

M00. Шпаргалка

Окружение, модели и инструменты

Команды курса

Команда	Что делает
make help	все команды
make setup	первичная настройка
make llm-check	проверка моделей
make lab M=00	JupyterLab в Docker
make native M=00	JupyterLab нативно
make test M=00	проверка упражнений
make check M=00	прогон ноутбуков
make opencode	OpenCode с .env

uv

Команда	Что делает
uv sync	окружение по uv.lock
uv add pkg	добавить зависимость
uv run cmd	запуск в окружении
uv python install 3.12	поставить Python

Docker

Команда	Что делает
docker compose up -d	поднять сервисы
docker compose ps	что запущено
docker compose logs -f	логи
docker compose down	остановить

Из контейнера хост — это `host.docker.internal`.

Ollama

Команда	Что делает
make ollama	сервер с контекстом 16K
ollama pull qwen3:8b	скачать модель
ollama list	скачанные модели
ollama ps	загруженные в память

.env

```
LLM_BASE_URL=...      # OpenAI-совм. endpoint
LLM_API_KEY=...
LLM_MODEL=...
LLM_LOCAL_MODEL=qwen3:8b
EMBED_MODEL=bge-m3
LLM_CACHE=on          # off|on|record|refresh|replay
```

mlcourse

```
from mlcourse import (get_client, model_name,
                      no_thinking, chat, embed)

client = get_client()          # роль main
r = client.chat.completions.create(
    model=model_name(),
    messages=[{"role": "user",
               "content": "Привет"}],
    extra_body=no_thinking())
r.choices[0].message.content
chat("Привет", role="local")  # коротко
embed(["текст 1", "текст 2"]) # [2, 1024]
```

Формулы

Верхняя граница скорости генерации:

$$\text{ток/с} \lesssim \frac{\text{пропускная способность, ГБ/с}}{\text{размер весов, ГБ}}$$

Память весов: $N \cdot b/8$ байт.

KV-кэш: $2 \cdot L \cdot H_{kv} \cdot d_{\text{head}} \cdot s \cdot T$ байт.

Ответ модели

Поле	Смысл
finish_reason: stop	закончила сама
finish_reason: length	упёрлась в max_tokens
finish_reason: tool_calls	зовёт инструмент
usage	токены промпта и ответа

SSE-поток: строки `data: {...}`, конец — `data: [DONE]`.