ЛОКАЛЬНЫЙ ИИ

для творчества на Windows



Савелий Орлов

Савелий Орлов

Локальный ИИ для творчества на Windows

«Автор»
2026

Орлов С.

Локальный ИИ для творчества на Windows / С. Орлов — «Автор», 2026

Как превратить компьютер с Windows в творческую студию с локальным ИИ? Этот самоучитель проведёт вас от идеи до готового ролика. Вы подготовите сценарий, озвучите текст, сделаете субтитры, создадите музыку и звуковые эффекты, оживите кадр и соберёте всё в монтаже. Отдельные главы посвящены созданию 3D-объектов и их доработке в Blender.В книге разобраны LM Studio, Piper, Whisper, ACE-Step и ComfyUI — от скачивания моделей до экспорта готовых файлов. Пошаговые инструкции, схемы и практикумы помогут выбрать инструменты под возможности вашего ПК, проверить работу без интернета и решить, какие этапы удобнее выполнять локально, а для каких пригодится облачный сервис.

© Орлов С., 2026
© Автор, 2026

Содержание

Навигатор по задачам 5
Вступление 7
Быстрый старт 9
Часть I. Выбор моделей и работа с текстом 11
Глава 1. Какие задачи можно перенести на свой компьютер 11
Глава 2. Первая локальная LLM под Windows 17
Глава 3. Текст, который можно использовать в проекте 20
Практикум I. От выбора модели до покадрового листа 23
Часть II. Голос, клонирование и субтитры 24
Глава 4. Озвучка русского текста 24
Глава 5. Свой голос и длинная озвучка 28
Конец ознакомительного фрагмента. 31

Савелий Орлов
Локальный ИИ для творчества на Windows

Навигатор по задачам

Выберите результат, который нужен вам сейчас, и откройте соответствующую главу. У каждой задачи свой маршрут установки. Если вы впервые запускаете модель на компьютере, начните с раздела «Быстрый старт». За один сеанс вы пройдёте путь от загрузки программы до сохранения первого текста.

Написать или отредактировать текст

Начните с главы 2, чтобы запустить языковую модель. В главе 3 вы сможете подготовить черновик, доработать реплики и разбить сценарий на кадры. Результат можно сохранить в TXT или перенести в редактор, в котором вы ведёте рукопись. Если у вас ноутбук со встроенной графикой, первым делом посмотрите маршрут с компактной моделью для центрального процессора.

Озвучить текст или подготовить субтитры

В главе 4 вы получите озвучку на русском языке и сохраните её в WAV. В главе 5 разберётесь, как подготовить запись собственного голоса или запись, которую вам разрешили использовать, и как заменить отдельную реплику в длинной озвучке. Если голос уже записан и нужны только субтитры, переходите сразу к главе 6. Итогом будет отдельный файл SRT. При работе без дискретной видеокарты ищите в требованиях к выбранному инструменту режим для центрального процессора.

Сделать подложку, песню или звуковой эффект

В главе 7 вы подготовите инструментальную дорожку, а в главе 8 получите песню со своим текстом. Глава 9 посвящена коротким звуковым эффектам и их месту в монтаже. Готовый звук сохраняйте в WAV, чтобы затем открыть его в редакторе. На карте с 4–6 ГБ видеопамяти сначала сверяйте требования конкретной сборки к памяти. При 8 ГБ также учитывайте, какие компоненты программы остаются загруженными одновременно. Для звуковых эффектов сначала установите сам ComfyUI по главе 10. Архив и распакованная папка вместе займут около 7,7 ГБ. Веса FLUX для этого маршрута не нужны.

Получить иллюстрацию или исправить изображение

Откройте главу 10, если хотите получить изображение по своему описанию. В главе 11 вы сможете использовать опорное изображение, или референс, чтобы задать композицию и подготовить серию кадров. В главе 12 разберётесь, как изменить отдельный участок изображения и увеличить готовый кадр. Результат сохраняется в PNG. При ограниченной видеопамяти начните с одного кадра и подберите размер по требованиям маршрута.

Сделать короткое видео

Если у вас уже есть исходный кадр, начните с главы 13. Для генерации по тексту и настройки движения откройте главу 14. Подрезку и экспорт в MP4 вы найдёте в главе 15. У видеомодели сверяйте требования вместе с разрешением и числом кадров. Для карты с 8 ГБ видеопамяти смотрите, указан ли подходящий режим для вашей серии видеокарт. Если модель выходит за пределы возможностей компьютера, в главе 15 есть способ создать движение камеры по статичному изображению в редакторе.

Получить 3D-модель из картинки

Начните с главы 16, чтобы получить меш, то есть трёхмерную сетку, и осмотреть его со всех сторон. В главе 17 вы доработаете геометрию и материалы в Blender. Глава 18 поможет подготовить объект для игры или рендера и проверить экспорт GLB. Для выбора маршрута сверяйте требования к геометрии и автоматическому текстурированию по отдельности. У этих операций могут быть разные требования к видеопамяти. Рабочую сцену сохраняйте в BLEND, а GLB используйте для передачи объекта в другую программу.

Разобраться с файлами, памятью и сбоями

Если пока не знаете, сколько памяти у вашего компьютера и какая видеокарта выполняет вычисления, начните с главы 1. Когда нужно перенести результаты или снова открыть проект с прежними настройками, поможет глава 19. При сбое запуска откройте главу 20. Если вы подбираете замену конкретной подписке, глава 21 поможет сравнить качество, время до пригодного результата и объём ручной правки.

Вступление

Генерация на собственном компьютере начинается с того же вопроса, что и работа в онлайн-сервисе. Что вы хотите получить и что будете делать с результатом дальше? Ответ определяет выбор инструмента. Чтобы вычитать черновик, нужна языковая модель. Для озвучки этого черновика понадобится модель речи. Если озвучка уже есть и к ней нужно добавить субтитры, задача меняется ещё раз. Теперь программе предстоит распознать речь и привязать строки к моментам звучания.

Когда генерация идёт локально, вычисления выполняет ваш компьютер. Интернет пригодится для загрузки программы, модели и её компонентов. Возможность работать без сети зависит от выбранного режима. В одном интерфейсе могут соседствовать локальная генерация и функции, которые отправляют данные на сервер. В инструкциях книги учитывается весь путь от исходного материала до сохранённого файла.

В этом пути участвуют модель, её веса и программа запуска. Модель определяет, какие данные инструмент принимает и какой вид результата может вернуть. Веса — сохранённые параметры, которые модель получила при обучении. Программа запуска загружает эти параметры, принимает ваш запрос и показывает или сохраняет результат. Набор настроек задаёт условия работы. От него могут зависеть размер изображения, длительность фрагмента, расход памяти и время ожидания.

На схеме 1 видно, за что отвечает каждый из четырёх компонентов.



Схема 1. Модель, веса, программа и настройки

Удобно подбирать инструменты по отдельным операциям. Если вам нужен только голос, имеет смысл оценить, сколько времени занимает вся озвучка с исправлениями и сохранением файла. Если цель — ассет для игры, важен и следующий шаг. Откройте экспорт в своей рабочей программе и посмотрите, как перенеслись масштаб, геометрия и материалы. Чем раньше вы проверите этот переход, тем проще будет понять, какой этап требует доработки.

Время одной генерации даёт представление о скорости компьютера. Для своей работы полезно также засечь время до дубля, который вы готовы использовать. В него входят повторные попытки, исправление дефектов и экспорт. По этому времени можно судить, какую часть работы удобно перенести на свой компьютер и когда для конкретной задачи вам по-прежнему подходит подписочный сервис.

Все главы рассчитаны на ваши материалы. Вы сами выбираете, какой текст отредактировать, какой кадр оживить и какой объект превратить в меш. После генерации откройте сохранённый файл там, где будете с ним работать. После этой проверки проще решить, оставить результат или вернуться к настройкам и исходному материалу.

Быстрый старт

Первый текст на своём компьютере

Вы установите LM Studio, скачаете языковую модель Qwen3-4B и сохраните её ответ в текстовый файл. Подготовьте короткий фрагмент своей работы, который хотите отредактировать, или сформулируйте задачу для нового текста.

Порядок первого запуска LM Studio показан на схеме 2.

