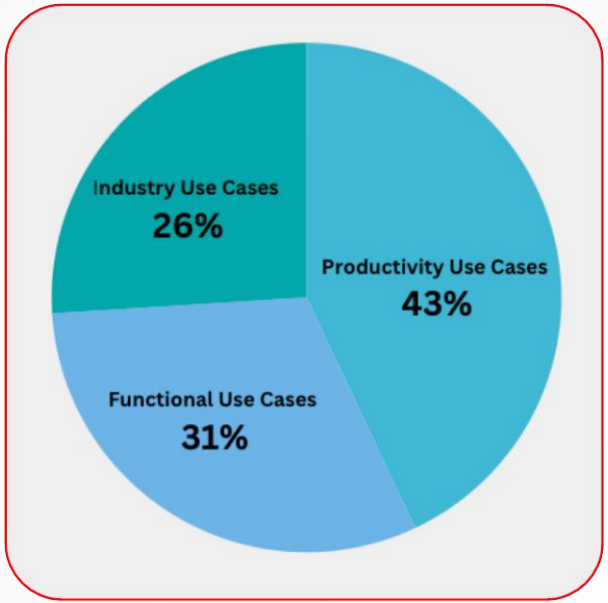
Пехтерев Алексей
Руководитель группы специальных проектов

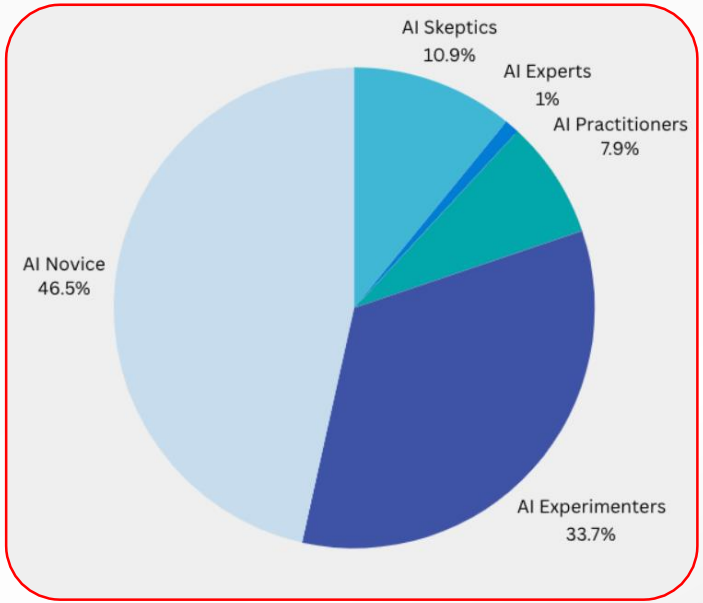
Парамонов Алексей
Ведущий программист

Особенности внедрения собственного LLM-сервиса внутри компании

Вебинар

ИИ агенты про генеративный ИИ





Key Generative AI Statistics and Trends for 2025 [as of May 28, 2025]

О чем сегодня поговорим

Часть 1:

- Публичный API vs Self-hosted LLM
- Как выбрать языковую модель
- Железо - GPU

Часть 2:

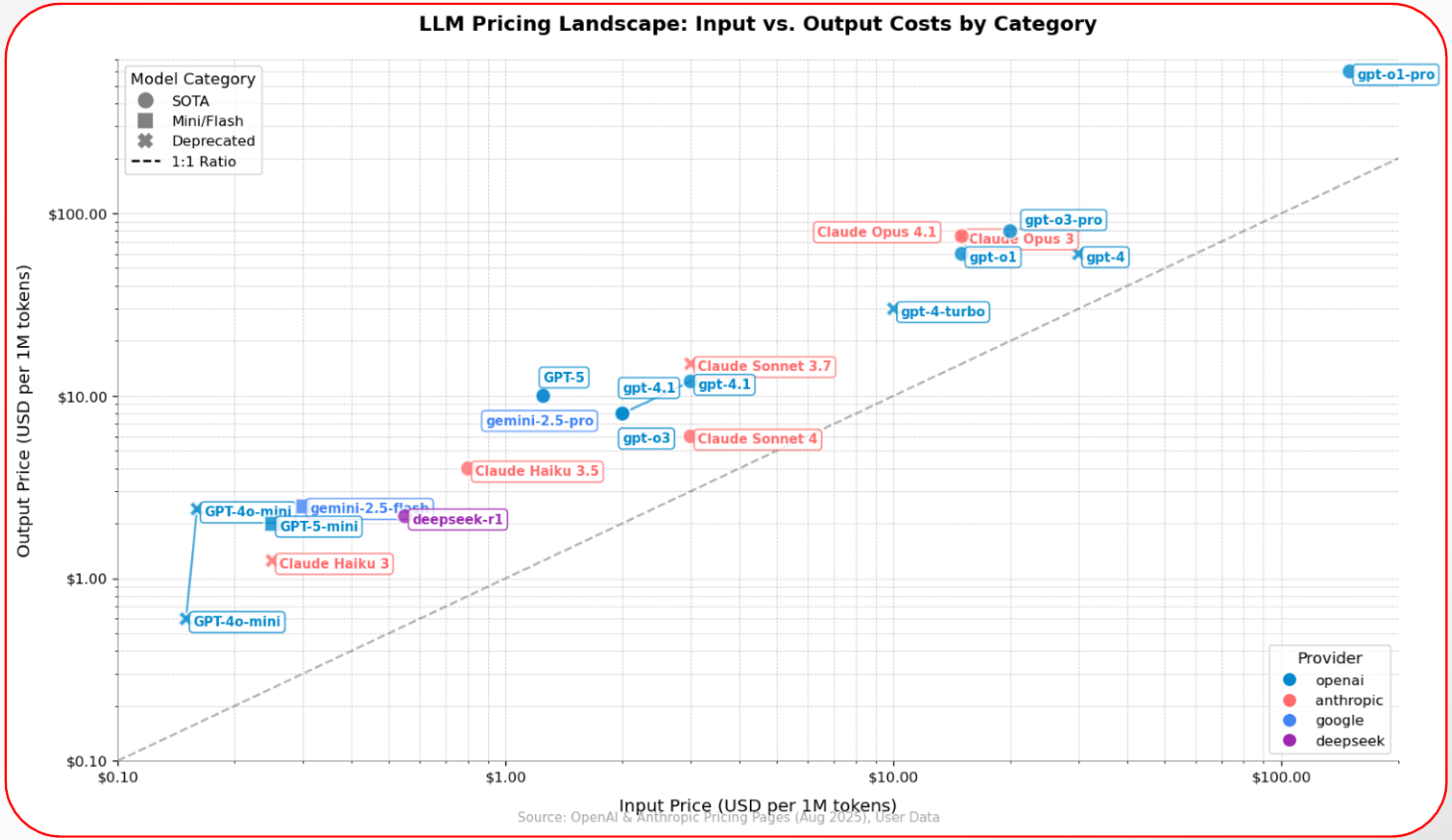
- Бэкенд для развертывания
- Зачем нужен LLM Gateway
- Как оценить нагрузку
- Метрики для мониторинга

