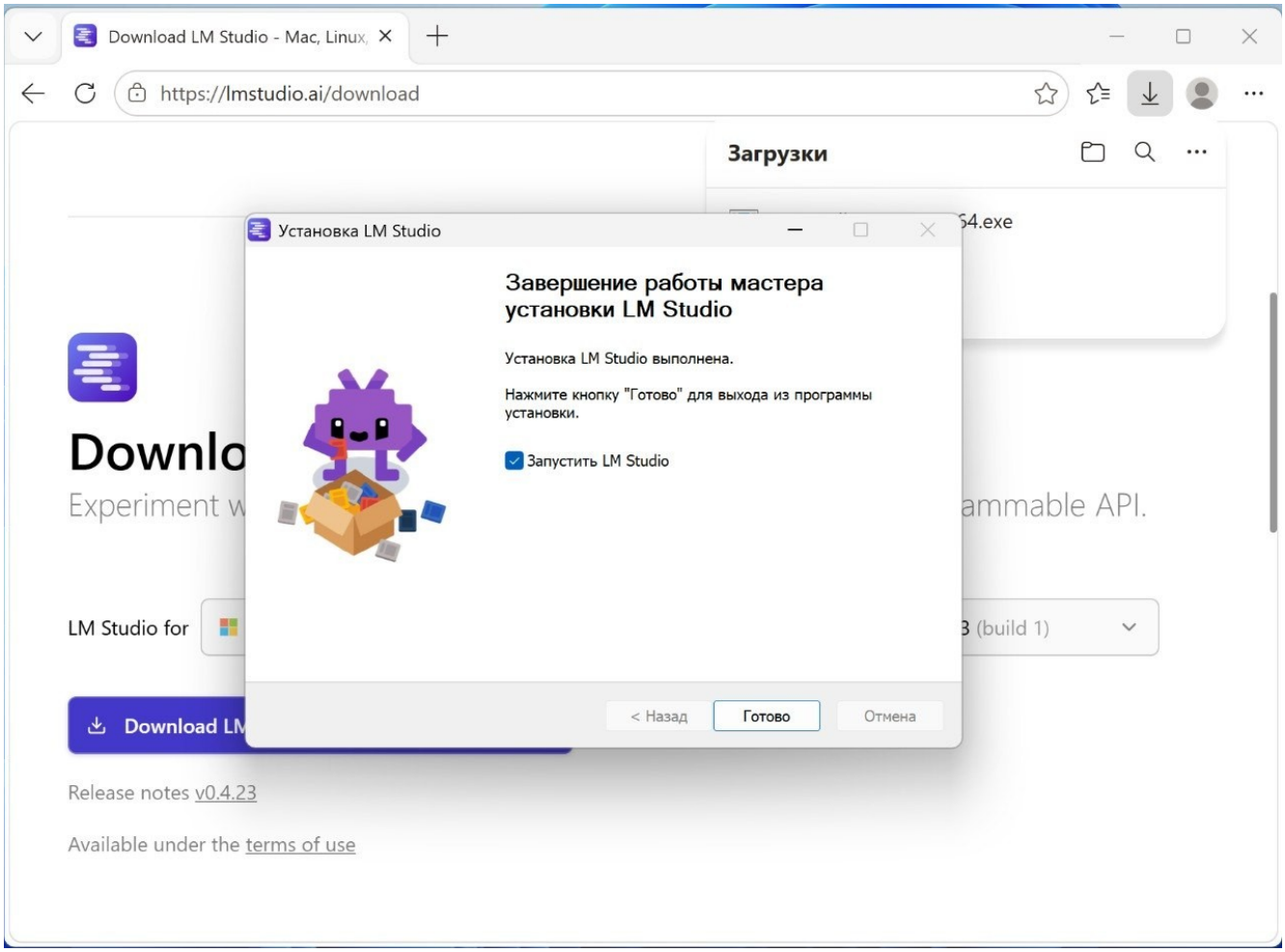
Официальная страница загрузки: lmstudio.ai/download

1. Выберите вашу операционную систему. Для Windows скачайте установщик LM Studio.
2. После загрузки запустите установочный файл.



