Инструкция по локальному инференсу (запуску) LLM

Чтобы запускать нейронки, которые могут делать что-то полезное, уже не обязателен топовый комп. Нейронки стали меньше и умнее. Но и на старье они не пойдут.

Минимальные требования на текущий момент (июнь 2026) — ноутбук с 16G RAM, процессором уровня Intel 10 поколения с 8 логическими ядрами, и какой-никакой дискретной видеокартой.

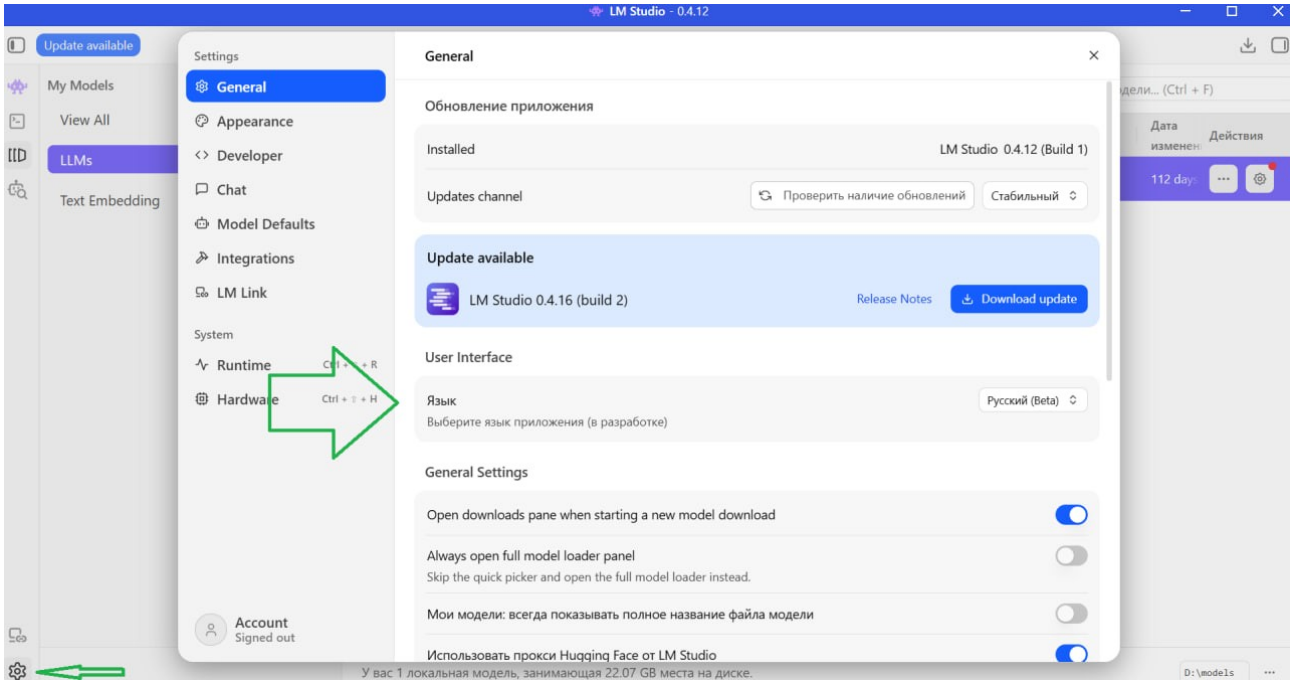
Общее правило — вы не сможете запустить нейронку большего размера, чем сумма RAM+VRAM.

