

Local AI

**Запуск моделей с открытым исходным кодом
на личном компьютере**

Лекция и практикум

2026

Llama · Mistral · Gemma | LM Studio · Ollama

План занятия

Теория (40 мин)

1. Почему локальный ИИ
2. Открытые модели: Llama, Mistral, Gemma
3. Как модель помещается в ноутбук

Практикум (50 мин)

4. Установка LM Studio
5. Установка Ollama
6. Работа без интернета
7. Анализ конфиденциального документа

Цель

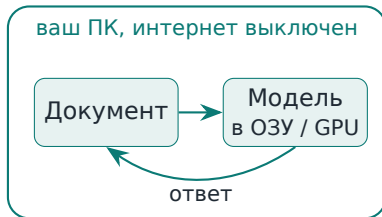
Запустить модель на своём ноутбуке, отключить интернет и проанализировать документ так, чтобы ни один байт не покинул компьютер.

Куда уходят ваши данные

Облачный чат-бот



Локальная модель



Локально: данные не передаются третьим лицам, нет журналов на чужих серверах, нет трансграничной передачи.

Облако или локально?

Критерий	Облачный сервис	Локальная модель
Где обрабатываются данные	На серверах провайдера	В памяти вашего ПК
Интернет	Нужен всегда	Только для скачивания
Качество	Максимальное	Ниже, но достаточно для многих задач
Стоимость	Подписка / токены	Бесплатно после покупки «железа»
Версия модели	Меняется без спроса	Неизменна, пока вы не обновите
Скорость	Не зависит от ПК	Зависит от CPU, GPU и ОЗУ

Когда облако всё же лучше

Нет секретных данных, нужна максимальная точность, слабый компьютер.

Конфиденциальность и закон

Закон РУз «О персональных данных» (ЗРУ-547, 2019)

Персональные данные граждан Узбекистана, обрабатываемые с помощью ИТ, должны храниться на технических средствах на территории республики.

- Похожая логика: GDPR (ЕС), 152-ФЗ (Россия)
- Загрузка кадровых документов или клиентской базы в зарубежный чат-бот может нарушать требования
- Локальная модель снимает вопрос трансграничной передачи

Типичные секретные данные: личные дела, медицинские заключения, финансовая отчётность, тендеры, исходный код, судебные материалы.

Перед внедрением в организации сверьтесь с юристом и действующей редакцией закона.

«Открытые» модели: open source или open weights?

Open weights

Опубликован файл обученной модели — его можно скачать и запустить у себя. Данные и код обучения обычно закрыты.

Что такое модель

Таблица из миллиардов чисел (параметров), предсказывающая следующий токен. **8B** = 8 млрд параметров.

Для конфиденциальности важно одно:

веса лежат у вас, запрос никуда не уходит

Основные семейства (октябрь 2026)

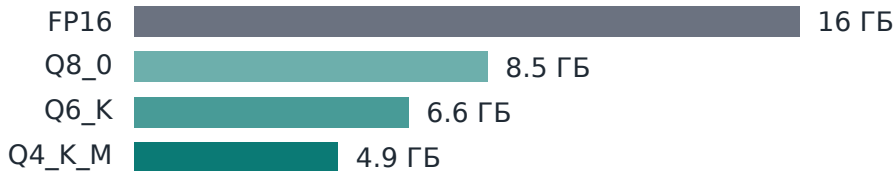
Семейство	Разработчик	Для ноутбука	Лицензия	Сильные стороны
Gemma 4	Google	E2B, E4B, 12B	Apache 2.0	Текст + изображения, лучший выбор для ноутбука
Mistral	Mistral AI	Ministral 3 (3-14B)	Apache 2.0*	Европейская юрисдикция, многоязычность
Llama	Meta	3.2 3B, 3.1 8B	Llama License	Огромная экосистема
Qwen	Alibaba	4B-27B	Apache 2.0	100+ языков, хороший русский, код
DeepSeek	DeepSeek	7B-14B (дистилл.)	MIT	Рассуждения, математика

- Gemma 4 (апрель 2026) — свободная лицензия Apache 2.0; Llama 4 (2025) для ноутбука великовата
- * Mistral Medium 3.5 — своя лицензия; Llama — ограничения для сервисов с 700+ млн пользователей

Как 16 ГБ превращаются в 5 ГБ: квантизация

Веса округляются: вместо 16 бит на число — 8, 5 или 4 бита.

$$\text{Память (ГБ)} \approx \frac{N_{\text{млрд}} \times \text{бит на вес}}{8} + 1-2 \text{ ГБ}$$



Модель 8V: размер файла весов

Q4_K_M — стандарт по умолчанию: качество почти не страдает, размер в 3 раза меньше.

