**Локальные языковые модели (LLM)**

**🔹 Параметры и размер**

* **Параметры (B = миллиардов)** → условная «мозговитость» модели.
  + 7B → лёгкие, запускаются на ноутбуках.
  + 13B → баланс качества и ресурсов.
  + 34B+ → медленно на домашнем ПК, лучше сервер/облако.
  + 70B+ → дата-центры (только через аренду).
* **Размер (FP16)** ≈ Параметры × 2 байта.
  + 7B → ~13 ГБ
  + 13B → ~26 ГБ
  + 34B → ~65 ГБ
  + 70B → ~130 ГБ

**🔹 Квантизация**

* Сжатие весов для работы на слабом железе.
* Основные типы:
  + **FP16** → оригинал (требует много VRAM).
  + **INT8** → ~½ размера.
  + **INT4** → ~¼ размера (оптимально для RTX 3070 Ti 8 ГБ).
* Примеры:
  + **Q4\_K\_M** — популярный компромисс (качество + экономия).
  + **Q5\_K\_M** — лучшее качество, больше памяти.
  + **Q8\_0** — почти без потерь, требует VRAM ≥ 24 ГБ.

**🔹 Способы запуска**

* **LM Studio** → GUI-клиент для Windows/Mac/Linux. Удобен, поддерживает GGUF.
* **Ollama** → лёгкий CLI-сервер, есть под Mac/Win/Linux.
* **llama.cpp** → C++ библиотека для запуска LLaMA-подобных моделей.
* **text-generation-webui** → веб-интерфейс для запуска разных моделей (поддержка HF, GGUF).

**🔹 Формат GGUF**

* Новый формат (сменил GGML).
* Универсальный, оптимизирован для llama.cpp / LM Studio / Ollama.
* Легко качать с HuggingFace и запускать без переконвертации.

**🔹 Популярные репозитории**

* **HuggingFace** → главный хаб моделей (https://huggingface.co).
* **TheBloke** → выкладывает готовые GGUF версии моделей.
* **NousResearch**, **MistralAI**, **Yandex** → авторы популярных моделей.

**🔹 Популярные модели**

* **LLaMA 2 (7B, 13B, 70B)** → универсальная база.
* **Mistral 7B / Mixtral 8x7B (MoE)** → быстрые, сильные на английском.
* **CodeLLaMA (7B, 13B, 34B)** → под кодинг.
* **DeepSeek-Coder (7B, 33B, 67B)** → лучший для кода.
* **Yi-34B** → мощный универсал (качество близко к GPT-3.5).
* **YaLM 2 (7B, 13B)** → от Яндекс, лучше на русском.

**🔹 Резюме под ПК (RTX 3070 Ti, 8 ГБ VRAM)**

* **Q4 версии 7B/13B моделей** — оптимальный вариант.

Я использую

LM Studio

Meta Llama 3.1 8B Instruct Q4\_K\_M 4.58 GB

YandexGPT 5 Lite 8B Instruct Q4\_K\_M 4.58 GB

Настраиваю так, чтобы все слои работали на GPU

При длине контекста 8-10 тыс. токенов загрузка памяти около 7,5-7,9 Гб