Сотрудничество



Движки: llama.cpp

- Inception: 10 марта 2023 года
- Этот проект имеет ту же значимость, что в свое время Linux
- Стек llama+free llms стал пригодным к использованию в начале 2025 года (QWEN-2.5, Deepseek R1)



Движки

- vLLM
- Ollama
- SGLang - серверный
- MLX — Apple

  Ничего о них не знаю!
- Но они примерно того же уровня что llama.cpp



Проблемы

Проблема llama.cpp и движков того же класса:

- Запуск только моделей, умещающихся в RAM + VRAM
- Негибкий offloading

Итог: эффективно работают только модели целиком помещающиеся в VRAM

Проблемы

- LM Studio: кривой харнесс
- Другие харнессы: тоже есть проблемы — не те акценты; только командная строка; телеметрия, итд

Итог: невозможен быстрый старт по сценарию: дал задачу и забыл



Проблемы

Харнесс:

- Нету единой мультимодальной среды: текст, звук, видео, картинки на входе и на выходе
- LM Studio: текст+картинка на входе, текст на выходе
- ComfyUI: ужс питон смерть поганая



Новые ребята на раене

- Strata
  https://github.com/Niko1221/Strata
- Project Maya
  https://github.com/mw00/project-maya
- Unsloth Studio+Desktop
  https://unsloth.ai



Новые ребята на раене

Движок:

- Гибкий offloading: стримим горячих экспертов в VRAM в реальном времени
- Запускаем модели > RAM+VRAM
- Запускаем топовые открытые модели: Qwen 3.8-Flash-Next, GLM-5.3, Deepseek 4.1, MiniMax 3.1 !!1



Новые ребята на раене

Харнесс (strata):

- Все включено, сел и поехал
- На борту: встроенные инструменты: работа с файлами/команды/веб-серфинг: попячено из харнесса DeepSeek https://github.com/deepseek-ai/deepseek-harness
- Коннектор к MCP
- Коннектор к внешним REST API в облаке
- Свой REST API → доступ снаружи по сети
- Скиллы (автовыбор)
- Авто компактификация
- Инференс картинок (заявлено, но не увидел в UI)



Новые ребята на раене

Харнесс (unsloth)

- Звук+картинка+текст на входе
- Генерация текста, картинок, видео (?) в одной среде
- **Обучение/дообучение моделей**