Публичный API: Вендоры

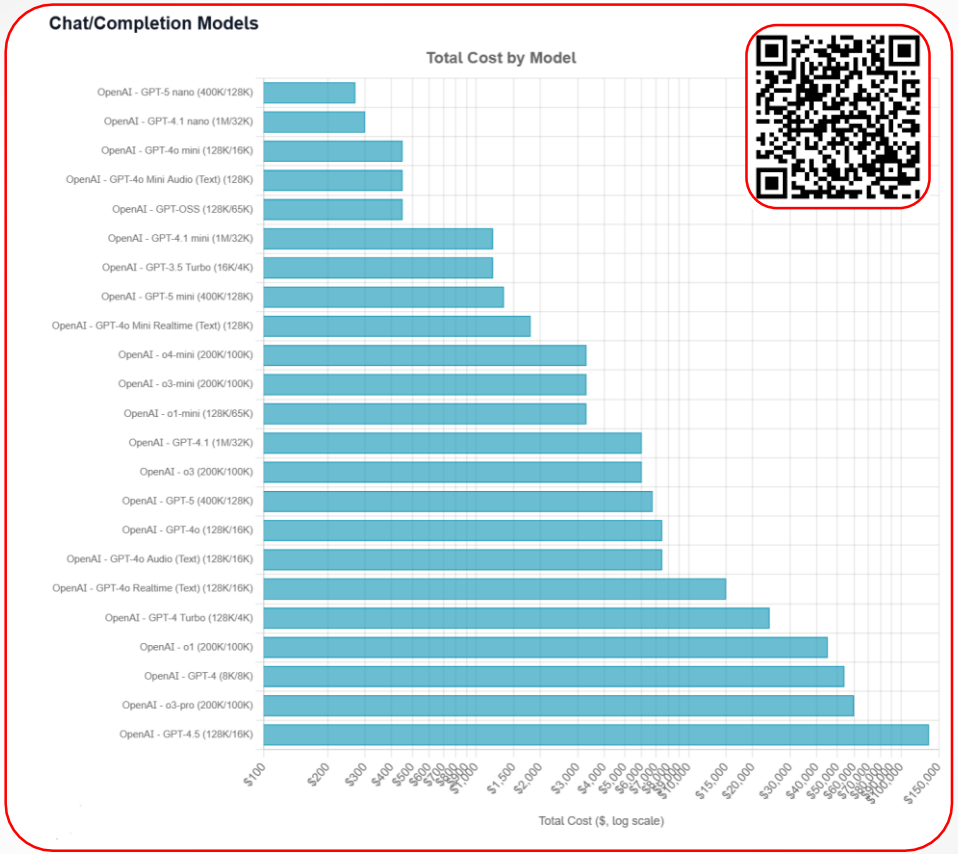
Публичный API: Очевидные плюсы

- State of the Art (SOTA) модели - лучшее качество из возможных
- Ready-to-use, простота интеграции
- Вендор отвечает за SLA
- Коммьюнити вокруг крупных вендоров
- Pay-as-you-go - платим за Input и Output токены

Публичный API: Цена вопроса



Публичный API: Примерный расчет



Daily Active Users: 1000
Messages per User: 10
Input / Output tokens: 2000 / 2000

Monthly price (gpt-5): 6 750 $
Monthly price (gpt-5-mini): 1 350 $

Calculated costs (per 1k queries):

Estimated search cost: **$15.000**

Estimated LLM cost: **$0.600**

Ratio (Search cost / LLM cost): **25.0x**

Публичный API: Лимиты

TIER	RPM	TPM	BATCH QUEUE LIMIT
Free		Not supported	
Tier 1	500	30,000	90,000
Tier 2	5,000	450,000	1,350,000
Tier 3	5,000	800,000	100,000,000
Tier 4	10,000	2,000,000	200,000,000
Tier 5	15,000	40,000,000	15,000,000,000

TIER	QUALIFICATION	USAGE LIMITS
Free	User must be in an allowed geography	$100 / month
Tier 1	$5 paid	$100 / month
Tier 2	$50 paid and 7+ days since first successful payment	$500 / month
Tier 3	$100 paid and 7+ days since first successful payment	$1,000 / month
Tier 4	$250 paid and 14+ days since first successful payment	$5,000 / month
Tier 5	$1,000 paid and 30+ days since first successful payment	$200,000 / month

Публичный API: Priority Tier

Requirements to advance tier

Usage Tier	Credit Purchase	Max Usage per Month
Tier 1	$5	$100
Tier 2	$40	$500
Tier 3	$200	$1,000
Tier 4	$400	$5,000
Monthly Invoicing	N/A	N/A

Committing to Priority Tier will involve deciding:

- A number of input tokens per minute
- A number of output tokens per minute
- A commitment duration (1, 3, 6, or 12 months)
- A specific model version