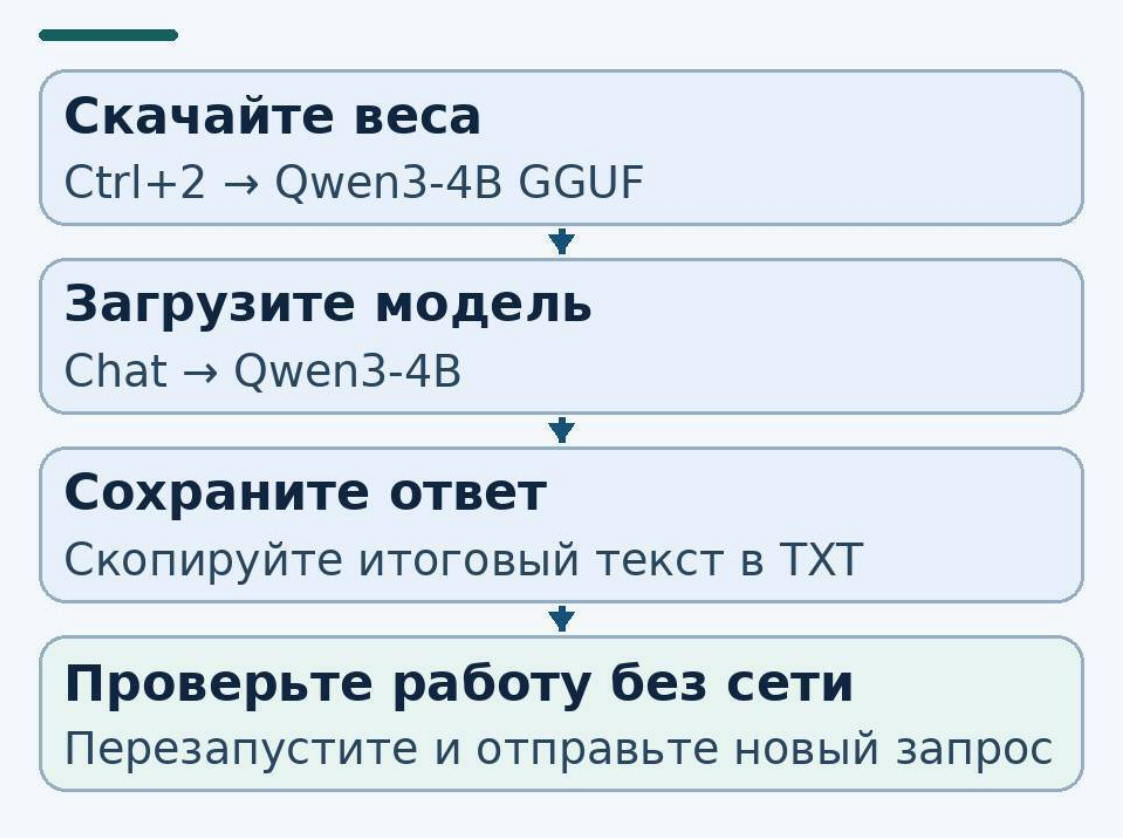

Схема 2. Первый локальный ответ в LM Studio

Этот маршрут рассчитан на Windows 11 x64. Для x64 нужны процессор с AVX2 и запас оперативной памяти. Разработчик рекомендует 16 ГБ. Название процессора и поддержку AVX2 можно сверить по главе 2. Для файла модели потребуется 2,5 ГБ на диске. Оставьте дополнительное место для установщика, программы и движка. Размер файла модели показывает, сколько нужно скачать, а расход памяти при запуске зависит ещё и от настроек. Системные требования LM Studio [1]

Скачайте LM Studio. На странице загрузки найдите блок **Download LM Studio**. Выберите Windows и x64. Для версии 0.4.29, сборки 2, файл называется LM-Studio-0.4.29-2-x64.exe. Сохраните его в папку «Загрузки» и дождитесь окончания скачивания. Если браузер сохранил установщик в другом месте, откройте список загрузок браузера и выберите действие «Показать в папке» рядом с этим файлом. Страница загрузки LM Studio [2] · Описание версии 0.4.29 [3] Если главная страница уже предлагает другую версию, возьмите сборку 2 по прямому

адресу разработчика. Установщик занимает около 655 МБ. Точный установщик LM Studio для Windows x64 [4]

Установите и откройте программу. В Проводнике откройте «Загрузки», дважды щёлкните по скачанному файлу и пройдите установку. Если окно приложения не открылось после установки, найдите **LM Studio** в меню «Пуск». Пока оставьте интернет включённым. Программа может предложить загрузить движок, который выполняет вычисления. Для этой модели нужен движок на базе llama.cpp с поддержкой GGUF. Первый запуск LM Studio [5] Если при первом запуске приложение предложит уровень интерфейса или стартовую модель, выберите простой режим и переходите к поиску модели из следующего шага. Загрузка предложенной модели необязательна.

Загрузите веса модели. В LM Studio нажмите **Ctrl+2** и вставьте в поле поиска Qwen/Qwen3-4B-GGUF. Откройте результат от **Qwen**. В списке файлов выберите Qwen3-4B-Q4_K_M.gguf размером 2,5 ГБ и нажмите **Download**. GGUF — формат файла модели, а Q4_K_M — обозначение варианта весов. Скачанный GGUF готов к загрузке в программу. Дождитесь, пока модель появится в **My Models**. Там же можно увидеть или изменить папку, в которой хранятся модели. Веса опубликованы под лицензией Apache 2.0. Поиск и загрузка в LM Studio [6] · Файл в репозитории Qwen [7]

Загрузите модель в чат. Откройте **Chat** и нажмите на поле выбора модели вверху окна. Выберите скачанную Qwen3-4B и дождитесь окончания загрузки в память. Если программа сообщит о нехватке памяти, вернитесь к настройкам загрузки и уменьшите **Context Length** до 2048 для первого короткого запроса. Контекст определяет, сколько текста модель может учитывать за раз. Если ошибка указывает именно на видеопамять, уменьшите **GPU Offload** и повторите загрузку. Эта настройка задаёт, какая часть слоёв модели размещается на видеокарте. При нулевом значении слои загружаются в оперативную память и вычисляются на процессоре. Загрузка модели в чат [5]

Отправьте свой запрос. Укажите, что нужно сделать, на каком языке ответить и какой объём текста вам подходит. Если редактируете фрагмент, вставьте его в сообщение и уточните, что следует сохранить при правке. Отправьте запрос и дождитесь завершения ответа. Прочитайте текст целиком. При редактировании сопоставьте его с исходником и проверьте, сохранились ли ваши факты и смысл.

Сохраните ответ в отдельный файл. После генерации выделите итоговый ответ и скопируйте его в Блокнот. Qwen3-4B может показать рассуждение перед ответом. Оставьте в TXT только текст, который идёт после этого блока. В папке «Документы» создайте папку «Локальный ИИ» и сохраните текст в ней в формате TXT с кодировкой UTF-8. Закройте Блокнот и снова откройте файл из Проводника. Убедитесь, что сохранился весь ответ и русские буквы отображаются правильно.

Повторите запуск без интернета. Закройте LM Studio, отключите интернет и снова откройте программу. Начните новый чат с той же локальной моделью, отключите сетевой поиск и внешние инструменты, если они включены, и отправьте новый короткий запрос. Если в чате появился новый ответ, программа смогла загрузить модель и запустить её с уже скачанными компонентами. Если при загрузке не хватает движка, подключитесь к сети, завершите его установку и повторите проверку. Работа LM Studio без интернета [8]

Настройки запуска и хранение моделей подробнее разобраны в главе 2, работа со своим текстом — в главе 3. За озвучкой переходите к главе 4, за музыкой — к главе 7, за изображениями — к главе 10, за видео — к главе 13, за 3D — к главе 16.

Часть I. Выбор моделей и работа с текстом

Глава 1. Какие задачи можно перенести на свой компьютер

Прежде чем скачивать модель, определите, на каком этапе она вам нужна и в какой программе вы откроете результат. От этого зависит, что считать удачной генерацией. У текста, который пойдёт в рукопись, важны смысл и редактура. У звука для монтажа — разборчивая речь, чистые границы реплик и сохранение в WAV. У трёхмерного объекта — форма со всех сторон и возможность перенести его в редактор с материалами. Так вы подбираете инструмент под конкретную работу и заранее знаете, что проверять после запуска.

Какие функции искать

Текст. Если вы пользуетесь ChatGPT или Claude для черновиков, правки и работы со структурой, ищите языковую модель с поддержкой диалога. В главе 2 этот маршрут начинается с Qwen3-4B в LM Studio. Для решения о замене важна работа с вашим русским текстом: сохраняются ли факты при правке, соблюдает ли модель ограничения и можно ли продолжить диалог от нужного места. Поиск в интернете и работа с файлами могут требовать дополнительных функций программы.

Голос и субтитры. Озвучка, воспроизведение тембра по записи и расшифровка речи — три разные задачи. Если вам в ElevenLabs нужна озвучка, смотрите поддержку русского языка, настройки голоса, возможность исправить отдельную реплику и экспорт в WAV. Для клонирования проверьте, может ли модель использовать запись, которую вам разрешили применять. Субтитры требуют модели распознавания речи и инструмента, который сохраняет текст с таймкодами в SRT. Эти маршруты разобраны в главах 4–6.