1. Качаем LM Studio: https://lmstudio.ai под вашу ОС

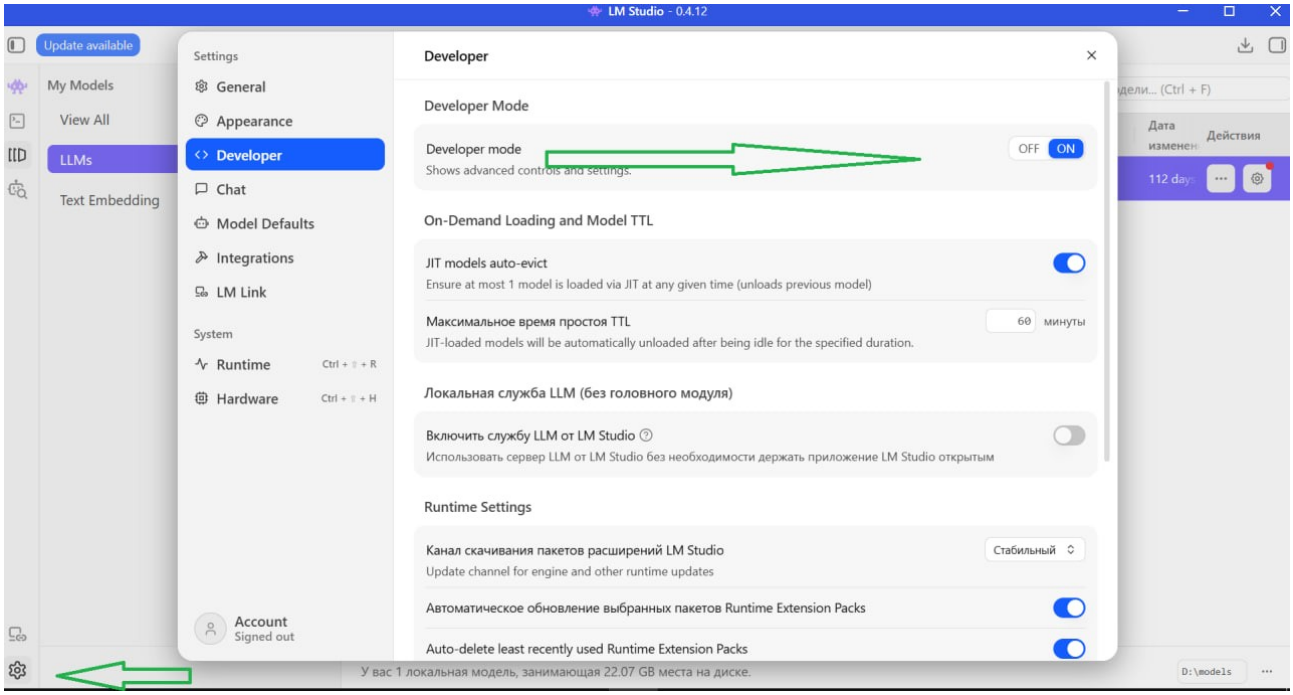
2. Предусловие: ваша ОС должна быть ободрана таким образом, чтобы размер занятой RAM в простое был минимальным. Если у вас 16G, должно быть занято не более 2G под системные нужды.

Как это сделать — отдельная тема. Например, подымайте архивы проекта https://ameliorated.info на wayback machine. Сейчас проект выродился и продвигает закрытое решение; примерно в 2022-2023 годах еще были открытые скрипты, которые позволяли убрать bloatware с компьютера.