ANTHROP\C

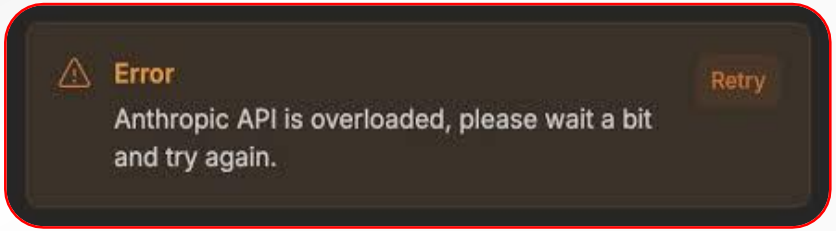
Tier 1 Tier 2 Tier 3 Tier 4 Custom

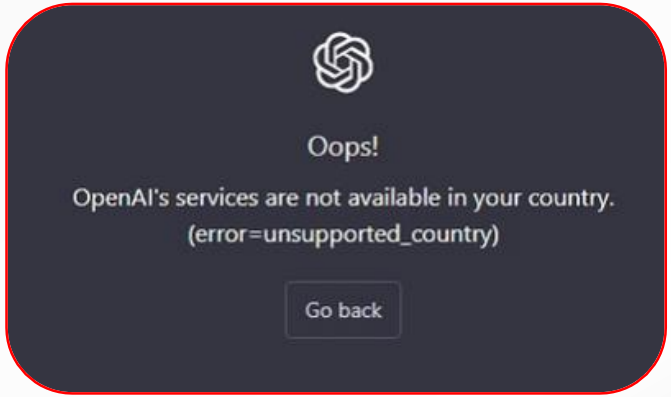
Model	Maximum requests per minute (RPM)	Maximum input tokens per minute (ITPM)	Maximum output tokens per minute (OTPM)
Claude Opus 4.x*	4,000	2,000,000	400,000
Claude Sonnet 4	4,000	2,000,000	400,000
Claude Sonnet 3.7	4,000	200,000	80,000
Claude Sonnet 3.5 2024-10-22	4,000	400,000†	80,000
Claude Sonnet 3.5 2024-06-20	4,000	400,000†	80,000
Claude Haiku 3.5	4,000	400,000†	80,000
Claude Opus 3	4,000	400,000†	80,000
Claude Sonnet 3	4,000	400,000†	80,000
Claude Haiku 3	4,000	400,000†	80,000

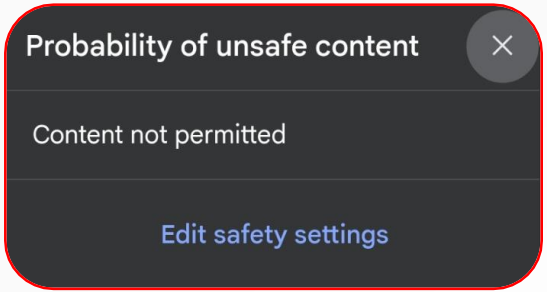
** - Opus 4.x rate limit is a total limit that applies to combined traffic across both Opus 4.0 and Opus 4.1.*

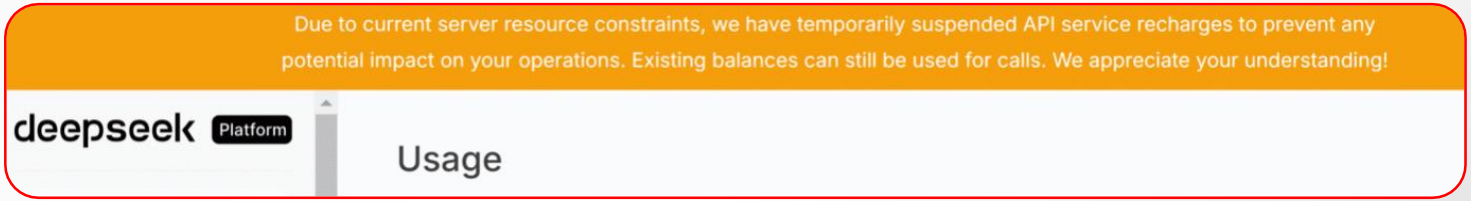
† - Limit counts `cache_read_input_tokens` *towards ITPM usage.*

Публичный API : Очереди, доступность



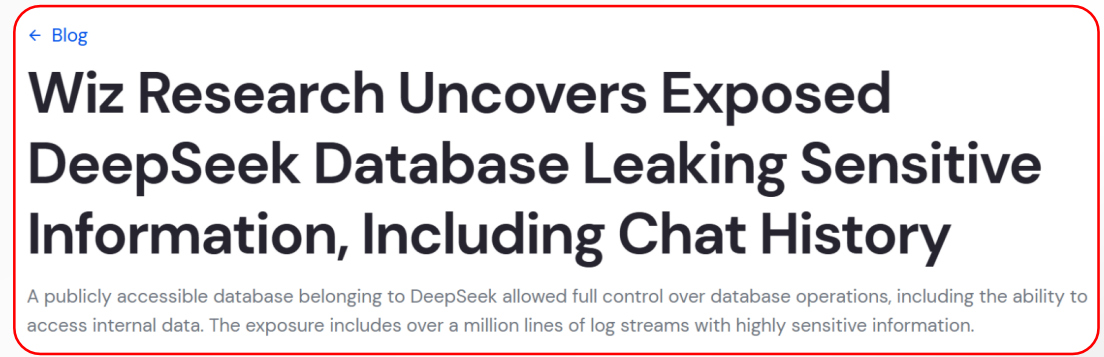






Главная неоднозначность. Приватность данных

Вебинар: “Остается ли в тайне ваш диалог с LLM? Как использовать большие языковые модели безопасно!”



Сравним подходы side-by-side.