Состояние модели

Ожидание · Чтение · **Генерация** · В очереди · Ошибка

Generating a reply

1.3k токенов · 60.5 токенов/с

Скорость

Измерено

60,5 t/s — Текущая генерация

401 t/s — Обработка этого запроса

Загрузка GPU

Измерено

98 %

NVIDIA GeForce RTX 4060 Laptop GPU

VRAM

Измерено

7,5 / 8 GB

Экспертов в кеше: 553

Темп. GPU

Измерено

65 °C

NVIDIA GeForce RTX 4060 Laptop GPU

Мощность

Измерено

62 W

из лимита 115 Вт

PCIe

Измерено

Gen4 x8

На GPU 3 701,6 МБ/с

CPU

Измерено

50 %

AMD Ryzen 9 7945HX with Radeon Graphics · 16 cores · 32 threads

Чтение диска

Измерено

0,0 МБ/с

Запись 0,2 МБ/с

Заполнение контекста

76%

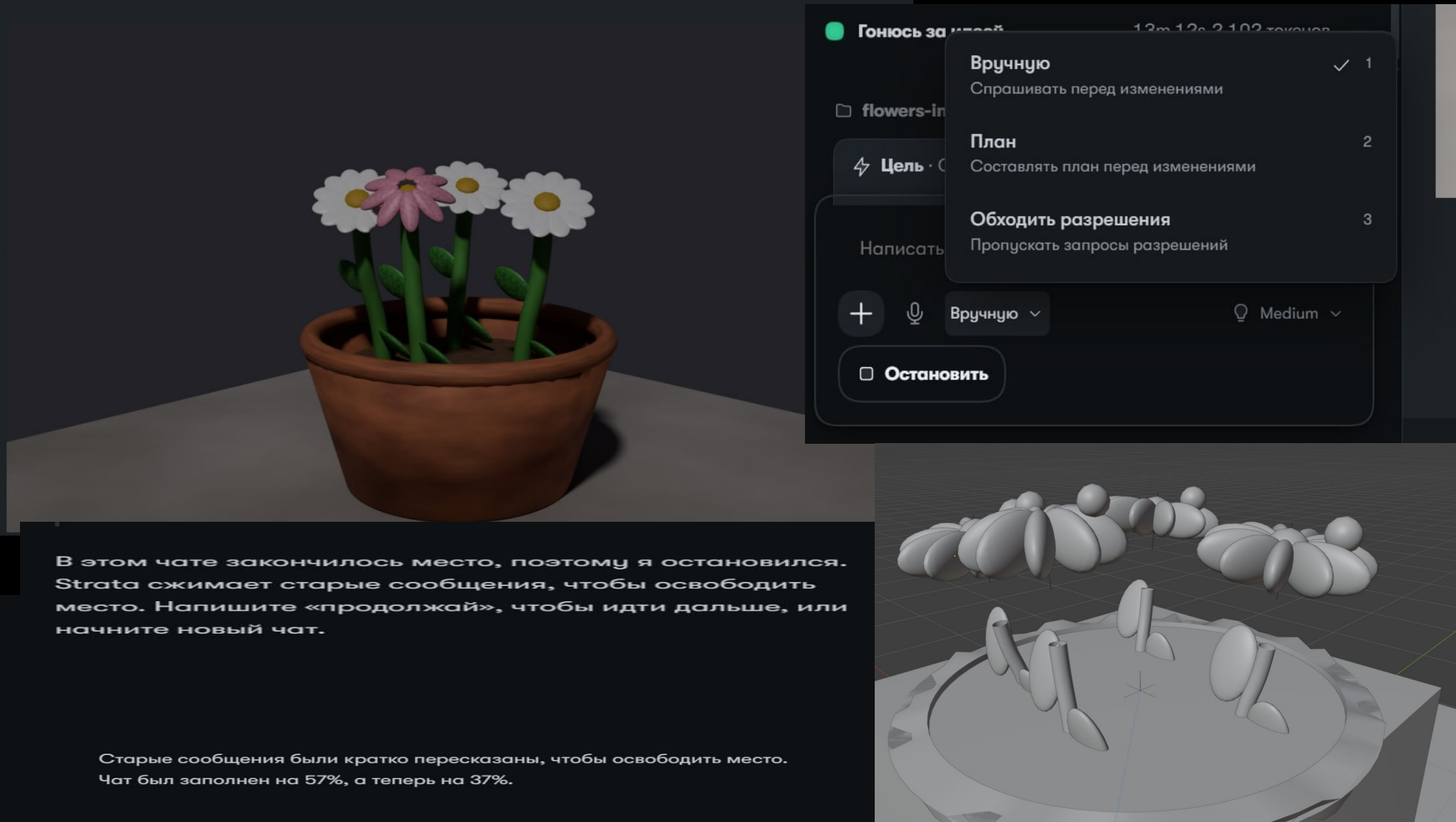
25k / 32K

Эксперты в VRAM — 553 · 0,8 GB

Недавние запросы

Показать все (141)

ВРЕМЯ	СТАТУС	ЗАПРОС	ИСПОЛЬЗОВАНО ПОВТОРНО	ВЫВОД	ТОКЕНОВ/С	ДОЛЯ ПОПАДАНИЙ В VRAM	ДЛИТЕЛЬНОСТЬ
01:00:21	Готово	15 132	6 126	2 867	56,2	—	1m 12s
00:58:57	Готово	23 606	15 589	2 867	43,6	—	1m 24s
00:57:08	Готово	28 939	15 589	2 359	30,4	—	1m 48s


Гонюсь за
Вручную
Спрашивать перед изменениями
1
План
Составлять план перед изменениями
2
Обходить разрешения
Пропускать запросы разрешений
3
flowers-in
Цель
Написать
Вручную
Medium
Остановить
В этом чате закончилось место, поэтому я остановился. Strata сжимает старые сообщения, чтобы освободить место. Напишите «продолжай», чтобы идти дальше, или начните новый чат.
Старые сообщения были кратко пересказаны, чтобы освободить место.
Чат был заполнен на 57%, а теперь на 37%.

Применяю правки: тонкие лепестки, более слабое освещение, экспозиция, толще стебли.



В этом чате закончилось место, поэтому я остановился. Strata сжимает старые сообщения, чтобы освободить место. Напишите «продолжай», чтобы идти дальше, или начните новый чат.

PY	**make_flower_pot.py** 15 KB	Открыть
BLEND	**flower_pot.blend** 239 KB	Открыть
BLEND1	**flower_pot.blend1** 234 KB	Открыть
PNG	**render.png** 870 KB	Открыть



flowers-in-pot

Цель · Create a textured flower pot with flowers model in Blender (headless s...

Написать Strata...