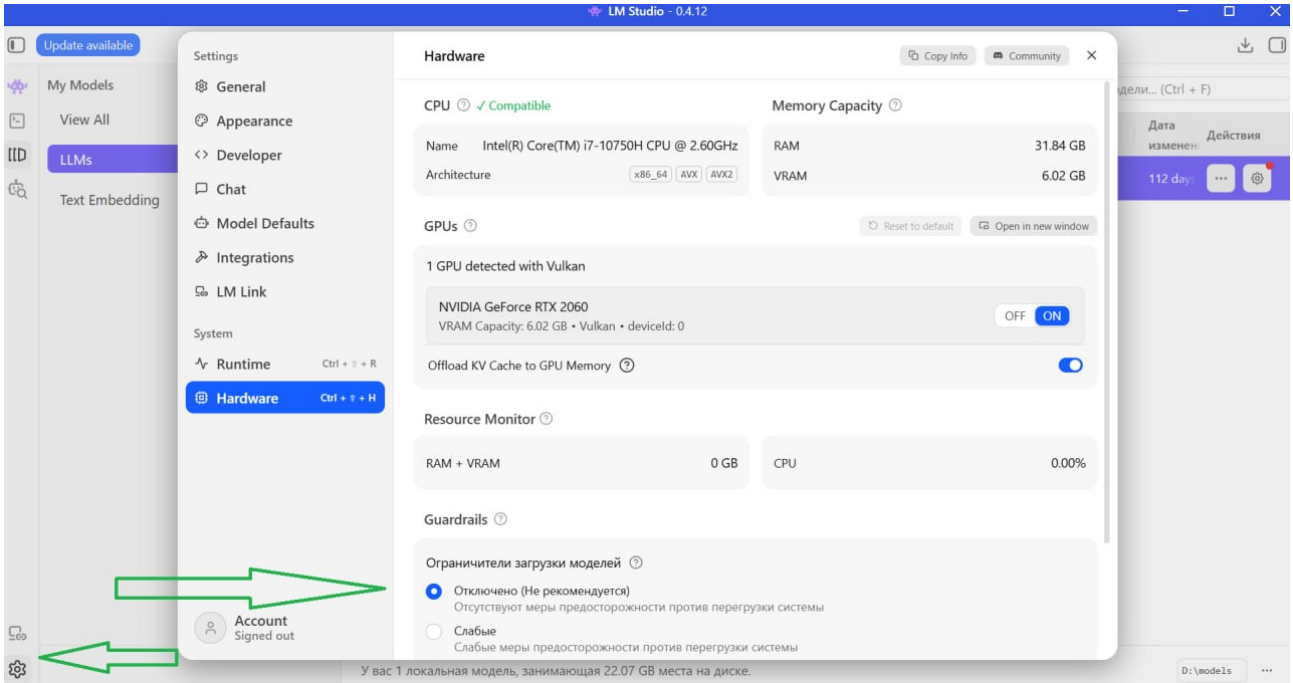
3. Выбираем русский язык:



4. Включаем режим разработчика, это откроет некоторые важные пункты настроек:

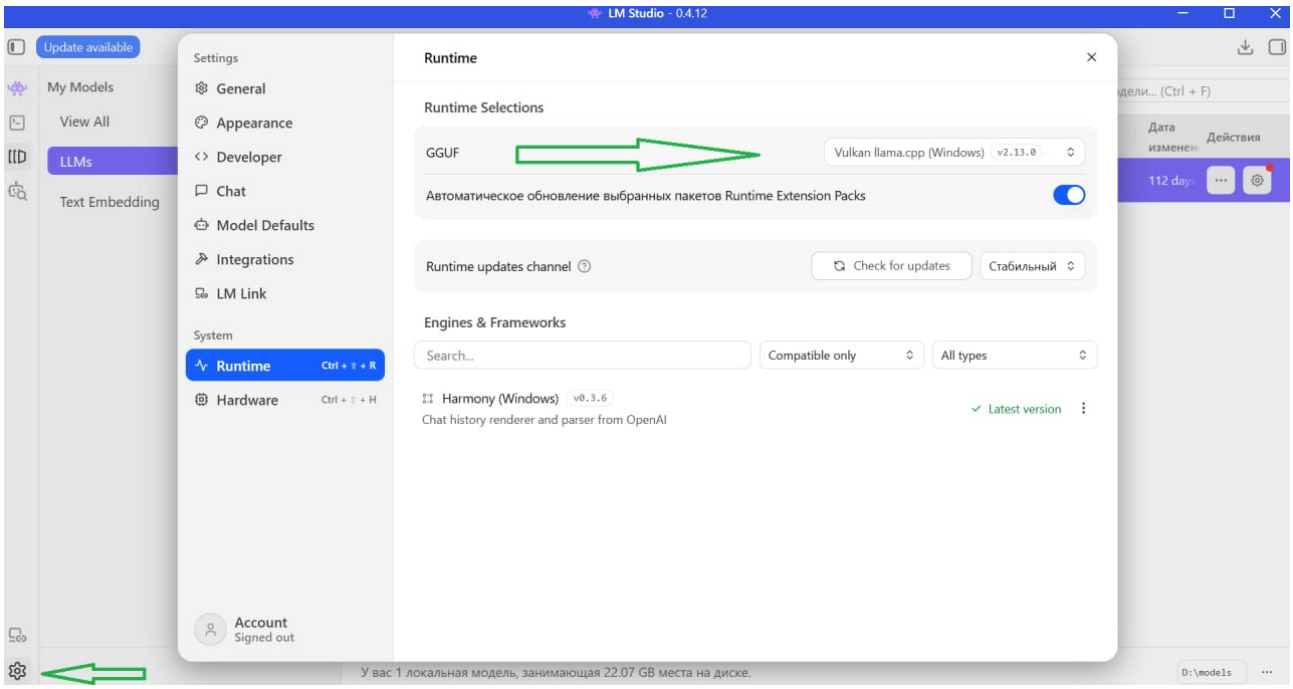


5. Отключаем ограничитель загрузок моделей:



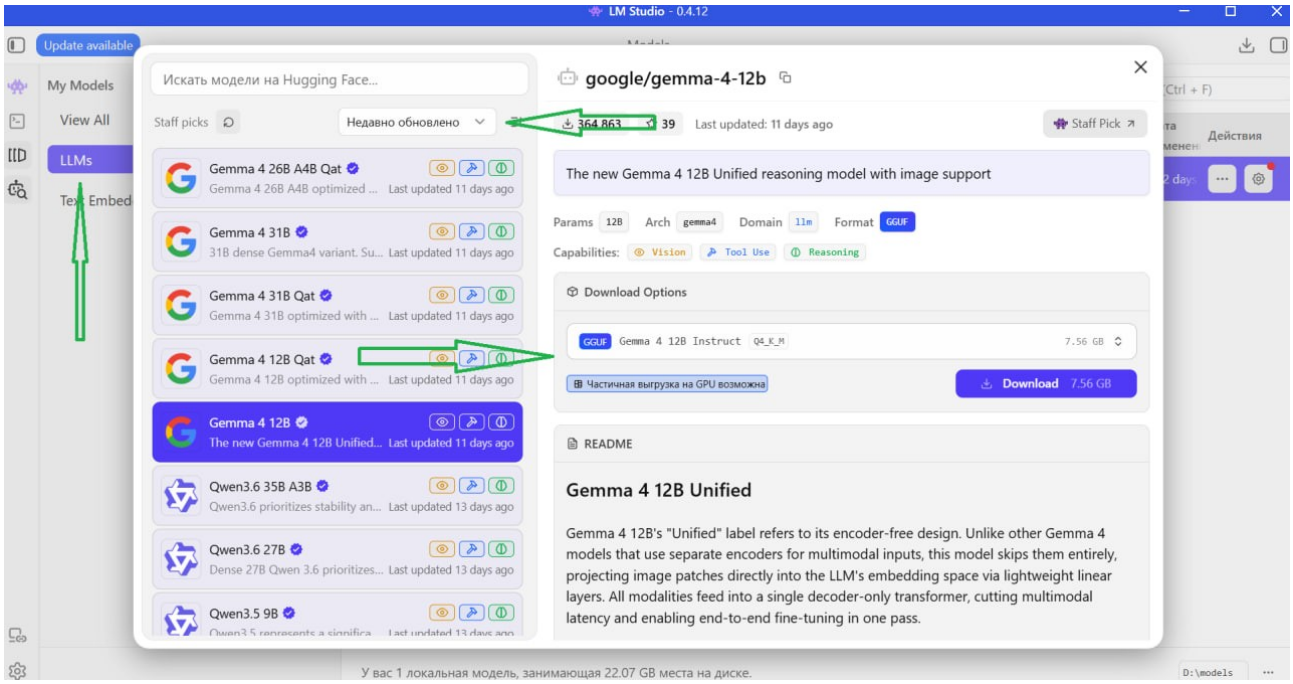
Если он включен, он не даст выбирать модели, которые при обработке рашпилем могли бы втиснуться в ваш комп.