Музыка и эффекты. Если вы обращаетесь к Suno за песней, выясните, умеет ли локальная модель принимать ваш текст и насколько разборчиво она его поёт. Если вам нужна подложка, проверьте инструментальный режим, возможность управлять длительностью и удобство монтажа под голос. Для коротких звуковых эффектов смотрите отдельные модели. В главах 7–9 вы подготовите дорожки для звукового редактора.

Изображения. Для работы, которую вы привыкли делать в Midjourney, отдельно оценивайте генерацию по описанию, работу с референсом и правку существующего кадра. Модель, которая делает удачный одиночный кадр, может плохо сохранять одни и те же детали в серии. В главах 10–12 вы получите PNG, проверите композицию и сможете определить, сколько работы остаётся после генерации.

Видео. Если в Kling, Runway или Higgsfield вы оживляете готовые кадры, начните с направления «из изображения в видео». Для генерации по одному описанию ищите поддержку «из текста в видео». При выборе учитывайте, сколько кадров нужно для монтажа и насколько важны форма объекта и управляемость движения. В главах 13–15 вы оцените короткий план, обрежете его при необходимости и сохраните в MP4. При длинной очереди дублей удобство облачного сервиса можно сравнить с временем, которое ваш ПК тратит на получение пригодного клипа.

3D. Если в Tripo AI или Meshy вам нужна геометрия из изображения, выясните, умеет ли локальная система экспортировать меш. Если важна готовая поверхность объекта, проверяйте текстурирование отдельно. В главах 16–18 вы осмотрите форму, доведёте сетку и материалы в Blender и проверите экспорт GLB. Для игровых ассетов в сравнение стоит включить

ручную правку сетки и подготовку масштаба. Эти этапы влияют на общую трудоёмкость даже при быстром первом запуске модели.

Из чего складывается инструмент

Модель принимает данные определённого типа и выполняет задачу, на которую рассчитана. Чтобы получить ответ в диалоге, нужна текстовая модель с подходящим режимом работы. Чтобы получить звук из текста, нужна модель синтеза речи. Возможности конкретной модели описывает её разработчик.

Веса — числовые параметры модели, сохранённые в одном или нескольких файлах. Размер этих файлов определяет, сколько места на диске займут веса. У одной модели могут быть выпуски с разным форматом и объёмом весов. При выборе обязательно смотрите точное имя файла. Совместимость и расход памяти зависят от конкретного выпуска.

Программа запуска даёт вам интерфейс для работы с моделью. LM Studio позволяет выбрать языковую модель и открыть чат. ComfyUI позволяет связать несколько операций в рабочую схему. Внутри программы вычисления выполняет движок. В документации его способ работы с конкретным устройством может называться **backend**. От движка и поддерживаемого им режима зависит, какие файлы он читает и на каком устройстве их запускает.

Настройки определяют, сколько данных загрузить и с какими параметрами их обработать. При генерации видео на расход памяти влияют размеры кадра и число кадров. В текстовом чате имеет значение длина контекста, то есть объём текста, с которым модель работает в данный момент. Смена настроек способна изменить и расход памяти, и качество результата, даже если веса остаются теми же.

Как узнать конфигурацию Windows

Откройте **Пуск → Параметры → Система → О системе**. В разделе «Характеристики устройства» найдите название процессора, установленную оперативную память и тип системы. В разделе «Характеристики Windows» посмотрите выпуск, версию и сборку ОС. Сохраните эти сведения. При загрузке программы вариант **x64** выбирают для 64-разрядной Windows с процессором архитектуры x64. Для ARM используются отдельные сборки, там, где их выпускает разработчик. Как найти сведения об устройстве в Windows [9]

Нажмите **Ctrl+Shift+Esc**, чтобы открыть Диспетчер задач. На вкладке **Производительность** выберите **Память**. Запишите установленный объём и значение «Доступно». Первое число показывает, сколько оперативной памяти установлено, а второе — сколько памяти Windows может предоставить приложениям в данный момент. Чтобы значение было полезно при выборе модели, смотрите его при открытых программах, которыми обычно пользуетесь во время работы. Числа по RAM в главах о ComfyUI — ориентиры по установленной памяти для одной короткой задачи на GPU под Windows x64 при закрытых ресурсоёмких программах. Фактический расход зависит от размера задания, видеопамяти и объёма данных, который ComfyUI переносит в RAM. Чтобы оценить запас на своём ПК, следите за значением «Доступно» при загрузке модели, генерации и сохранении результата.

На той же вкладке откройте разделы **GPU**. Windows нумерует устройства. Запишите название каждого и найдите строку «Выделенная память графического процессора». Для дискретной видеокарты это ориентир по собственной видеопамяти, или **VRAM**. Общая память графического процессора выделяется из системной RAM. Поэтому складывать оба значения и считать сумму объёмом собственной памяти карты нельзя. Если перед вами отношение занятой памяти к предельной, для оценки свободной видеопамяти вычтите текущее значение из

предельного. На странице GPU также доступна версия драйвера. Как Диспетчер задач показывает GPU и их память [10]

На схеме 3 показано, какие два показателя искать на странице GPU.



Схема 3. Где посмотреть VRAM и RAM

На компьютере с несколькими GPU номер карты в списке сам по себе не говорит, какая делает вычисления. Во время генерации откройте вкладку **Подробности** в Диспетчере задач. Щёлкните правой кнопкой по заголовку столбца, откройте выбор столбцов и включите отображение GPU и его ядра, если они скрыты. Сопоставьте номер у работающего процесса с номером карты на вкладке **Производительность**. Если приложение запускает отдельный процесс движка, смотрите его активность. Показатель ядра GPU отражает наиболее загруженное ядро процесса в момент наблюдения. Для подтверждения режима сверяйте его с настройками программы и изменением нагрузки во время обработки. Как читать данные о видеокарте в Диспетчере задач [10]

Теперь откройте **Проводник → Этот компьютер** и посмотрите, сколько места свободно на нужных дисках. Учитывайте отдельно диск, куда скачивается установщик, и диск, где программа будет хранить веса, временные файлы и результаты. Если загрузка идёт архивом, во время распаковки оригинальный архив и извлечённые файлы одновременно занимают место. Сравнивайте с требованиями размер загрузки, объём после установки и расход оперативной памяти при работе. Это три разные величины. Как проверить свободное место в Windows [11] Размеры загрузок в книге приведены в десятичных единицах. Один ГБ равен 1000 МБ. Windows может показывать то же количество байтов другим числом.

Как сверить драйвер NVIDIA. Для NVIDIA откройте «Пуск», найдите Windows PowerShell и выполните команду из любой папки.

nvidia-smi

Найдите свою карту в списке и запишите Driver Version. У NVIDIA это короткая версия вида 5xx.xx. Для сравнения с таблицами CUDA используйте её вместо длинной версии Windows из Диспетчера задач. Поле CUDA Version показывает максимальную версию, поддерживаемую драйвером. Версия библиотек, встроенных в архив приложения, указана отдельно. При ошибке запуска утилиты посмотрите сведения о драйвере в панели управления NVIDIA или приложении NVIDIA. Справка NVIDIA System Management Interface [12]

С какого маршрута начать на вашем ПК

Без дискретной видеокарты. Начните с текста на CPU по главе 2. Для голоса берите Piper из главы 4, для субтитров — whisper.cpp из главы 6. Эти два инструмента имеют CPU-маршруты и короткие проверки на Linux. Для клонирования по записи начните с одной реплики Qwen3-TTS из главы 5.

4–6 ГБ VRAM. Сначала выполните те же задачи на CPU. Для одной иллюстрации оцените, достаточно ли RAM и места для FLUX из главы 10. Загрузка описанного графа на такой карте зависит от выгрузки в оперативную память. Число 4–6 ГБ само по себе готовность к генерации не определяет.

8 ГБ VRAM. Добавляйте генерацию кадров и 3D-формы по одному маршруту за раз. Для монтажа готового кадра или клипа подходит глава 15. У RTX 2080 есть ограничения по типам вычислений, которые важны для музыки и видео. Они разобраны в следующем разделе.

12–16 ГБ VRAM и выше. Выбирайте короткое видео из главы 13 или 14 и начните с одного дубля. Для автотекстурирования сначала сверьте требования именно к материалам. Для Paint 2.1 Tencent указывает 21 ГБ VRAM, поэтому ручной маршрут Blender из главы 17 остаётся пригодным и на компьютерах с меньшей видеопамятью.