Форматы и «железо»

- **GGUF** — один файл: веса + токенизатор + настройки. Понимают llama.cpp, LM Studio, Ollama
- **MLX** — формат для Apple Silicon, быстрее на Mac
- Быстрее всего — в **VRAM** видеокарты; на Mac память общая
- Без GPU — на CPU, в 3-10 раз медленнее

Память	Что запускать
8 ГБ ОЗУ	Gemma 4 E2B, Llama 3.2 3B
16 ГБ / Mac 16 ГБ	Gemma 4 E4B, Qwen 3 8B, Llama 3.1 8B
32 ГБ / GPU 12-16 ГБ	Gemma 4 12B, Qwen 3 14B
64 ГБ / GPU 24 ГБ	Gemma 4 31B, Qwen 3.6 27B

Комфортная скорость чтения — 5-10 токенов/с.
1 страница русского текста $\approx$ 500-700 токенов.

Практикум

LM Studio · Ollama · офлайн · документы

Практикум 1: LM Studio — графический интерфейс

1. Скачать: `lmstudio.ai/download` (Windows, macOS, Linux; бесплатно, в т. ч. для работы)
2. Установить. На x64 нужен процессор с **AVX2**
3. Вкладка **Discover** → найти `gemma-4-e4b` или `qwen3-8b`
4. Выбрать квантизацию **Q4_K_M** — зелёная метка: модель поместится
5. **Download** (3–6 ГБ) → вкладка **Chat** → выбрать модель
6. Спросить: «Объясни простыми словами, что такое квантизация»

Ключевые настройки

Context Length — объём текста **GPU Offload** — максимум, что влезет **Temperature** 0,1–0,3 для документов

LM Studio: документы, сервер, перенос моделей

- **Документы:** перетащите PDF, DOCX или TXT в окно чата. Короткий файл вставляется целиком, длинный — фрагментами (встроенный RAG)
- **Локальный сервер:** Developer → Start Server API, совместимый с OpenAI: `http://localhost:1234/v1`
- **Без интернета:** файл `.gguf` копируется с флешки в `~/.lmstudio/models/издатель/модель/`
- **CLI:** `lms` — управление моделями из терминала

Практикум 2: Ollama — одна команда

Бесплатно, открытый код (MIT), работает как фоновая служба с API.

Установка

```
# Windows / macOS: установщик с ollama.com/download
# Linux:
curl -fsSL https://ollama.com/install.sh | sh
ollama --version
```

Первый запуск

```
ollama run gemma4:e4b          # скачает и откроет чат
>>> Объясни, что такое квантизация
>>> /bye
```

Основные команды Ollama

Команда	Что делает
<code>ollama pull qwen3:8b</code>	Скачать модель
<code>ollama run qwen3:8b</code>	Запустить чат
<code>ollama run qwen3:8b --verbose</code>	Чат + скорость (токенов/с)
<code>ollama list</code>	Скачанные модели
<code>ollama ps</code>	Модели в памяти, доля GPU/CPU
<code>ollama show qwen3:8b</code>	Параметры, контекст, лицензия
<code>ollama rm <модель></code>	Удалить модель

Каталог: `ollama.com/library` — llama3.2:3b, llama3.1:8b, mistral, qwen3:8b, gemma4:e4b

Модели: `~/.ollama/models/` API: `http://localhost:11434`

Ollama: свой ассистент через Modelfile

```
FROM qwen3:8b
PARAMETER temperature 0.2
PARAMETER num_ctx 16384
SYSTEM "Ты внимательный юрист. Отвечай на русском,
опирайся только на текст документа, указывай пункты."
```

```
ollama create jurist -f Modelfile
ollama run jurist
```

Анализ файла одной командой (Linux/macOS):

```
ollama run jurist "Найди риски для арендатора: $(cat dogovor.txt)"
```

LM Studio или Ollama?

	LM Studio	Ollama
Интерфейс	Графический	Терминал + простое окно чата
Порог входа	Ниже	Чуть выше
Документы	Встроенный RAG	Через команды или Open WebUI
Код программы	Закрытый, бесплатный	Открытый (MIT)
Автоматизация	API :1234, CLI <code>lms</code>	API :11434, удобно для скриптов

Совет

Начинайте с LM Studio, переходите на Ollama, когда понадобится автоматизация.