5. Выбираем ваш runtime.
Владельцы Nvidia могут выбирать CUDA, AMD – Vulkan.
Насчет NPU: вот https://gist.github.com/vsbogd/49e8c470a709c963454e20965beb12e6
вроде сейчас в LMStudio поддержки нет.



6. Выбираем модели
Включаем сортировку по дате, чтобы увидеть самые свежие модели.
Под слабые компьютеры идут модели Gemma 4, Phi, Nemotron.

В карточке модели есть выпадающее меню с вариантами.
В вариантах указана квантизация, она влияет на размер. Модель с размером больше чем ваше ОЗУ у вас не запустится.
Чем меньше размер кванта, тем быстрее модель и тем она тупее.

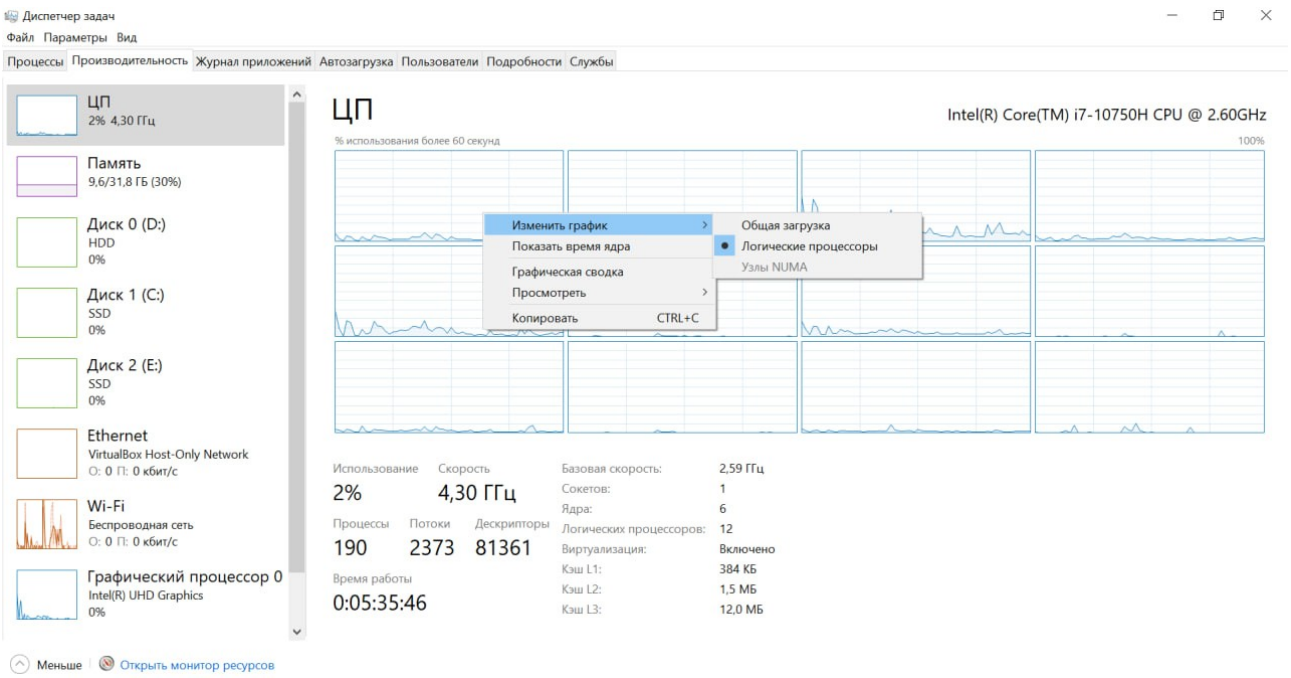


У них также есть отметки что они могут:
- молоточек - интегрируется как tool в агентов таких как hermes
- глаз - распознает картинки
- и со звуком там еще и видео, но это я не пробовал