Первые проверки на RTX 2080

RTX 2080 относится к поколению Turing с вычислительной возможностью CUDA 7.5. У карты есть аппаратное ускорение FP16, но нет аппаратного ускорения BF16 и FP8. Обычная CUDA-сборка FlashAttention 2 ориентирована на Ampere и новее. Файл с весами FP8 может работать через преобразование, если его поддерживает движок. Это зависит от сборки и модели. Видеокарты и вычислительная возможность CUDA [13] Типы данных Tensor Core [14] Требования FlashAttention 2 [15]

Проверьте короткую задачу после установки подходящего драйвера. В окне запуска ComfyUI найдите строку Device с названием RTX 2080. После Run сохраните строки журнала с dtype и manual cast, если они появились. В главе 7 режим CUDA в этой сборке ACE-Step выбирает BF16. Для первого опыта на Turing предусмотрен CPU-режим с float32, время которого нужно оценить отдельно.

Считайте проверку успешной, когда на выбранном устройстве появился новый файл и он открылся в следующей программе. Для FLUX это PNG заданного размера, для LTX или Wan — короткий MP4 из 49 кадров, для Hunyuan3D — GLB, который импортируется в Blender. Запишите драйвер, время и пики выделенной VRAM и RAM.

Почему генерация может резко замедлиться

При offload приложение само переносит данные в RAM. Кроме того, на Windows драйвер NVIDIA может использовать системную память при нехватке VRAM, если это позволяет его

политика CUDA – Sysmem Fallback Policy. Следите и за выделенной, и за общей памятью GPU. Если оперативная память исчерпана, Windows может использовать файл подкачки на диске, что замедляет расчёт. Файл подкачки не даёт запаса выделенной VRAM. Начните с уменьшения нагрузки и сверяйте расход RAM и VRAM, прежде чем менять системные настройки. Что означает Sysmem Fallback у NVIDIA [16]

Что посмотреть на странице модели

Начните с сайта программы и страницы разработчика модели. Оттуда переходите к опубликованным релизам, репозиторию или карточке весов. Поисковая выдача и каталоги помогают найти варианты. Автора, версию и условия использования лучше сверять в исходном репозитории.

На Hugging Face текстовое описание находится во вкладке **Model card**. В нём ищите назначение модели, поддерживаемые языки, ограничения и рекомендации по запуску. Во вкладке **Files and versions** найдите требуемый файл и его размер. Сверьте, что ссылка ведёт к нужному выпуску и владельцу репозитория. У программы и весов могут быть разные лицензии. Откройте условия для того выпуска, который собираетесь использовать, и сохраните ссылку вместе с названием модели. Что содержит карточка модели на Hugging Face [17]

Прежде чем скачивать, сопоставьте описание модели с документацией программы запуска. Проверьте формат весов, операционную систему, архитектуру, поддерживаемые GPU или CPU и версию движка. Если нужны вспомогательные файлы, учитывайте каждый из них. Когда в карточке приводят расход памяти или скорость, посмотрите, при каких настройках и на каком устройстве их получили. По числу без условий сложно понять, повторится ли результат на вашей машине.

Как узнать, где выполняется генерация

В ComfyUI рабочая схема, или **workflow**, состоит из узлов. Каждый узел выполняет операцию: загружает данные, запускает обработку или сохраняет файл. Часть узлов выполняет вычисления на компьютере, часть может вызывать внешний сервис. Партнёрские узлы **Partner Nodes** обращаются к внешним API. Для их запуска через Comfy требуется учётная запись и оплачиваемые кредиты. Партнёрские узлы в документации ComfyUI [18]

Перед запуском посмотрите документацию узлов, выполняющих генерацию и обработку. Ищите сведения о том, откуда загружаются веса, какой узел использует ключ API, какие материалы он передаёт и требуется ли платёж или вход в аккаунт. Проверяйте цепочку до самого экспорта. Наличие локальной модели в начале маршрута не определяет устройство, на котором выполняются последующие операции.

Для проверки автономности сначала скачайте нужные компоненты и завершите первый запуск. Затем закройте программу, отключите интернет, снова запустите её и измените входной материал, запрос или seed и запустите новый расчёт. Сохраните результат под новым именем, откройте его и убедитесь, что изменения видны или слышны. Так вы проверите конкретный маршрут вместе с его зависимостями и экспортом. Если выполнение остановилось, найдите шаг, на котором понадобилась сеть: от него будет зависеть, что скачать заранее или какую часть схемы нужно заменить, чтобы работать автономно.

Перед выбором модели

Сохраните описание своего компьютера в текстовом файле в папке «Документы». Укажите версию и сборку Windows, архитектуру, процессор, установленную и доступную RAM, название GPU, выделенную видеопамять и свободное место на выбранном диске. Добавьте туда название одной операции, которую хотите перенести на свой ПК, и формат ожидаемого результата. Запись готова, если по ней можно выбрать правильную сборку программы и сравнить требования модели с возможностями компьютера ещё до загрузки весов.

Глава 2. Первая локальная LLM под Windows

В «Быстром старте» вы получили первый ответ. Теперь настроим LM Studio так, чтобы к этому же тексту можно было вернуться после перезапуска, продолжить работу без сети и перенести модель на другой диск при необходимости. Если вы открыли главу сразу, начните с установки. Если Qwen3-4B уже отвечает в чате, переходите к разделу «Как подобрать параметры загрузки».

Для работы приготовьте короткий фрагмент своего текста или сформулируйте одну задачу. Поставьте себе понятную цель: к концу главы у вас должны быть сохранены исходный запрос и ответ, который можно открыть в редакторе. Для скачивания понадобится интернет. Обычный текстовый чат с уже установленным движком и скачанной моделью может работать без сети. Автономная работа LM Studio [8]

Установка программы

Маршрут описан для Windows 11 x64 и LM Studio 0.4.29, сборки 2. На странице загрузки выберите **Windows** и **x64** в блоке **Download LM Studio**. Установщик этой сборки называется LM-Studio-0.4.29-2-x64.exe. Кнопка на сайте ведёт к актуальной версии, поэтому имя файла может измениться. Версию уже установленной программы можно сопоставить с журналом выпусков. Загрузить LM Studio для Windows [2] · Журнал выпуска 0.4.29 [3]

Прежде чем открыть установщик, сверьте тип системы в **Параметры → Система → О системе**. Для сборки x64 процессор должен поддерживать AVX2. LM Studio рекомендует от 16 ГБ оперативной памяти, а для работы на видеокарте — от 4 ГБ выделенной видеопамяти. Эти рекомендации относятся к программе в целом. Фактический расход зависит от файла модели, длины контекста и устройства, на котором выполняются вычисления. Системные требования LM Studio [1]

Сохраните установщик в «Загрузках», запустите его из Проводника, завершите установку и откройте **LM Studio** через меню «Пуск». Держите соединение с интернетом, пока программа найдёт и скачает модель и необходимые компоненты. После установки открывается интерфейс LM Studio. Весами и движком LM Studio управляет отдельно, поэтому при первом запуске они могут загружаться в разное время. Установка и первый запуск LM Studio [5] · Загрузка движков [8]

Найдите точное название процессора в Параметры → Система → О системе. Откройте его страницу на сайте Intel или AMD и посмотрите наборы инструкций. Ищите именно AVX2. При отсутствии этого обозначения проверьте спецификацию точной модели, прежде чем выбирать сборку x64. Характеристики Intel [19] Характеристики AMD [20]

Выбор файла и папки для весов

В этой главе используется **Qwen3-4B** в варианте Qwen3-4B-Q4_K_M.gguf. Фрагмент Q4_K_M в имени обозначает вариант квантования. Квантование позволяет хранить параметры модели с меньшей точностью, чтобы уменьшить размер файла и потребность в памяти. Формат GGUF подходит для движка на базе llama.cpp. Репозиторий Qwen содержит несколько вариантов весов. Для этого маршрута достаточно одного указанного файла объёмом 2,5 ГБ. Это размер скачивания, а на диске дополнительно потребуется место для LM Studio, движка и временных данных. Лицензия весов — Apache 2.0. Карточка модели и список файлов Qwen [21]

Итого на диске. Около 3,16 ГБ на скачанный установщик и GGUF. Оставьте дополнительное место для установки, движка, кэша и экспорта. Этот объём зависит от выбранного движка.

Если на системном диске мало места, до скачивания откройте **My Models** и задайте папку для весов на другом диске. Узнайте её полный путь, чтобы позднее найти файл в Проводнике. Храните в этой папке модели. Для текстов, которые будете редактировать и экспортировать, используйте отдельную папку в «Документах». Управление папкой моделей [6]

Нажмите **Ctrl+2**, чтобы открыть поиск моделей, и введите Qwen/Qwen3-4B-GGUF. Откройте репозиторий от **Qwen** и выберите вариант **Q4_K_M**. Сверьте имя файла с официальной страницей, затем нажмите **Download**. Полоса загрузки должна дойти до конца. После этого модель появится в **My Models**. Если скачивание прервалось, проверьте свободное место именно на диске, где находится папка моделей, и повторите загрузку из этого же репозитория. Поиск и скачивание моделей в интерфейсе [6] · Страница файла Qwen3-4B-Q4_K_M.gguf [7]

LM Studio организует веса по издателю и названию репозитория. Если решите перенести уже скачанную модель, закройте программу и в Проводнике скопируйте папку издателя вместе с папкой модели в новую папку для весов. Сохраните вложенность. Затем откройте LM Studio, укажите новую папку в **My Models** и проверьте, видна ли модель и открывается ли чат. Старую копию можно удалить, когда убедитесь, что новая загрузилась с нужного диска. Как LM Studio организует папки с моделями [22]

Как подобрать параметры загрузки

Запишите отдельно версию LM Studio и название с версией движка. Приложение отвечает за интерфейс, а движок загружает GGUF и выполняет вычисления. Если результат изменился после обновления, эти две записи помогут проверить, какой компонент поменялся.