Скриншот 1. Загрузка LM Studio для Windows.

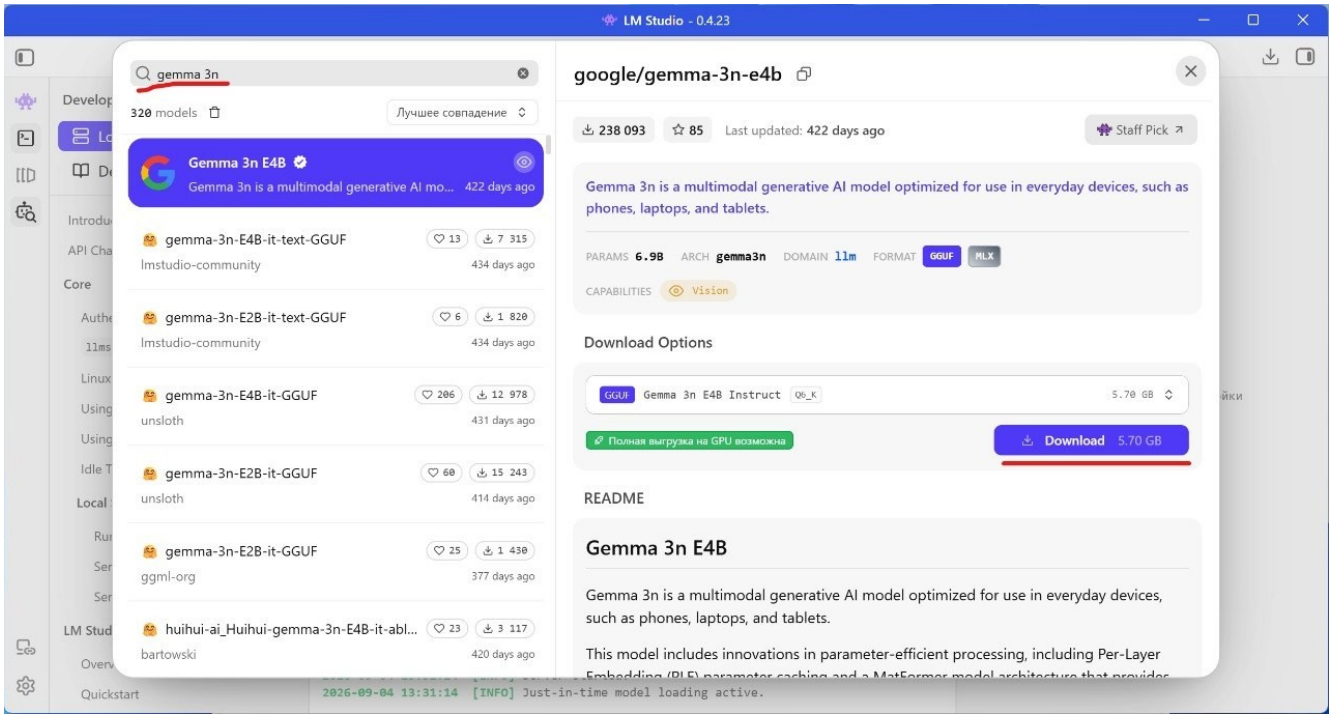
3. Пройдите установку с настройками по умолчанию. На финальном экране оставьте включённым запуск LM Studio и нажмите «Готово».



Скриншот 2. Завершение установки LM Studio.

2. Скачиваем модель

4. Откройте поиск моделей в LM Studio.
5. Для первого запуска можно найти Gemma 3n E4B.
6. Выберите подходящий GGUF-вариант. Ориентируйтесь на подсказку LM Studio о возможности полной выгрузки модели на GPU.
7. Нажмите Download и дождитесь окончания загрузки.