8. Настройка

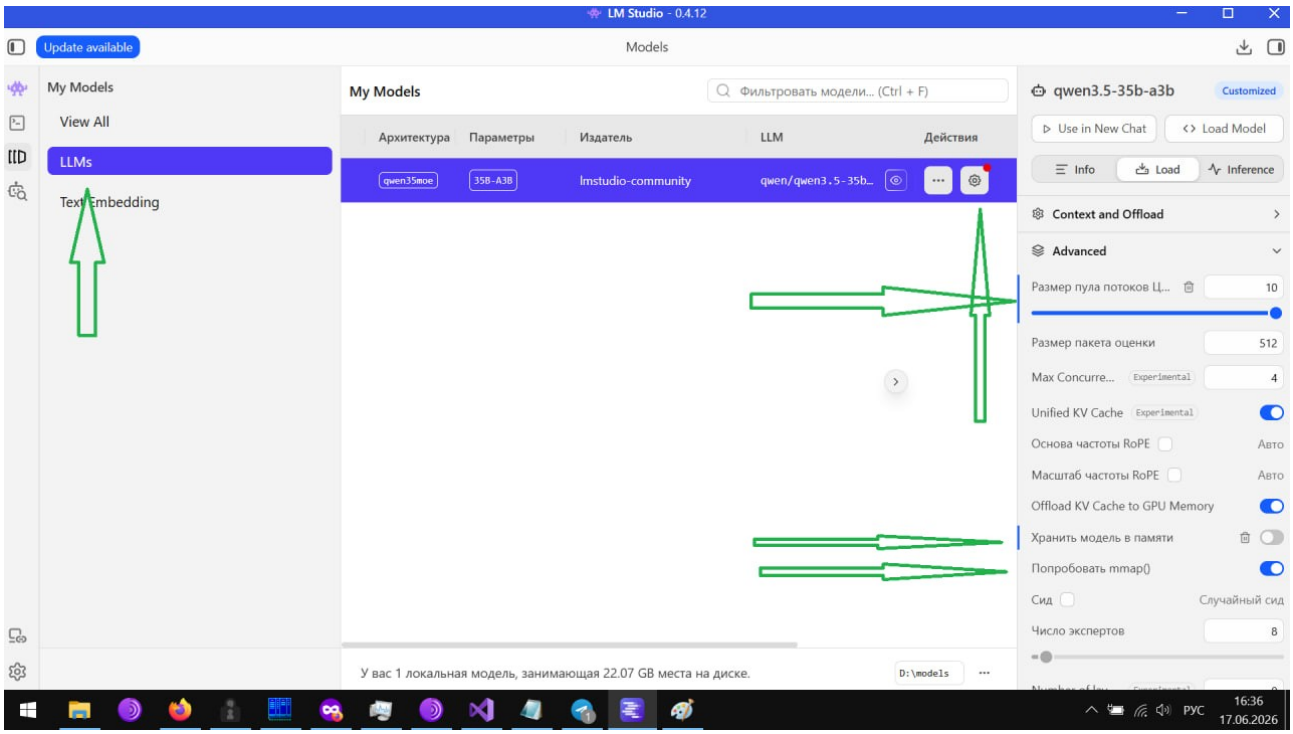
Перед тем как запускать, надо будет провести донастройку.
Вашим лучшим другом будет диспетчер задач.

Нужно знать сколько у вас физических ядер ЦП.