Откройте **Chat**. Вверху окна раскройте список моделей и выберите установленную Qwen3-4B. После выбора LM Studio помещает веса и рабочие данные в оперативную память и, в зависимости от настроек, в видеопамять. Дождитесь окончания загрузки. Если программа предложит загрузить движок, пока не отключайте интернет. Для этой модели нужен движок с поддержкой Qwen3 и формата GGUF. Как загрузить модель в чат [5] · Движки LM Studio [8]

Размер контекста задаёт настройка **Context Length**. Контекст измеряется в токенах, то есть единицах текста, на которые модель разбивает сообщения. В пределах этого объёма должны поместиться ваш запрос, используемая история диалога, служебные инструкции и формируемый ответ. Чем больше контекст, тем больше памяти может потребоваться. Начиная с LM Studio 0.4.16 приложение использует по умолчанию 8192 токена, если настройки модели не задали другое значение. При нехватке памяти уменьшите Context Length до 2048 для первого короткого запроса. Для длинной работы сначала сократите лишнюю историю или берите текст по фрагментам. Если информации всё равно мало и у компьютера есть запас памяти, увеличьте контекст и загрузите модель заново. Что означает длина контекста [23] Перед повторной загрузкой сверьте число, которое фактически показывает поле. Изменение контекста в LM Studio 0.4.16 [24]

Настройка **GPU Offload** определяет, сколько слоёв модели программа попытается разместить на видеокарте. Если возникает ошибка нехватки видеопамяти, уменьшите это значение и повторите загрузку. При снижении GPU Offload возрастает роль оперативной памяти и процессора. Если используете ноутбук со встроенной графикой, для запуска на центральном процессоре выберите доступный для Windows вариант движка llama.cpp для CPU, когда LM Studio предлагает выбор движка, и установите GPU Offload на ноль. Скорость оцените по своему короткому запросу. Как LM Studio управляет вычислительными движками [25] · Настройки переноса слоёв на GPU [26]

Параметры можно менять перед загрузкой. Чтобы сохранить их для следующих сеансов, откройте **My Models**, нажмите значок шестерёнки рядом с моделью и измените её настройки по умолчанию. Новые значения применяются при следующей загрузке. Если модель уже открыта в чате, выгрузите её из памяти и снова выберите в списке. После запуска посмотрите нагрузку в Диспетчере задач по инструкции из главы 1. Это поможет заметить, что вычисления ушли на CPU или на другой GPU, если скорость и расход памяти отличаются от ожидаемых. Параметры конкретной модели в LM Studio [27]

Первый ответ и продолжение диалога

Введите свой запрос в поле сообщения. Укажите действие, нужный язык ответа, желаемую форму и допустимый объём. Если редактируете, приложите свой фрагмент в том же сообщении и обозначьте, какие сведения и особенности стиля должны сохраниться. Отправьте запрос и дождитесь окончания генерации. В первом ответе оцените соответствие задаче и проверьте, не появились ли новые факты, которых не было в исходном тексте.

Если ответ надо уточнить, отправьте короткое продолжение в том же чате. Назовите конкретный участок и нужное изменение. Для другой задачи откройте новый чат. LM Studio хранит историю разговоров, а модель получает сообщения из текущего диалога в пределах доступного контекста. Если разговор стал длинным, перенесите в новый чат только исходный отрывок, текущую редакцию и те требования, которые ещё действуют. Как LM Studio хранит и использует чаты [28]

Если перед ответом появилось рассуждение, дождитесь конца генерации и копируйте текст после него. На первом запуске вид блока может зависеть от настроек отображения и шаблона чата. Режим рассуждений Qwen3-4B [29]

Создайте в папке «Документы» отдельную папку для работы с текстом. Скопируйте исходный запрос из чата в один файл TXT, а пригодный ответ — в другой. При сохранении в Блокноте используйте кодировку UTF-8. Если хотите позже повторить работу, рядом запишите версию LM Studio, имя GGUF-файла и длину контекста. Историю можно открыть и в самой программе. Чтобы увидеть место её хранения, щёлкните по чату правой кнопкой и выберите **Show in File Explorer**. LM Studio хранит разговоры в JSON. Для текущей работы проще пользоваться чатом и своими TXT-файлами. Расположение истории чатов [28]

Проверка после перезапуска без сети

Закройте LM Studio после того, как модель уже выдала ответ. Отключите интернет, снова откройте программу и загрузите ту же модель в **Chat**. Откройте прежний диалог и отправьте новое уточнение. Проверьте, что пришёл именно новый ответ и он относится к вашему тексту. Если LM Studio сообщает об отсутствующем движке, снова подключитесь к интернету, установите нужный компонент и повторите проверку. Поиск новых моделей и скачивание движков требуют сети. Что доступно в LM Studio без интернета [8]

Откройте сохранённые TXT-файлы из Проводника. У вас должны читаться и формулировка задачи, и текст, который получился в ответ. Проверка закончена, когда модель отвечает после перезапуска без сети, а результат открывается в вашем редакторе. В следующей главе эта же связка пригодится для работы с черновиком, репликами и покадровым листом.

Глава 3. Текст, который можно использовать в проекте

В предыдущей главе вы получили и сохранили ответ Qwen3-4B. Теперь возьмите материал, над которым действительно работаете. К концу главы у вас будет черновик сценария, отредактированная реплика и покадровый лист. Откройте **Chat** в LM Studio и загрузите модель из главы 2. Исходные заметки держите в отдельном файле, чтобы в любой момент сравнить с ними новую версию.

Из идеи в сценарий

Запишите, о чём вы хотите рассказать, для кого предназначен текст и что должно произойти к концу. Если у материала уже есть формат и ограничение по времени, добавьте их. Отдельно отметьте сведения, которые нельзя менять. Для художественного текста это могут быть события, мотивация и точка зрения персонажа. Для материала, который опирается на реальные сведения, сохраните даты, имена, числа и источники.

В новом чате попросите модель сначала разложить вашу идею по смысловым эпизодам. Эпизод можно описать через действие и его последствие. Посмотрите, чем закончился каждый и вызывает ли этот конец следующее действие. Если два эпизода повторяют одну мысль, объедините их. Если переход между ними требует события, которого в исходнике не было, сами решите, как его построить. После такой правки попросите развернуть утверждённый порядок в сценарий.

Если сценарий будут озвучивать, определите, какая информация должна прозвучать вслух и что будет видно на экране. Прочитайте черновик своим обычным темпом и засеките время. Если текст длиннее, чем позволяет ролик, сначала уберите повторяющиеся пояснения и оставьте сведения, без которых следующая мысль повиснет в воздухе. Длительность лучше проверять чтением или черновой озвучкой. Подсчёт слов годится для предварительной оценки, а темп, паузы и произношение меняют итоговый хронометраж.

Как отредактировать реплику и сохранить её голос

Прочитайте реплику вслух вместе с предыдущей и следующей. Определите, где именно она мешает сцене. Проблемой может быть сбивчивый ритм, лишнее объяснение или смысл, который слушатель понимает слишком поздно. Берите на один проход одну задачу. Отправьте модели саму реплику, короткий контекст и ограничения. Укажите, какие сведения, намерение говорящего и уровень разговорности нужно сохранить.

Затем прочтите исходную и новую реплики на одном месте в сцене. Отметьте, где новый текст легче произнести, и посмотрите, осталось ли у персонажа то же намерение. Если реплика стала короче, но исчезла причина следующего поступка, восстановите смысл до того, как приметесь за ритм. Если модель изменила манеру речи, укажите, какая черта потерялась, и верните ей на правку исходную реплику вместе с тем фрагментом диалога, где нужная манера слышна лучше всего.

Для одиночной правки достаточно местного контекста. Рядом с репликой дайте модели тот кусок, из которого следует, кто говорит, кому и с какой целью. Такой запрос проще проверить и легче уместить в контекст. Если надо проследить, как речь меняется от сцены к сцене, сначала выпишите несколько устойчивых признаков голоса персонажа в рабочую заметку. Используйте эту заметку для каждой сцены и после правки прочитайте сцены подряд.