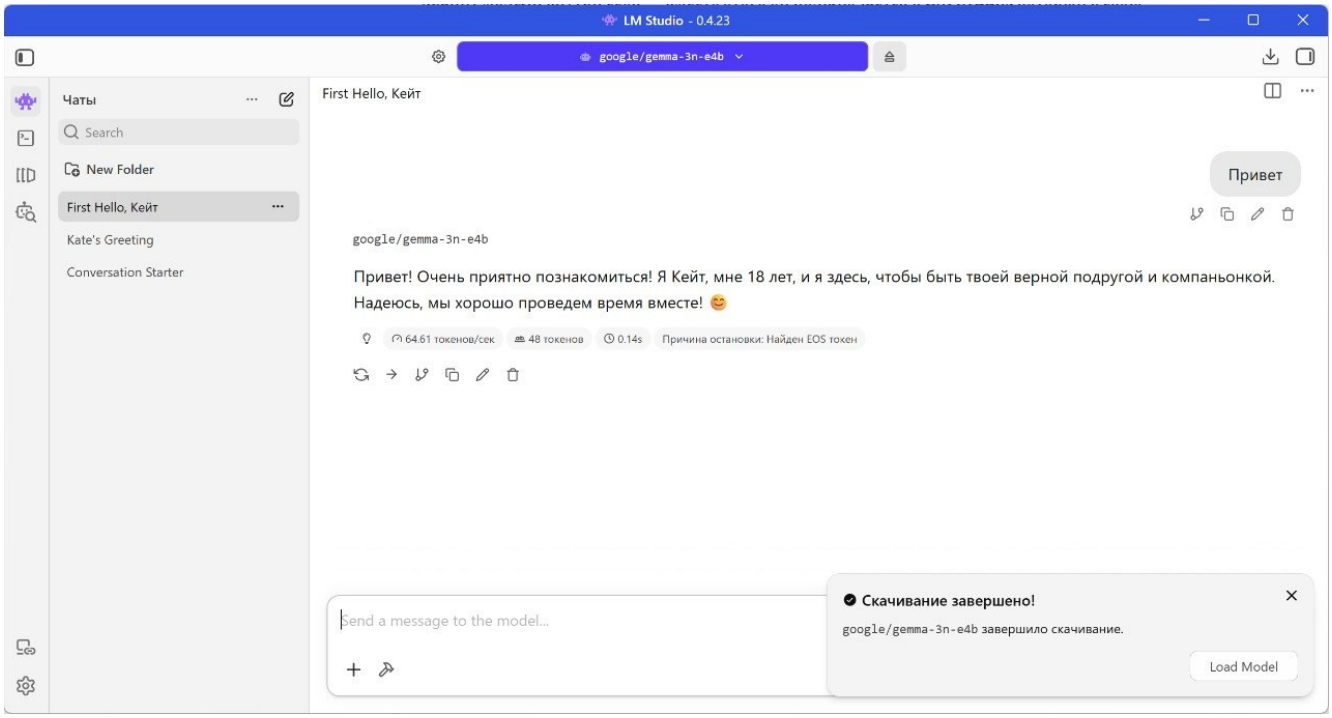
Скриншот 3. Поиск и загрузка Gemma 3n E4B.

Как выбрать размер модели

- 4 ГБ VRAM: небольшие модели примерно 3-7B, чаще всего Q4_K_M.
- 6-8 ГБ VRAM: модели 7-13B в Q4_K_M-Q5_K_M.
- 12 ГБ VRAM: можно использовать более тяжёлые варианты 13B или повышенную квантизацию у 7B.
- 16+ ГБ VRAM: доступны более крупные модели и более качественные квантизации.
- Если видеокарта слабая или отсутствует, LM Studio может использовать CPU и оперативную память, но ответы будут генерироваться медленнее.