	Публичный API	Собственный хостинг
Данные	Отправляются 3-му лицу. **Риск утечек.**	Остаются внутри компании. **Полный контроль.**
Затраты	Оплата за токен. **Зависят от объемов использования.**	Фиксированные. **Прогнозируемы.**
Кастомизация	Ограниченная. **Требует доп затрат.**	Глубокая. **Конкурентное преимущество.**
Скорость	Зависит от провайдера. **Возможны задержки.**	Прогнозируема и контролируема.
Усилия	**Минимальные.** Быстрый старт.	**Высокие.** Нужна команда и время.
Качество	**State of the Art.**	Уступают по качеству.
Выбор для	**Прототипов и данных без ограничений**	**Ноу-хау продуктов/фичей и sensitive данных.**

Чек-лист: Когда выбрать self-hosted вариант?

☐ Работаем со **строго конфиденциальными данными** (PII, коммерческая тайна и тд)?

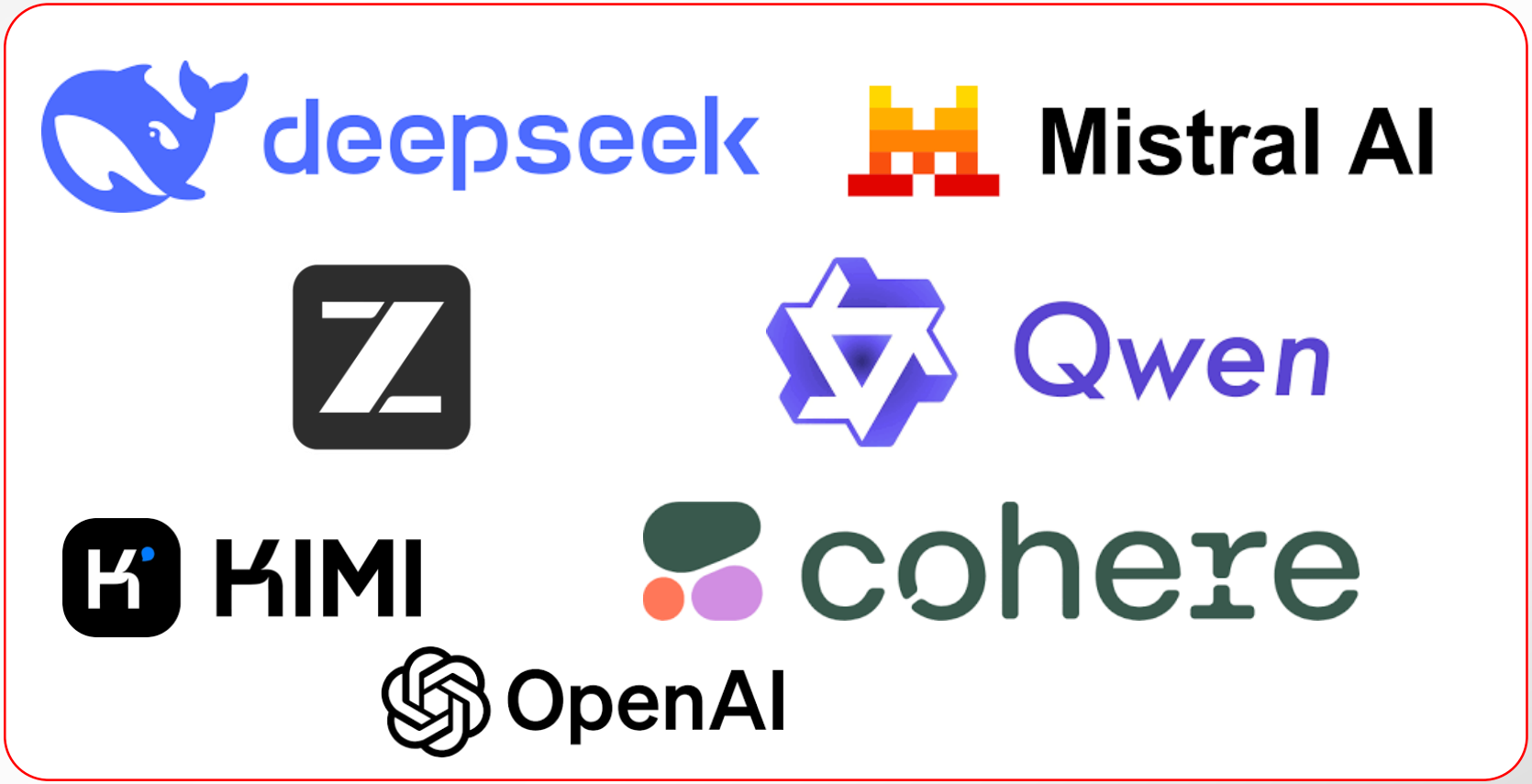
☐ Планируется **высокая и постоянная нагрузка длительное время**, при которой расходы на публичный API станут невыгодными?