Практикум 3: отключаем интернет

1. Модель скачана? (`ollama list / My Models`)
2. Выключить Wi-Fi и кабель
3. `ping 8.8.8.8` → ошибка
4. Задать вопрос модели → **ответ есть!**
5. Диспетчер задач: растёт нагрузка CPU/GPU

Проверка соединений

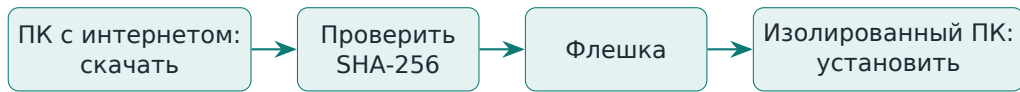
```
# Windows
netstat -ano
# macOS / Linux
lsof -i -P | grep -iE "ollama|lm"
```

Только `localhost:11434 / :1234`,
никаких внешних адресов.

В сеть обращаются только

обновления, каталог моделей и загрузки — отключите
автообновление.

Полностью изолированный компьютер (air-gap)



```
certutil -hashfile model.gguf SHA256      # Windows  
shasum -a 256 model.gguf                  # macOS / Linux
```

Сравните результат с контрольной суммой на странице модели (Hugging Face).

Практикум 4: анализ секретного документа

Целиком в контекст

До 10–20 страниц. Модель видит весь текст — точнее. Context Length 16–32 тыс.

RAG

Большие документы и архивы. В модель подаются только фрагменты, похожие на вопрос. Риск: нужный кусок не найдётся.

- Учебный материал: **вымышленный** трудовой договор или медзаключение
- Сканы: сначала OCR (Tesseract) или мультимодальная Gemma 4
- Для организации: Open WebUI / AnythingLLM поверх Ollama

Шаблоны запросов

1. **Резюме:** «Составь резюме документа в 5 пунктах, для каждого укажи номер раздела»
2. **Извлечение:** «Найди стороны, сумму, сроки, штрафы. Дай таблицу. Если чего-то нет — „не указано“»
3. **Риски:** «Ты юрист со стороны работника. Перечисли невыгодные условия и процитируй пункты»
4. **Проверка:** «Есть ли пункт о конфиденциальности? Да/нет и цитата»
5. **Обезличивание:** «Замени ФИО, паспорта, телефоны, адреса на [ФИО], [ПАСПОРТ]...»

Проверяйте ответ

Сверяйте цитаты с оригиналом, задавайте вопрос дважды, сравнивайте 2–3 модели. Малые модели «галлюцинируют».

Локально — не значит автоматически безопасно

- Модели — только из официальных источников; **GGUF** безопаснее старых .bin/.pt (pickle)
- **Не открывайте** порты 11434 / 1234 в сеть: в интернете тысячи открытых серверов Ollama
- Чаты и документы на диске — включите шифрование (BitLocker, FileVault, LUKS)
- Промпт-инъекции: документ может содержать скрытые инструкции для модели
- Обновляйте LM Studio и Ollama — закрываются уязвимости

Ограничения: слабее облака в сложных задачах, знания «заморожены», скорость зависит от железа, ноутбук греется.

Чек-лист перед работой с секретным документом

- ☐ Модель из официального источника, SHA-256 совпадает
- ☐ Интернет отключён / исходящие соединения запрещены
- ☐ API слушает только localhost
- ☐ Диск зашифрован, учётная запись под паролем
- ☐ Температура 0,1-0,3, в промпте — «цитируй текст»
- ☐ Каждый факт сверен с оригиналом
- ☐ Ненужные чаты удалены

Задания

1. Установите LM Studio или Ollama, запустите модель 3-8B, запишите скорость (токенов/с)
2. Отключите интернет и докажите, что модель работает (скриншот)
3. Проанализируйте учебный договор тремя шаблонами; найдите хотя бы одну ошибку модели
4. Сравните две модели на одном документе: точность, полнота, качество русского языка
5. *Дополнительно:* создайте Modelfile «юрист» или «HR-ассистент»

Контрольные вопросы

1. Чем open-weight модель отличается от open source?
2. Почему локальная обработка снижает риски? Какие риски остаются?
3. Сколько памяти нужно модели 12B в Q4_K_M?
4. Что такое GGUF и зачем нужна квантизация?
5. Как проверить, что программа не отправляет данные в сеть?
6. Чем подача документа целиком отличается от RAG?
7. Почему нельзя открывать порт Ollama в интернет?

Итоги

- Открытые модели (Gemma, Mistral, Llama, Qwen) работают на обычном ноутбуке
- Квантизация Q4 — ключ к 8–16 ГБ памяти
- LM Studio — просто, Ollama — гибко
- Интернет нужен только для скачивания
- Локальный ИИ защищает данные, но ответы надо проверять

Спасибо!

Вопросы?

Ссылки: lmstudio.ai · ollama.com · ollama.com/library · huggingface.co · lex.uz (ЗРУ-547)