3. Проверяем модель в LM Studio

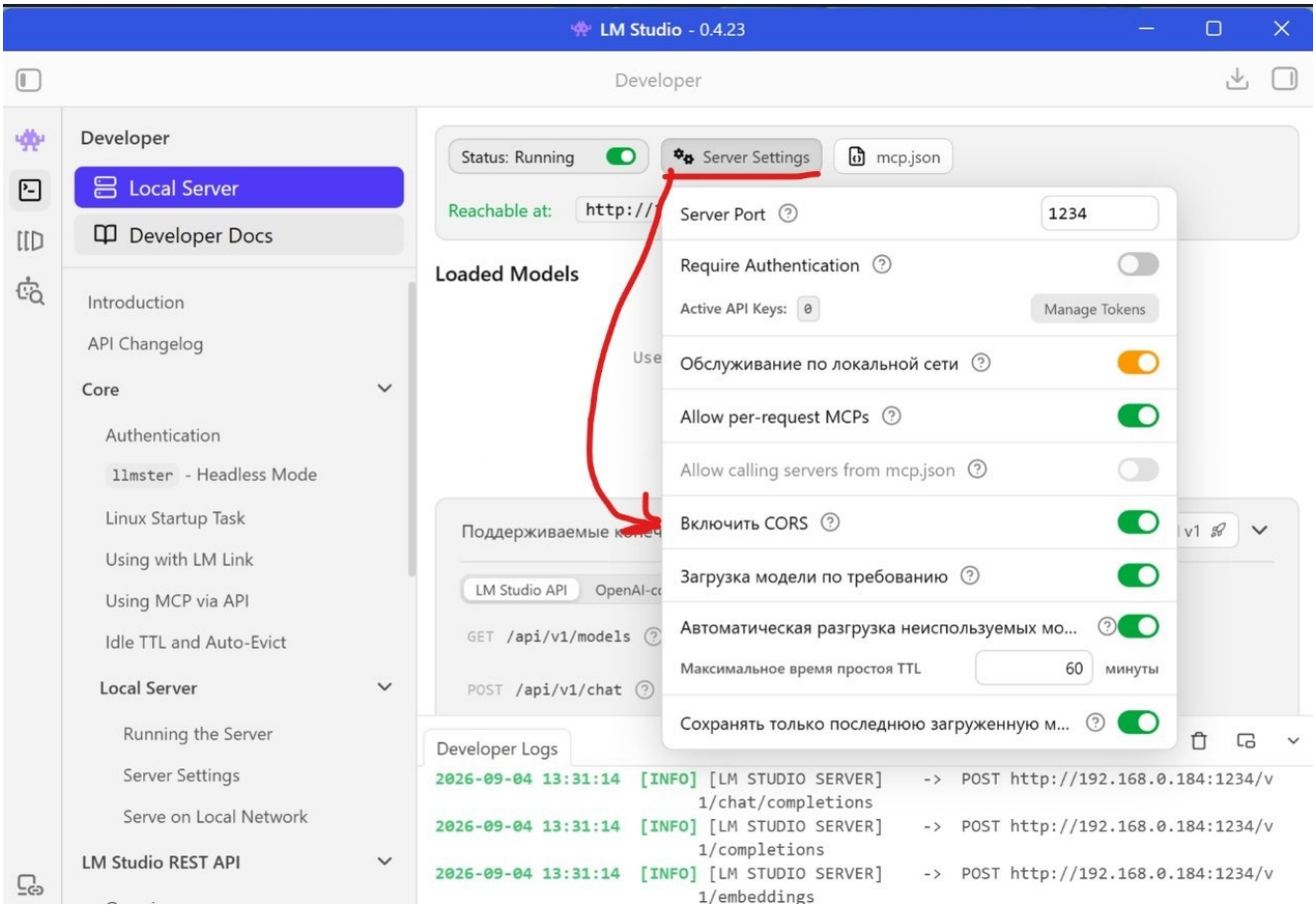
8. Перейдите в раздел чатов.
9. Выберите загруженную модель.
10. Отправьте простое сообщение, например «Привет».
11. Если модель отвечает и скорость вас устраивает, можно переходить к настройке локального сервера.