Как сверить факты с исходником

Сначала сравните имена, даты, числа и единицы измерения. Затем проверьте утверждения о причине и следствии. Модель может добавить причинную связь там, где в исходнике описывалась только последовательность событий. Отдельного внимания требуют уверенные формулировки, появившиеся вместо предположений, и детали, добавленные ради связности.

Попросите модель отметить утверждения, которые она добавила при переработке. Это поможет быстрее пройти по черновику, однако каждое значимое утверждение всё равно нужно сверить самостоятельно. Держите исходную ссылку или заметку рядом с соответствующим абзацем. Для факта, которого нет в ваших материалах, найдите источник или уберите его. В художественном тексте так же проверяют непротиворечивость событий. Особое внимание уделите тому, что персонажи уже знают в этот момент и какие действия они могли совершить за отведённое время.

Как составить покадровый лист

Когда сценарий выдерживает чтение вслух, разделите его на участки, примерно соответствующие будущим планам. Поручите модели выделить моменты, для которых нужно отдельное изображение, движение или смена точки зрения. Для каждого плана укажите номер, ориентировочную длительность, видимое действие, соответствующую реплику или звук и то, что должно совпасть со следующим кадром. В текстовом файле удобнее отвести каждому плану отдельный абзац.

Пройдите по листу от первого плана к последнему. Проверьте, что суммарная длительность соответствует формату, у каждой реплики есть место, а последовательность действий можно понять по изображению. Если один план содержит несколько действий, разберите, какие должны остаться в нём и что лучше перенести в соседний. Для генерации кадров полезно держать общие параметры сцены в отдельной заметке. Тогда цвет, свет и элементы, которые должны повторяться, можно сверять по одной записи.

В описании сцены для художника или генератора изображений оставьте наблюдаемые признаки. Укажите, кто или что в кадре, где расположен объект, откуда на него смотрят, как он освещён и что находится позади. Если работаете по референсу, сначала разберите его сами. Запишите, какие свойства нужно перенести и какие можно изменить. В Qwen3-4B из главы 2 передайте эту разметку текстом. При работе с моделью, которая принимает изображения, проверьте её описание референса глазами перед тем, как превратить его в запрос для генератора. В окончательном запросе оставьте условия, относящиеся к одному кадру. Описание движения и изменения между кадрами пригодится отдельно, в видеочасти книги. Модель Qwen3-4B [30]

Как подготовить черновик куплета

Если вам нужен текст для музыкальной главы, используйте тот же подход в меньшем масштабе. Определите тему куплета, точку зрения и мысль, с которой должна закончиться последняя строка. Сообщите модели, какой объём вы готовы спеть и какую роль куплет играет относительно припева. Когда получите черновик, прочитайте его в предполагаемом темпе или пропойте под свою подложку. Исправьте места, где приходится ускорять произношение или переносить естественное ударение ради рифмы. Сохраните текст куплета отдельно от описания жанра, темпа и инструментов. Эти сведения понадобятся в разных полях при работе с музыкальной моделью.

Как решить, где доводить текст

Для коротких переписываний и обсуждения одного фрагмента удобно держать Qwen3-4B в локальном чате. Редактор текста удобнее для сравнения версий и точной ручной правки. Если работе нужно сопоставление нескольких источников, а их содержание не помещается в текущий контекст или требуется поиск в сети, используйте браузер для источников либо облачный инструмент, где такой режим доступен. Существенные утверждения проверяйте по найденным материалам до переноса в сценарий.

Если выбираете между локальным и облачным чатом для регулярной работы, дайте обоим одну и ту же собственную задачу с одинаковым фрагментом и ограничениями. Сравните число смысловых ошибок, объём ручных исправлений и время, за которое удалось получить версию для проекта. Результат сравнения будет относиться к выбранной задаче, версии модели и условиям запуска.

Сохраните сценарий и покадровый лист отдельными файлами. Откройте их в рабочем редакторе и сверьте, что после правок совпадают порядок действий, реплики и предполагаемый хронометраж. Чистовой текст, который должен прозвучать вслух, перенесите в отдельный файл. Он станет входом для следующей главы.

Практикум I. От выбора модели до покадрового листа

Определите конфигурацию компьютера. Сохраните версию Windows, архитектуру, название CPU, доступную RAM, модель GPU, выделенную VRAM и свободное место на диске. По записи должно быть понятно, какую сборку программы выбирать и куда скачивать веса.

Выберите файл модели. Найдите на странице издателя точное имя и размер совместимого файла. Сверьте формат весов с движком программы, а размер с местом на выбранном диске. Сохраните в TXT точное имя файла, его размер, формат и название движка. По этим четырём полям должно быть понятно, что и для какой программы вы скачиваете. Если формат расходится, вернитесь к разделу выбора весов в главе 2.

Получите первый текст. Задайте модели собственную короткую задачу и сохраните запрос и ответ в двух отдельных TXT-файлах. Откройте оба в редакторе. Русские буквы должны отображаться правильно, а по запросу должно быть понятно, что вы хотели получить.

Отредактируйте свою реплику. Укажите цель правки и смысл, который нужно сохранить. Прочитайте оба варианта в сцене. Задание выполнено, когда новая реплика решает указанную проблему, сохраняет нужные сведения и подходит говорящему.

Подготовьте покадровый лист. Разбейте свой сценарий на планы. У каждого должны быть номер, действие, длительность и отметка по звуку. Сверьте общую длительность со своим форматом и убедитесь, что можете по листу начать работу над кадрами.

Часть II. Голос, клонирование и субтитры

Глава 4. Озвучка русского текста

В этой главе вы превратите свой русский текст в звуковую дорожку WAV. Возьмите короткий абзац, который можно прочитать вслух целиком. Если у вас уже готов сценарий, выберите из него реплику с теми словами, на которых особенно важно проверить произношение. Сохраните авторский текст отдельно. Для озвучки вы будете работать с копией, где написание можно менять ради звучания.

Как выбрать голос и программу

Для первого запуска под Windows возьмём **Piper 1.4.2** и модель **Irina medium**. Piper — программа синтеза, а Irina — готовый женский голос для русского языка. Он рассчитан на одного диктора, поэтому отдельный номер голоса и переключатель языка здесь не нужны. Выбор русского языка происходит при выборе модели с кодом ru_RU. Работать будем на центральном процессоре. В этом маршруте не требуются дискретная видеокарта и CUDA. Piper 1.4.2 [31] · Карточка голоса Irina [32]

Если в ElevenLabs вы получаете озвучку набранного текста и затем монтируете её самостоятельно, у Piper есть нужная пара операций: прочитать текст и сохранить WAV. Готовый локальный голос имеет смысл оценивать по разборчивости, числу исправлений и тому, насколько он подходит вашему материалу. В ElevenLabs доступны разные модели и голоса. Это удобно, когда выбор подачи важнее автономной работы. Из локальных моделей **Qwen3-TTS CustomVoice** тоже предлагает несколько готовых тембров и заявляет поддержку русского языка. Версия **1.7B CustomVoice** дополнительно принимает инструкции по стилю. Какая связка лучше для вашего текста, определит прослушивание одного и того же фрагмента. Если понадобится озвучка по записи собственного голоса, это отдельная задача следующей главы. Синтез речи в ElevenLabs [33] · Варианты Qwen3-TTS [34] Для чтения текста готовым русским голосом начните с Piper. Base из главы 5 нужен, когда надо передать тембр по разрешённой записи. CustomVoice имеет смысл рассмотреть, если нужен выбор готовых тембров. В этой книге отдельный маршрут CustomVoice не разбирается.

Подготовка Windows

Для описанного маршрута нужна Windows 11 x64 и 64-разрядный Python 3.13. На странице Python 3.13.17 найдите **Windows installer (64-bit)**. Имя файла — python-3.13.17-amd64.exe, размер загрузки — 28,6 МБ. Запустите установщик и выберите **Install Now**. Сохраните выбранный по умолчанию компонент запуска Python, если установщик предлагает его установить. Он нужен для команды py. Если 64-разрядный Python 3.13 уже установлен и команда ниже показывает эту версию, продолжайте с ним. Загрузка Python 3.13.17 [35] · Запуск Python в Windows [36]

На Windows вычислительная библиотека ONNX Runtime требует среду Visual C++. Если её нет, установите Microsoft Visual C++ Redistributable для x64, файл vc_redist.x64.exe, по ссылке из документации Microsoft. Установщик Python, пакет Piper и голос скачиваются по отдельности. Windows-пакет Piper занимает 13,8 МБ, голос — примерно 63 МБ. Кроме них

установятся зависимости Python. Оставьте запас для установленных пакетов и звуковых файлов. Требования ONNX Runtime [37] · Пакет Visual C++ для Windows [38]

В папке «Документы» создайте папку **Piper** и откройте её в Проводнике. Щёлкните по адресной строке Проводника, введите powershell и нажмите **Enter**. Откроется PowerShell с рабочей папкой Piper. Её путь будет указан слева от курсора в строке после PS. Команды ниже выполняйте в этом окне. Вставляйте только строку команды и нажимайте **Enter**.