☐ Нужен экспертный продукт/фича в **уникальной нишевой сфере** (внутренняя база знаний, агентские системы, автоматизирующие внутреннюю экспертизу)?

☐ Для продукта/фичи хотим сами **контролировать минимальную задержку**?

☐ Не хотим принимать **риски передачи данных** 3-му лицу?

Первая дилемма. Как выбрать OSS модель



Первая дилемма. Как выбрать OSS модель

Цель — найти не «лучшую» модель, а «правильную» именно для вас.

Четыре ключевых критерия оценки:

1. **Качество на задаче:** Насколько хорошо модель справляется именно с вашей задачей?
2. **Размер модели и железо:** Что вы можете себе позволить?
3. **Лицензия:** Можете ли вы легально использовать модель в коммерческих целях?
4. **Экосистема и поддержка:** Насколько активно сообщество вокруг модели?

Качество: LLM Benchmarks

Бенчмарк	Что оценивает
General Reasoning — Humanity's Last Exam, MMLU+Pro, BIG-Bench Hard, GPQA, SuperGPQA, MMMU+Pro, HLE, CharXiv-Reasoning	Проверка знаний, логики и мультимодальных навыков
Conversation & Instructions — Scale MultiChallenge, COLLIE, IFEval	Оценка диалога и следования инструкциям
Coding & Math — MATH, AIME 2025, FrontierMath, HMMT, SWE-bench (Full/Lite/Verified/Live), Aider Polyglot, GSM8K	Генерация кода, решение инженерных задач, математика
Safety & Ethics — TruthfulQA, HHH, AdvBench, AILuminate	Проверка guardrails, честности и отсутствия предвзятости
Specialized Suites — BenchHub, LegalEval-Q, IberBench, LocalValueBench, HealthBench +Hard	Унифицированная, доменно-специфическая и мультиязычная оценка

Качество: LLM Arena Leaderboard

Lmarena.ai

Lmarena.ru

*Упоминаемый продукт произведен компанией Meta, которая признана экстремистской и запрещена в РФ

Качество: LLM Arena Leaderboard

Lmarena.ai

Lmarena.ru

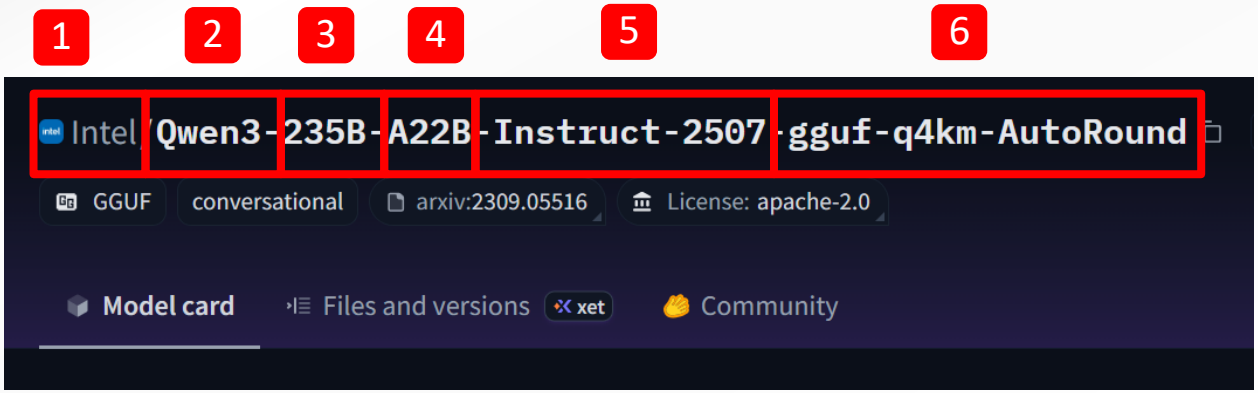
* Упоминаемый продукт произведен компанией Meta, которая признана экстремистской и запрещена в РФ

Качество: Базовые знания, специфика, свой бенчмарк

Универсальная модель или домен-специфичная:

- проприетарные SOTA модели - универсальны и лучшие по качеству
- не все задачи требуют универсальности знаний, а скорее наоборот, требуют специфики или экспертности
- это можно решить путем дообучения или построения RAG системы
- в основе не обязательно должна лежать сверх умная модель, достаточно хороших baseline знаний в домене вашей задачи
- специфика языка очень важна
- в идеале иметь свой бенчмарк

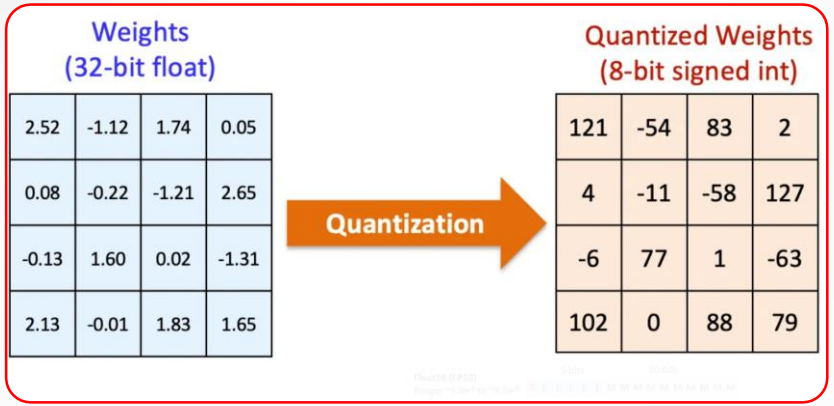
Железо: Количество параметров



1 – Владелец репозитория
2 – Название и версия модели
3 – Количество параметров (весов) модели
4 – Количество активных параметров (для MoE)
5 – Тип модели
6 – Формат квантизации

Hugging Face

Железо: Квантование и vRAM GPU



FP32 (Full Precision)	32 бит	4 байта
FP16 / BF16 (Half Precision)	16 бит	2 байта
INT8 (8-bit Quantized)	8 бит	1 байт
INT4 (4-bit Quantized)	4 бит	0.5 байт

Примерный расчет:

*vRAM = Количество параметров * Количество байт +* ***~20-25% на KV Cache****

Пример 1: **Qwen-235b-a22b**

*vRAM ~ 235 * 2 + 0.2 * (235 * 2) = 564 GB*

Пример 2 : **Qwen-235b-a22b-gguf-q4km**

*vRAM ~ 235 * 0.5 + 0.2 * (235 * 0.5) = 141 GB*

***минимум, если хотим динамически масштабироваться, то нужно больше*

Железо: Достаточно ли одной GPU?

Для качественного продакшена, к сожалению, нет.

Идеальный вариант:

- Дополнительная vRAM для обеспечения масштабируемости в периоды пиковой нагрузки
- Дополнительные GPU для обеспечения роллирования
- Дополнительные стенды как fallback
- Dev / Test стенды

Железо: Где взять GPU ресурсы?

Вариант А: Аренда (Облако)	Вариант Б: Покупка (On-Premise)
Плюсы: Оплата по факту, доступ к новейшим GPU (H100/A100), нет расходов на обслуживание. **Минусы:** Очень дорого при работе 24/7. **Провайдеры:** Yandex Cloud, Selectel, VK Cloud.	**Плюсы:** Фиксированная стоимость (CapEx), предсказуемый бюджет, максимальная безопасность. **Минусы:** Высокие первоначальные вложения, затраты на обслуживание, устаревание железа. **Железо:** NVIDIA A100/H100/L40S/etc (Enterprise), NVIDIA RTX 4090/5090 (Prosumer/Dev). **Альтернативы:** AMD ROCs, Intel XPU

Железо: Цены

GPU	Selectel	Yandex Cloud	VK Cloud	Покупка	Окупаемость
Nvidia H100	359 000 руб/мес	-	-	30 000 $	7 мес.
Nvidia A100	180 000 руб/мес	300 000 руб/мес	225 000 руб/мес	24 000 $	11 мес.
Nvidia L40S	-	-	149 000 руб/мес	12 500 $	7 мес.

Вывод: Арендуйте для экспериментов. Покупайте для предсказуемых, постоянных продакшн-нагрузок.