ВЕРСИЯ

Сохранённые настройки

Управление моделями

Swift IQ2_XS Скачано — 36.5 GB RAM + VRAM

РЕКОМЕНДУЕТСЯ	RAM + VRAM	СКАЧАТЬ
IQ3_S (GSQ-RCO) · выбор для этого ПК Скачано	**51.3 GB**	83.6 GB
Gyro-S · самая маленькая	**27.6 GB**	58.5 GB
✓ **Swift IQ2_XS** · хороший баланс Скачано	**36.5 GB**	68 GB
Unsloth UD-Q4_K_XL · выше качество Использует SSD 21 ГБ	**78 GB**	111.3 GB

Модели Swift 3

✓ **Swift IQ2_XS** Скачано	**36.5 GB**	68 GB
Swift IQ3_XXS	**43.9 GB**	75.8 GB
Swift IQ3_S	**51.3 GB**	83.6 GB

Модели GSQ-RCO 5

Модели NVFP4 3

Модели Gyro 2

Модели Unsloth 5

Модели Qwen3.6 6

Использовать свой GGUF...

Этот компьютер

NVIDIA GeForce RTX 4060 Laptop GPU · 8 GB видеопамяти · 63.7 GB ОЗУ

Видеопамять **8 GB** из 8 GB

ОЗУ **33.5 GB** из 63.7 GB

Полностью помещается в память.

Файлы модели

Strata готовит все части, черновой слой и общие файлы. Уже скачанные файлы используются повторно.

`F:\Strata\models`

Настройки хранилища

Другие локальные модели...

Обзор | Модели

Все | 30d | 7d



Qwen3.8-Flash-Next (Unsloth) UD-Q4_K_XL	49.2M ввод · 486k вывод	100.0%
DeepSeek-V4.1-Flash IST-DAS-2.00bpw	13k ввод · 2 вывод	0.0%

strata

Qwen3.8-Flash-Next (Unsloth) UD-Q4_K_XL · unsloth-ud-q4_k_xl · 128K · 29.10.2026, 11:20:31

Скорость
Измерено

52,2 t/s
Генерация последнего запроса

177 t/s
Обработка последнего запроса

Загрузка GPU
Измерено

0 %
NVIDIA GeForce RTX 5090 Laptop GPU

VRAM
Измерено

23,1 / 24 GB
Экспертов в кеше: 4 921

Темп. GPU
Измерено

36 °C
NVIDIA GeForce RTX 5090 Laptop GPU

Мощность
Измерено

6 W
из лимита 150 Вт

PCIe
Измерено

Gen5 x16
На GPU 0,0 МБ/с · idle Gen1

CPU
Измерено

5 %
Intel(R) Core(TM) Ultra 9 275HX · 24 cores · 24 threads

Чтение диска
Измерено

0,0 МБ/с
Запись 0,0 МБ/с

Энергия с запуска
0,0 Wh

Частота GPU
180 MHz

Температура SSD
36 °C

Project Maya v1.1.23

This PC | Choices | Packages | Engine | Model | Pack | Images | Config D:\Projects\ai\project-maya

Step: 2 of 8 · Your choices
Now: Waiting for your answer (below)
Progress: ━━ 1 of 8 steps done

Output

✓ CUDA toolkit 13.4: C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v13.4\bin\nvcc.EXE
✓ NVIDIA driver 610.88
✓ C++ compiler: Visual Studio Build Tools (C:\Program Files (x86)\Microsoft Visual Studio\2022\BuildTools\VC\Auxiliary\Build\vcvars64.bat)
✓ CMake 3.30
✓ Python 3.10.5 (D:\Projects\ai\project-maya\.venv)
✓ RAM: 191 GB, 183 GB available now
✓ CPU: Intel(R) Core(TM) Ultra 9 275HX (AVX2, 24 threads)

— Your choices

ПОТРАЧЕНО

maya-setup.log

Context?

Context length = how much text the model sees at once (the chat, files, tool output). A longer one takes VRAM from the expert cache, so answers get a little slower:

1) 8K tokens
2) 32K tokens *recommended*
3) 64K tokens
4) 128K tokens

[Enter→ Choose] [↑↓→ Move] [1-4→ Pick] [Tab→ Steps] [Ctrl+C→ Stop]

Движок

Как это работает: двойной маппинг

https://habr.com/ru/articles/1101852/

- mmap() файла с весами в RAM → дешевый page’ing и стриминг весов с диска
- Video shared memory → rebar → веса в RAM видны GPU, результат подкачки страниц виден на скорости шины
- (старые движки: трансферы в dedicated VRAM двухфазные и дорогие → стриминг горячих весов не делали)



Истории успеха

- RAM: 64G
- VRAM: 8G RTX4060 mobile
- AMD Ryzen 9 7945HX 16 cores 32 threads
- Быстрый SSD

  → 30..60 токенов/сек, Qwen-3.8-Flash-Next, 32k контекста одной суб-сессии, длинные сессии