Для Piper закреплена ветка Python 3.13 x64. Пакет Piper 1.4.2 имеет совместимую Windows-сборку с ABI Python 3.9 и выше. Для воспроизведения этого маршрута используйте указанную версию Python и ссылку на её выпуск.

Итого на диске. Около 106 МБ на установщик Python, пакет Piper и файл голоса. Установленный Python, зависимости, файлы WAV и при необходимости Visual C++ Redistributable потребуют дополнительного места.

В PowerShell из папки **Piper** проверьте, что Windows видит установленный Python 3.13. В ответе должна появиться его версия.

```
py -3.13 --version
```

В том же PowerShell, из папки **Piper**, создайте отдельную среду для озвучки. После выполнения в рабочей папке появится каталог .venv. В него будут устанавливаться пакеты этой главы.

```
py -3.13 -m venv .venv
```

В PowerShell из папки **Piper** установите Piper и его зависимости. Установка загружает пакет piper_tts-1.4.2-cp39-abi3-win_amd64.whl для Windows x64 и нужные ему библиотеки. Дождитесь окончания установки и возврата к строке, в которую можно ввести следующую команду. Пакет Piper распространяется по GPL-3.0-or-later. Файлы выпуска Piper 1.4.2 [39]

```
.\.venv\Scripts\python.exe -m pip install "piper-tts==1.4.2"
```

В PowerShell из папки **Piper** откройте справку. Если появится описание параметров, включая --model, --input-file и --output-file, программа готова принимать команды. Если установка завершилась ошибкой, начните с проверки версии Python в этом окне и убедитесь, что путь в команде ведёт к python.exe внутри рабочей папки. При ошибке загрузки DLL для ONNX Runtime проверьте установку Visual C++ Redistributable x64. Командный интерфейс Piper [40]

```
.\.venv\Scripts\python.exe -m piper --help
```

Загрузка русского голоса

Создайте внутри папки **Piper** папку **models**. На странице голоса скачайте два файла: ru_RU-irina-medium.onnx и ru_RU-irina-medium.onnx.json. У первого объём около 63,2 МБ, у второго — около 4,8 КБ. Положите оба в **models** с исходными именами. ONNX содержит модель, а JSON — настройки, которые Piper загружает рядом с ней. Для обоих файлов используйте кнопку скачивания вместо сохранения страницы браузера. Если не видите расширения, включите их в Проводнике через **Вид → Показать → Расширения имён файлов**. Файлы голоса Irina medium [41]

В репозитории голосов указан MIT. В карточке Irina отдельно указано, что лицензия набора данных неизвестна. Это разные сведения об источнике. Храните название и ссылку на карточку вместе с материалами проекта, чтобы вернуться к ним при выборе голоса для публикации. Карточка голоса и сведения о данных [32]

Первый WAV

Откройте Блокнот, вставьте свой текст и сохраните его в папке **Piper** под именем text.txt с кодировкой **UTF-8**. Для первой проверки положите весь короткий абзац на одну строку файла.

Предложения разделите обычными знаками препинания. Заголовки, указания по монтажу и пометки для редактора оставьте в исходном документе. При синтезе они могут прозвучать вместе с репликой.

В PowerShell из папки **Piper** запустите следующую команду. Параметр -m указывает файл модели, -i — файл с вашим текстом, -f — имя результата. После возврата к строке ввода в папке **Piper** появится voice.wav. Если программа сообщает, что голос не найден или не может прочитать настройки, сначала убедитесь, что оба файла находятся в **models**, и сравните полные имена с теми, которые указаны на странице загрузки. Параметры запуска Piper [40]

.\.venv\Scripts\python.exe -m piper -m ".\models\ru_RU-irina-medium.onnx" -i ".\text.txt" -f ".\voice.wav"

Откройте WAV двойным щелчком из Проводника. Прослушайте начало, середину и конец. По карточке модели Irina выдаёт звук с частотой дискретизации 22 050 Гц. Если редактор использует другую частоту проекта, проверьте звук после импорта и слушайте результат на той дорожке, где он будет в монтаже. Параметры звука Irina [32]

Правка произношения

Числа и сокращения. Прослушайте, как голос читает даты, суммы, единицы измерения и сокращения. В копии для озвучки запишите проблемное число словами в том падеже, в котором оно звучит внутри предложения. Раскройте сокращение или передайте буквами именно то произношение, которое нужно услышать. После правки проверьте число вместе с соседним существительным и убедитесь, что величина и единица остались прежними.

Имена. Сначала выясните нужное произношение имени. Для слова на латинице попробуйте записать его русскими буквами по звучанию. Оценивайте имя внутри предложения: перед ним и после него могут стоять звуки, от которых зависит разборчивость. Если имя повторяется, используйте удачную запись в копии для озвучки одинаково. Правильное авторское написание сохраните в сценарии и субтитрах.

Ударение. Начните с обычной пунктуации и буквы «ё» там, где она нужна. Если слово допускает разные прочтения, попробуйте добавить естественный контекст или заменить его однозначным словом, если смысл позволяет. Для имени или термина, который требуется оставить, проверяйте варианты фонетической записи на одной короткой реплике. Оставьте вариант, который звучит верно в полной фразе. Когда нужное ударение не удаётся получить, озвучьте реплику отдельно другим подходящим голосом или запишите её сами, если это допустимо в монтаже.

Паузы. Сначала расставьте знаки препинания так, как должен звучать текст. Если между соседними предложениями нужна более заметная пауза, можно добавить параметр --sentence-silence. Значение задаётся в секундах. В Piper 1.4.2 добавочная тишина вставляется между предложениями внутри одной строки входного файла. Перенос строки в TXT сам по себе не задаёт паузу нужной длины. Точную паузу в одном месте удобнее выставить в звуковом редакторе между двумя отдельно сгенерированными фрагментами.

После правок сохраните свой текст в том же text.txt. В PowerShell из папки **Piper** запустите следующий вариант команды, если хотите добавить 0,2 секунды между предложениями. Новый файл voice-revised.wav появится рядом с первым. В имени файла сохранена отдельная версия, чтобы можно было прослушать оба дубля подряд.

.\.venv\Scripts\python.exe -m piper -m ".\models\ru_RU-irina-medium.onnx" -i ".\text.txt" -f ".\voice-revised.wav" --sentence-silence 0.2

Проверка результата без сети

Сохраните текст для озвучки рядом с выбранным WAV. Если вы хотите проверить автономность, закройте PowerShell, отключите интернет, снова откройте PowerShell из папки **Piper** и повторите команду генерации с новым именем WAV после -f. Откройте получившийся файл, чтобы убедиться, что Piper смог выполнить синтез после перезапуска. Для этого этапа используются уже установленные пакеты и два локальных файла голоса.

Импортируйте выбранный WAV в свою программу монтажа. Проверьте, что звук читается целиком, существенные имена и числа произносятся правильно, а паузы укладываются в хронометраж. При сравнении с другим голосом засеките время от сохранения текста до пригодного дубля. Включите в него перегенерацию исправленных реплик и правку пауз. Именно этот отрезок показывает, удобно ли вам использовать голос в следующем проекте.

Глава 5. Свой голос и длинная озвучка

В предыдущей главе Piper прочитал текст готовым голосом. Здесь исходной точкой станет запись вашего голоса или голоса человека, который разрешил использовать его для этого проекта. Вы подготовите короткий референс, озвучите новый текст, соберёте из отдельных реплик минутную дорожку и замените одну реплику после монтажа.

Прежде чем озвучивать текст целиком, сохраните короткий WAV и прослушайте его. Если голос звучит ровно и слова разборчивы, переходите к длинной дорожке.