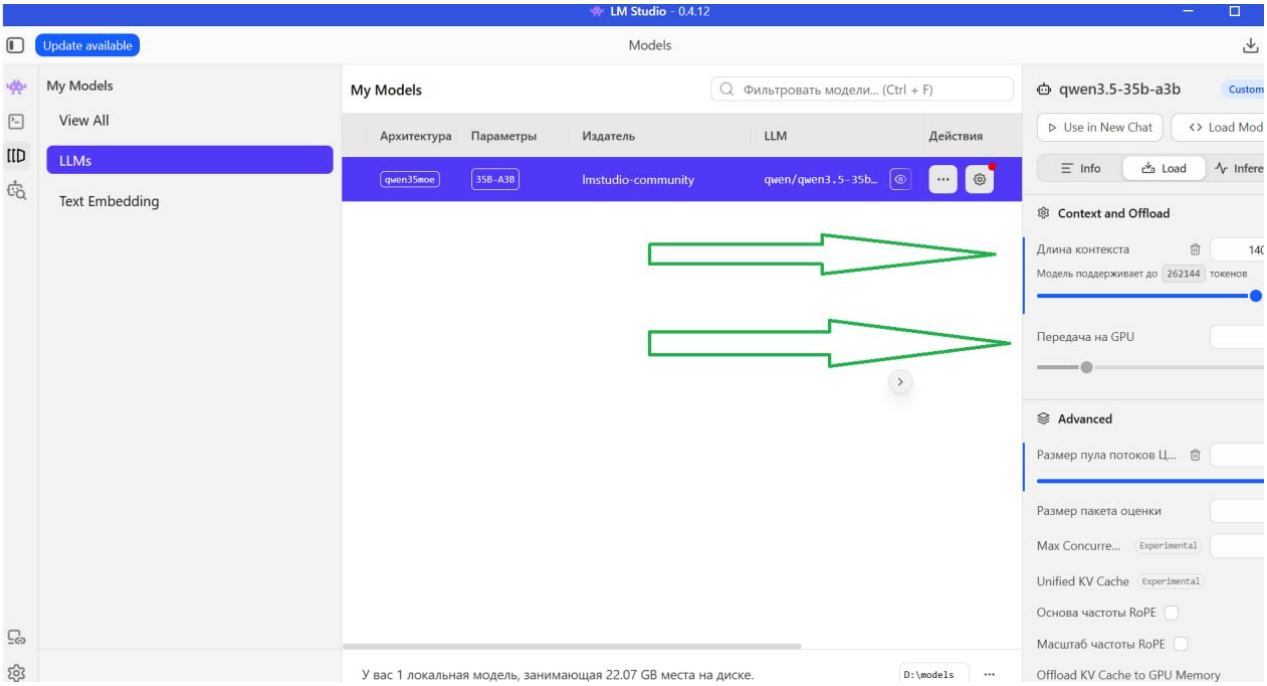


Также нужно будет постоянно мониторить потребление RAM, VRAM (выделенной).

8.1. Ставим "Размер пула потоков ЦП" равный количеству логических ядер минус 2.