Истории успеха

- 64G RAM
- 2xTesla M40 24G = 48G VRAM
- Уродская мать: только один разъем x16, остальные x1
- 64к контекста
- Человек очень доволен



Сценарии

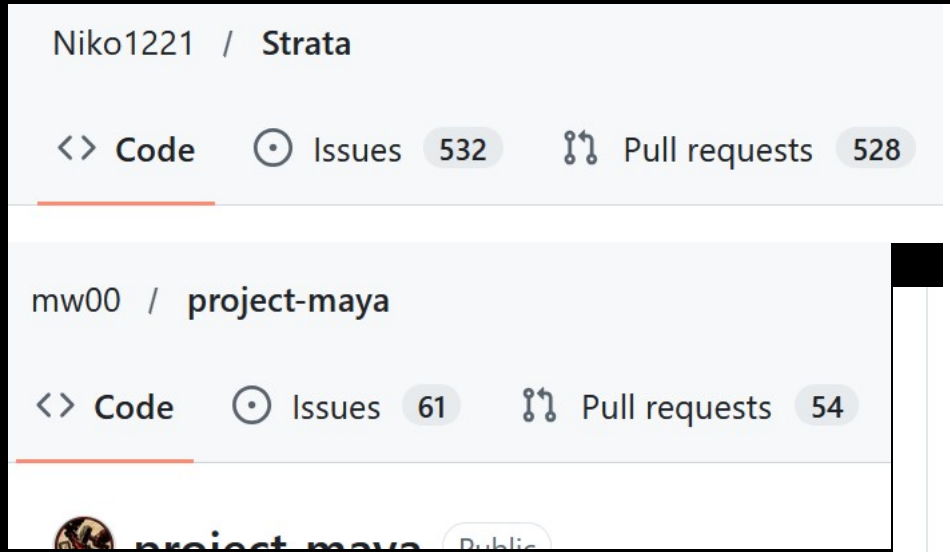
- Сделай рефакторинг в большом проекте (100k строк)
- Нарисуй картинку в Блендере процедурно
- Найди данные в сети

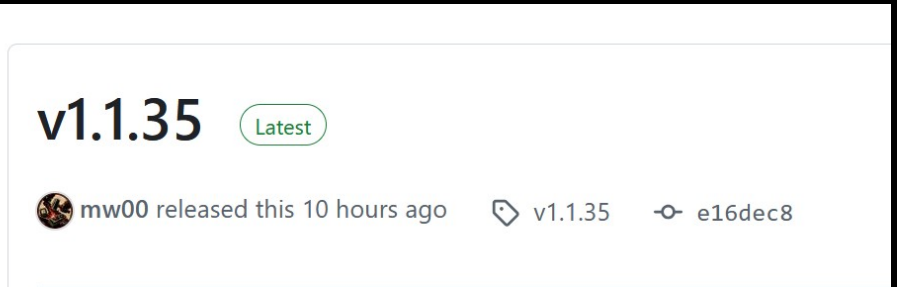
→ под ключ!

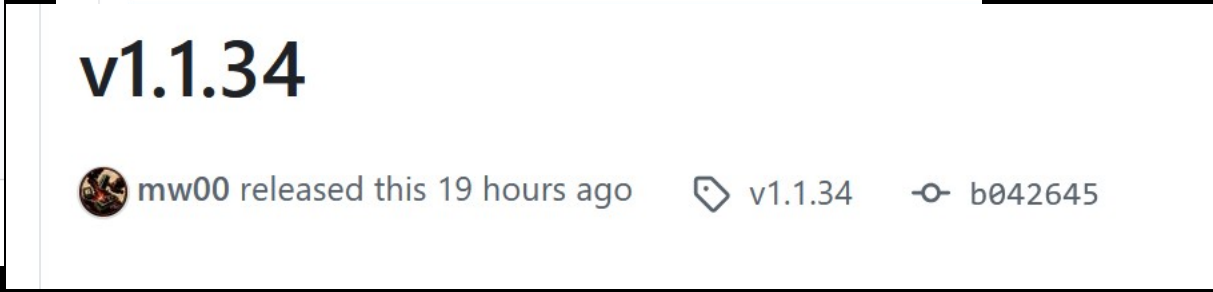


Проблемы

- Это вайб-кодят с гигантской скоростью







Проблемы

- Maya не взлетела
- Strata достаточно стабильна, но не всё из рекламы в ней работает



Минусы

Амортизация железа:

- Надо много VRAM для комфортного контекста (не меньше 24G)
- Греется
- 24/7/365 работать не выйдет — см.наработку на отказ карт для майнинга