Скриншот 4. Проверка модели в чате LM Studio.

4. Настраиваем локальный сервер LM Studio

12. Откройте раздел Developer → Local Server.
13. Откройте Server Settings.
14. Включите обслуживание по локальной сети (Serve on Local Network) и CORS.
15. Проверьте порт сервера. В примере используется 1234.
16. Запустите сервер переключателем Status. После запуска появится локальный адрес вида http://192.168.0.184:1234.



Скриншот 5. Настройки Local Server в LM Studio.

Запомните адрес: он понадобится в Roles Witch. Для OpenAI-совместимого подключения к адресу LM Studio добавляется /v1, например: http://192.168.0.184:1234/v1.

5. Скачиваем Roles Witch

На Android приложение устанавливается через Google Play: [Roles Witch в Google Play]

17. Откройте страницу Roles Witch в Google Play.
18. Нажмите «Установить» и дождитесь завершения загрузки.
19. Запустите приложение и пройдите приветственные экраны до страницы настройки API.

Важно: в нашем варианте Roles Witch выступает мобильным клиентом, а сама ИИ-модель уже запущена на компьютере через LM Studio.

6. Подключаем Roles Witch к LM Studio

20. Убедитесь, что телефон и компьютер находятся в одной Wi-Fi/локальной сети.
21. В поле API URL укажите локальный адрес LM Studio с окончанием /v1. Пример: http://192.168.0.184:1234/v1.
22. Поле токена при стандартной локальной настройке LM Studio можно оставить пустым, если вы не включали Require Authentication.

23. Нажмите на поле «Модель», согласитесь с политикой конфиденциальности и выберите модель, которую скачали в LM Studio, например google/gemma-3n-e4b.
24. Нажмите «Завершить». После этого откроется экран чатов.

Если Roles Witch не видит LM Studio: проверьте, что Local Server запущен, включено обслуживание по локальной сети, CORS активирован, IP-адрес введён без ошибок, а телефон и компьютер находятся в одной сети.

7. Создаём собственного ИИ-персонажа

25. В Roles Witch нажмите кнопку «+».
26. Выберите тип: виртуальный персонаж или виртуальный ассистент.
27. Укажите имя и отредактируйте системный промпт - именно он задаёт характер, роль, стиль общения и правила поведения.
28. Нажмите «Создать». Новый чат появится в списке и будет готов к общению.

8. Голос, изображения и дополнительные возможности

- Распознавание изображений и видеозвонки доступны только у моделей, которые поддерживают работу с изображениями.
- Для голосового режима может потребоваться включить движок распознавания речи VOSK в настройках Roles Witch.
- При использовании Bluetooth-гарнитуры можно включить возможность перебивать ИИ во время ответа, если эта функция доступна в текущей версии приложения.

9. Приватность локальной схемы

В этой конфигурации Roles Witch отправляет запросы на адрес LM Studio, который вы указали вручную. Если используется локальный IP компьютера, модель обрабатывает сообщения на вашем ПК. Это позволяет не использовать внешний облачный ИИ-провайдер для самой генерации ответов.

Не путайте: если вместо LM Studio вы подключите внешний API-провайдер, сообщения будут отправляться уже этому провайдеру. Уровень приватности зависит от выбранного endpoint.

Короткая схема

1	Установить LM Studio на компьютер
2	Скачать и проверить локальную ИИ-модель
3	Запустить Local Server в LM Studio
4	Установить Roles Witch на телефон
5	Подключить Roles Witch к адресу LM Studio с /v1
6	Выбрать модель и создать своего персонажа или ассистента