8.2. Контекст



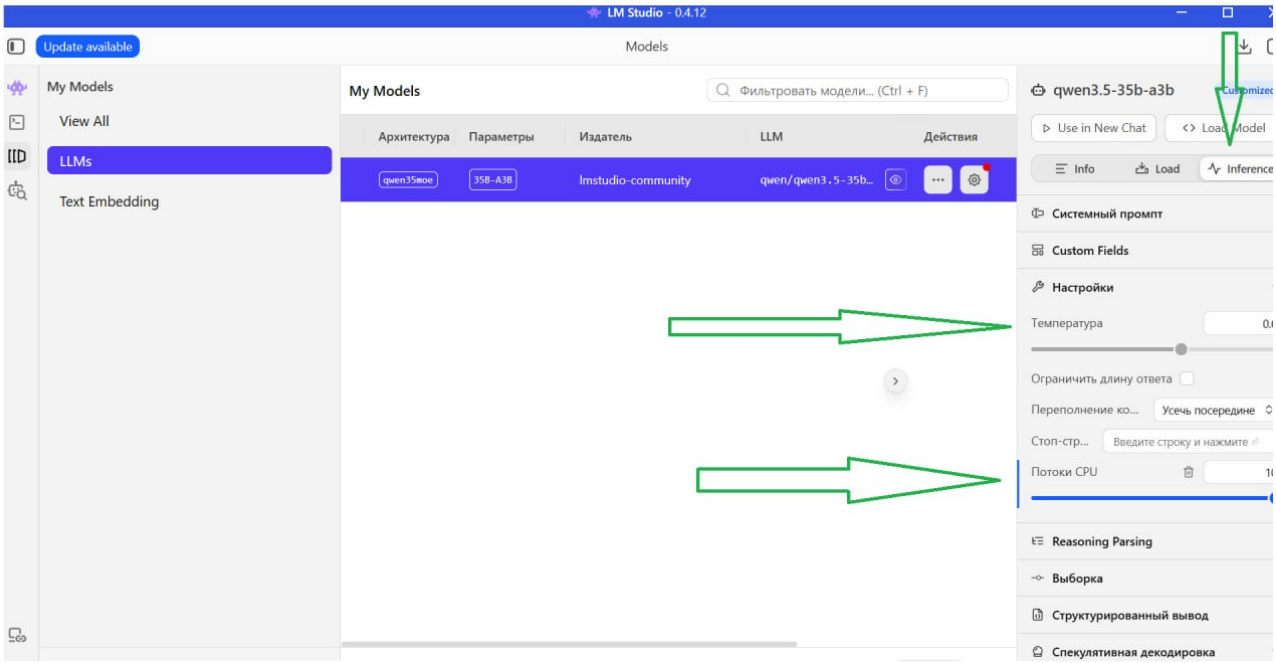
два важнейших параметра
- размер контекста. Чем больше контекст, тем большие тексты обрабатывает модель. Считаем условно 1 токен = 1 слог.

Контекст зависит от размера ОЗУ. Начинаем с малого контекста, убеждаемся что модель грузится и работает, и поэтапно увеличиваем контекст до максимального значения. Я увеличиваю в 2 раза за попытку.

- Кол-во слоев, которые обрабатываются на видеокарте. Методом проб и ошибок, от минимального значения, увеличиваем и пробуем.
В диспетчере задач нужно открыть вкладку видеокарты и смотреть потребление памяти. Чем больше занято - тем лучше. Но если перебрать, модель не загрузится и вы получите ошибку.

При тюнинге запуска, нужно по отдельности настраивать каждый параметр, перед тем как переходить к следующему. Я советую начинать со второго параметра, а контекст оставить на потом

8.3. Температура



температура влияет на степень точности или безумия модели. 1 - полное творчество, 0 - полное воспроизведение результатов при одинаковых запросах. Обычно оставляют по умолчанию рекомендованное производителем. Просто на понимание что это такое.

Кол-во потоков ЦП должно быть равно тому, что вы выставите на предыдущей вкладке по кол-ву ядер.

Также тут можно установить системный промпт.

8.4. Выключите параметр «Хранить модель в памяти» и оставьте «Использовать mmap”.

в принципе, это основные вещи.
Правильно настройте эти параметры и пробуйте работать.

Важно добиться хоть какого-то запуска, когда модель просто загружается как есть, и начинает хоть как-то отвечать. Это самый основной тест. Если он пройден, то дальше уже делаете тюнинг для увеличения скорости.
Если ничего вообще не работает, либо
- сама модель глючная (так бывает, что выдают кривые карточки моделей)
- неверно выбран runtime (вулкан нерабочий например, так тоже бывает)

тогда надо смотреть логи, гуглить и разбираться

9. Чем мощнее у вас железо, тем больше открывается возможностей.

Использовать локальные LLM для рабочих задач **более чем реально**.

Контекст в 100к-256к токенов — это размер проекта порядка 10-20к строк кода, который можно загрузить в нейронку целиком.

Для серьезной работы нужен компьютер с характеристиками:
RAM: 32-64G
CPU: не менее 10 поколения Intel, лучше 13-14, чем мощнее и больше ядер тем лучше
Video: NVIDIA 3080+
Это игровой ноутбук 2-3 летней давности.