Как выбрать инструмент

Для клонирования возьмём **Qwen3-TTS-12Hz-0.6B-Base**. Версия **Base** принимает аудиореференс и умеет синтезировать по нему речь на русском. Вариант **CustomVoice** позволяет выбрать один из готовых голосов. В этой главе используется локальная веб-панель Qwen3-TTS. Она открывается в браузере, а модель при запуске по инструкции работает на вашем компьютере. Версии Qwen3-TTS и поддержка языков [34]

Piper подходит, если для текста достаточно готового диктора и удобно быстро сохранять отдельные WAV. У Qwen3-TTS добавляется работа с референсом и контроль того, как тембр держится в разных фразах. Облачный сервис вроде ElevenLabs удобен, когда нужен готовый интерфейс для управления голосами и обработки длинного материала. При выборе учитывайте, можно ли передавать записи и сценарий внешнему сервису, сколько удачных реплик получается и сколько времени уходит на их замену в монтаже. Локальный режим Qwen3-TTS [42] · Работа с голосами в ElevenLabs [43]

Как записать референс

Возьмите собственный голос или запись с разрешением диктора на синтез и использование результата в вашем проекте. Подберите 5–10 секунд чистой речи одного человека. Это отправная точка для прослушивания. Важнее, чтобы слова были разборчивы, речь не перекрывалась музыкой, а начало и конец фрагмента не приходились на середину слова. Для спокойно читаемого текста ищите запись с близким темпом и подачей. Если диктор записывался в нескольких помещениях, сравните участки по шуму и отражениям. Дайте модели материал, который сами готовы были бы поставить в голосовую дорожку.

Расшифруйте выбранный фрагмент дословно, включая произнесённые слова, которых не было в сценарии. Сложные имена и числа запишите так, как диктор действительно их прочитал. Расшифровка относится только к звуку референса. Текст, который нужно озвучить, будет введён отдельно.

Для монтажа возьмём **Audacity 3.7.7**. Если его ещё нет, скачайте audacity-win-3.7.7-64bit.exe со страницы выпуска, запустите файл и завершите установку. Размер установщика — около 18,5 МБ. Откройте исходную запись в Audacity, выделите выбранный участок и в **File → Export Audio** выберите WAV и экспорт выделенного участка (**Current Selection**). Сохраните референс отдельным файлом. Откройте его для проверки и прослушайте от первого звука до последнего. Выпуск Audacity 3.7.7 [44] · Экспорт WAV [45]

Установка Qwen3-TTS на Windows

Маршрут рассчитан на **Windows 11 x64, Python 3.12 x64 и CPU**. Он использует PyTorch 2.8.0, qwen-tts 0.1.1 и Gradio 5.49.1 для интерфейса. Для модели и её звукового токенизатора потребуется скачать около 2,5 ГБ. Отдельное место займут Python, пакеты и звук, который вы сохраните. Расчёт памяти для видеокарты здесь не нужен. В режиме CPU модель и рабочие данные размещаются в оперативной памяти. На настольном ПК с 16 ГБ RAM или больше разумно начать с короткой реплики при закрытых ресурсоёмких приложениях. Это ориентир для первого прогона. Скорость зависит от процессора и длины текста. Файлы модели [46] · Требования пакета qwen-tts [47]

Если у вас пока только Python 3.13 из главы 4, установите Python 3.12 параллельно. На официальной странице выберите **Windows installer (64-bit)**, файл python-3.12.10-amd64.exe размером около 25,7 МБ. Запустите установщик, оставив включённым компонент запуска Python, если он доступен в вариантах установки. При наличии уже установленного Python 3.12 x64 используйте его. Команда py -3.12 выбирает ветку 3.12, даже если на компьютере есть другая версия. Загрузка Python 3.12.10 [48] · Запуск Python в Windows [49]

Python 3.12 x64 выбран для сочетания qwen-tts 0.1.1 с CPU-пакетами PyTorch и TorchAudio 2.8.0. Возьмите установщик конкретного выпуска по ссылке в этом разделе, чтобы сохранить совместимость зависимостей.

В «Документах» создайте папку **QwenVoice**. Откройте её в Проводнике, в адресной строке введите powershell и нажмите **Enter**. В PowerShell из папки **QwenVoice** создайте отдельную среду. После команды должна появиться подпапка .venv. Если команда не видит Python 3.12, проверьте, что установщик завершил работу, и откройте новое окно PowerShell из той же папки.

Итого на диске. Около 2,55 ГБ на установщик Python и папку модели. Понадобится дополнительное место для PyTorch, Gradio, виртуального окружения и WAV. Для монтажа учтите отдельный установщик Audacity (около 18,5 МБ) и место для установленного редактора.

py -3.12 -m venv .venv

В PowerShell из папки **QwenVoice** установите версии PyTorch и torchaudio для CPU. Команда обращается к официальному каталогу сборок PyTorch. Дождитесь окончания установки и возврата к строке ввода без ошибки. Сборки PyTorch 2.8.0 [50]

.\.venv\Scripts\python.exe -m pip install "torch==2.8.0" "torchaudio==2.8.0" --index-url https://download.pytorch.org/whl/cpu

В PowerShell из папки **QwenVoice** установите Qwen3-TTS и версию Gradio, которая даёт интерфейс для загрузки записи. По завершении в .venv\Scripts должны появиться qwen-tts-demo.exe и hf.exe. Если установка прервалась, проверьте свободное место на диске и выполните команду ещё раз из той же папки. Пакет qwen-tts 0.1.1 [47] · Пакет Gradio 5.49.1 [51]

.\.venv\Scripts\python.exe -m pip install "qwen-tts==0.1.1" "gradio==5.49.1"

В PowerShell из папки **QwenVoice** скачайте модель целиком, включая папку speech_tokenizer. После завершения в подпапке **model** будут два файла с именем model.safetensors: один в корне модели и второй внутри speech_tokenizer. Оба нужны для работы. Остальные скачанные файлы содержат конфигурацию и данные для обработки текста. Веса опубликованы под Apache 2.0. Файлы и лицензия модели [46] · Загрузка через клиент Hugging Face [52]

.\.venv\Scripts\hf.exe download Qwen/Qwen3-TTS-12Hz-0.6B-Base --revision 5d83992436eae1d760afd27aff78a71d676296fc --local-dir ".\model"

Первая реплика

В PowerShell из папки **QwenVoice** вставьте следующие три строки. Первая включает автономный режим Hugging Face, вторая выключает аналитику панели Gradio. Третья запускает модель на CPU, со стандартной точностью float32 и без FlashAttention 2. Интерфейс будет доступен только на этом компьютере по адресу http://127.0.0.1:7860. Подождите, пока PowerShell покажет локальный адрес, и откройте его в браузере. Если программа сообщает, что не находит модель, убедитесь, что скачивание завершилось и внутри **model** есть корневой файл весов и папка **speech_tokenizer**. Если модель надо загрузить заново, откройте новое окно PowerShell из папки **QwenVoice**, чтобы команда скачивания выполнялась без ранее включённого автономного режима. Локальная панель Qwen3-TTS [42] · Автономный режим Hugging Face Hub [53]

```
$env:HF_HUB_OFFLINE = "1"

$env:GRADIO_ANALYTICS_ENABLED = "False"

.\.venv\Scripts\qwen-tts-demo.exe ".\model" --device cpu --dtype float32 --no-flash-attn --ip 127.0.0.1 --port 7860 --no-share
```

На вкладке **Clone & Generate** загрузите подготовленный WAV в **Reference Audio**. В **Reference Text** внесите расшифровку именно этой записи. Оставьте **Use x-vector only** выключенным, чтобы модель могла воспользоваться и звуком, и его расшифровкой. В **Target Text** введите первую новую реплику, а в **Language** выберите **Russian**. Нажмите **Generate**. Когда появится сообщение **Finished**, прослушайте **Output Audio** и сохраните WAV кнопкой скачивания у плеера. Поместите его в папку проекта и дайте имя, по которому найдёте реплику при монтаже. Интерфейс клонирования Qwen3-TTS [54]

Сравните WAV с референсом по тембру, ясности слов и естественности пауз. Если тембр странно меняется или первые слова звучат неразборчиво, проверьте, точно ли введён текст референса и нет ли в нём обрезанного слова. Затем повторите ту же реплику с более чистым отрывком разрешённой записи и сравните результаты. За один прогон меняйте одно условие. Так проще понять, помогла ли замена записи или изменение написания новой реплики.

Сборка минутной дорожки

Разбейте свой текст на фразы или небольшие смысловые блоки. Каждая часть должна быть достаточно полной, чтобы по ней было понятно, где падает интонация и где уместна пауза. Синтезируйте по одному блоку, используя тот же референс и тот же язык. Сохраняйте дубли отдельно от исходной записи. Присвойте файлам последовательные номера, чтобы их порядок сохранился при импорте.

В Audacity импортируйте WAV через **File → Import → Audio**. Если они открылись на отдельных дорожках, перемещайте клипы за верхнюю полосу и выстройте их на общей шкале времени. Сначала выставьте порядок реплик, потом паузы. Прослушайте начало, середину и конец подряд. Отметьте резкие изменения громкости, высоты голоса и темпа. Если часть резко выбивается, попробуйте один новый дубль с той же подачей исходной записи и тем же текстом. Для нужной длительности убирайте лишние паузы и фразы, если они не нужны по смыслу. Проверьте конец по шкале времени и доведите общую длину до примерно одной минуты. Сохраните проект Audacity и экспортируйте собранную дорожку через **File → Export Audio**

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, купив полную легальную версию на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.