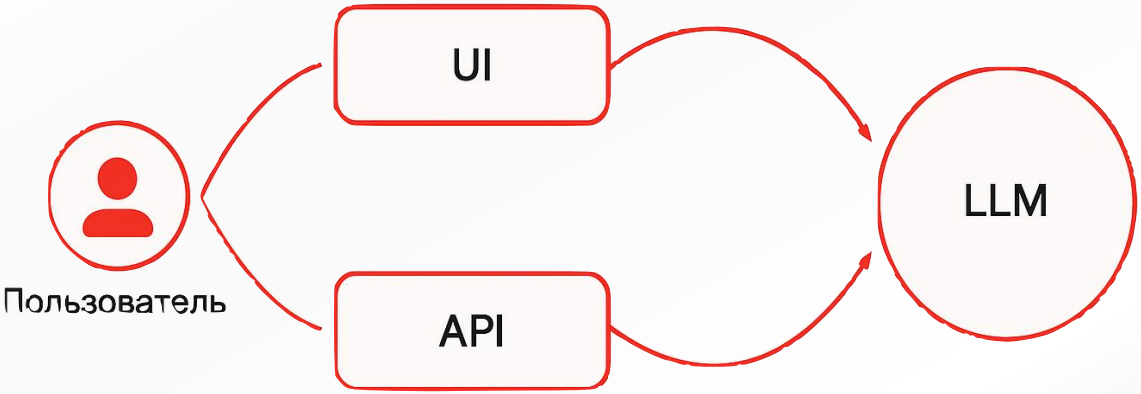
Лицензия: “читайтемелкийшрифт”.

Модель	Тип лицензии	Краткое описание
Qwen3	Apache 2.0	Свободная лицензия, разрешает коммерческое использование, модификацию, распространение, требует указания копирайта
Llama*	Собственная коммерческая лицензия Llama Community License	Лицензия с ограничениями; не считается открытым исходным кодом, содержит Acceptable Use Policy и ограничения по использованию.
Mistral Large	Mistral AI Research License	Лицензия Research, частично открытая; использование требует соблюдения условий, влияют санкции.
DeepSeek-R1	MIT	Открытая лицензия; позволяет коммерческое использование, модификацию, распространение, дистилляцию.
Gemma3-27B	Использование по лицензии от Google	Доступ предоставляется при принятии условий Google; не является открытой лицензией, содержит ограничения для «Hosted Services», влияют санкции.
gpt-oss	Apache 2.0	Свободная лицензия

* Упоминаемый продукт произведен компанией Meta, которая признана экстремистской и запрещена в РФ

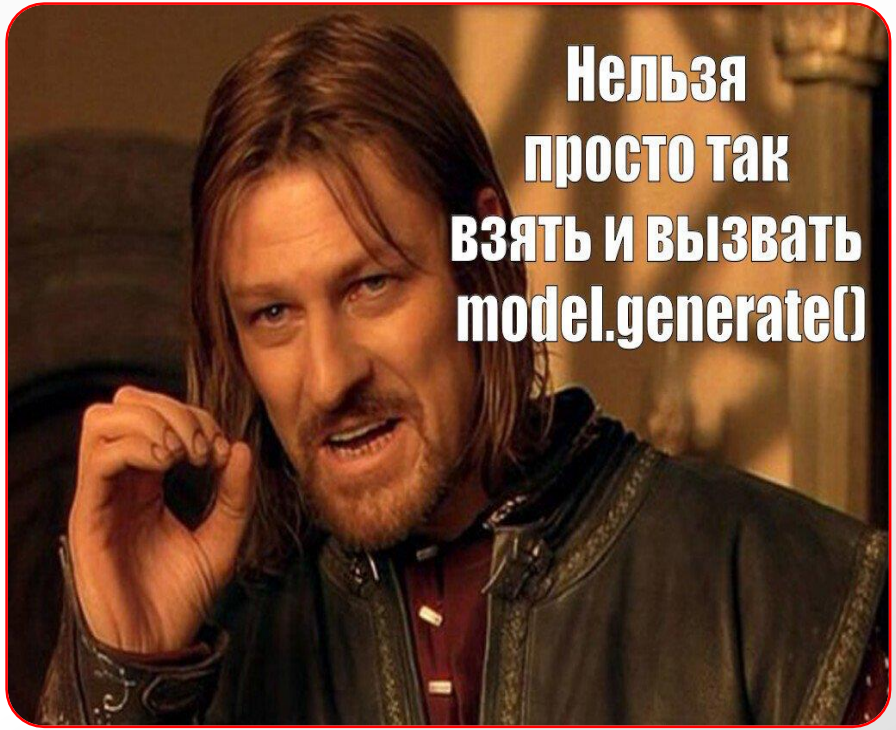
Что делать после выбора модели?

Как выглядит общение с языковой моделью для пользователя



Почему model.generate() не работает

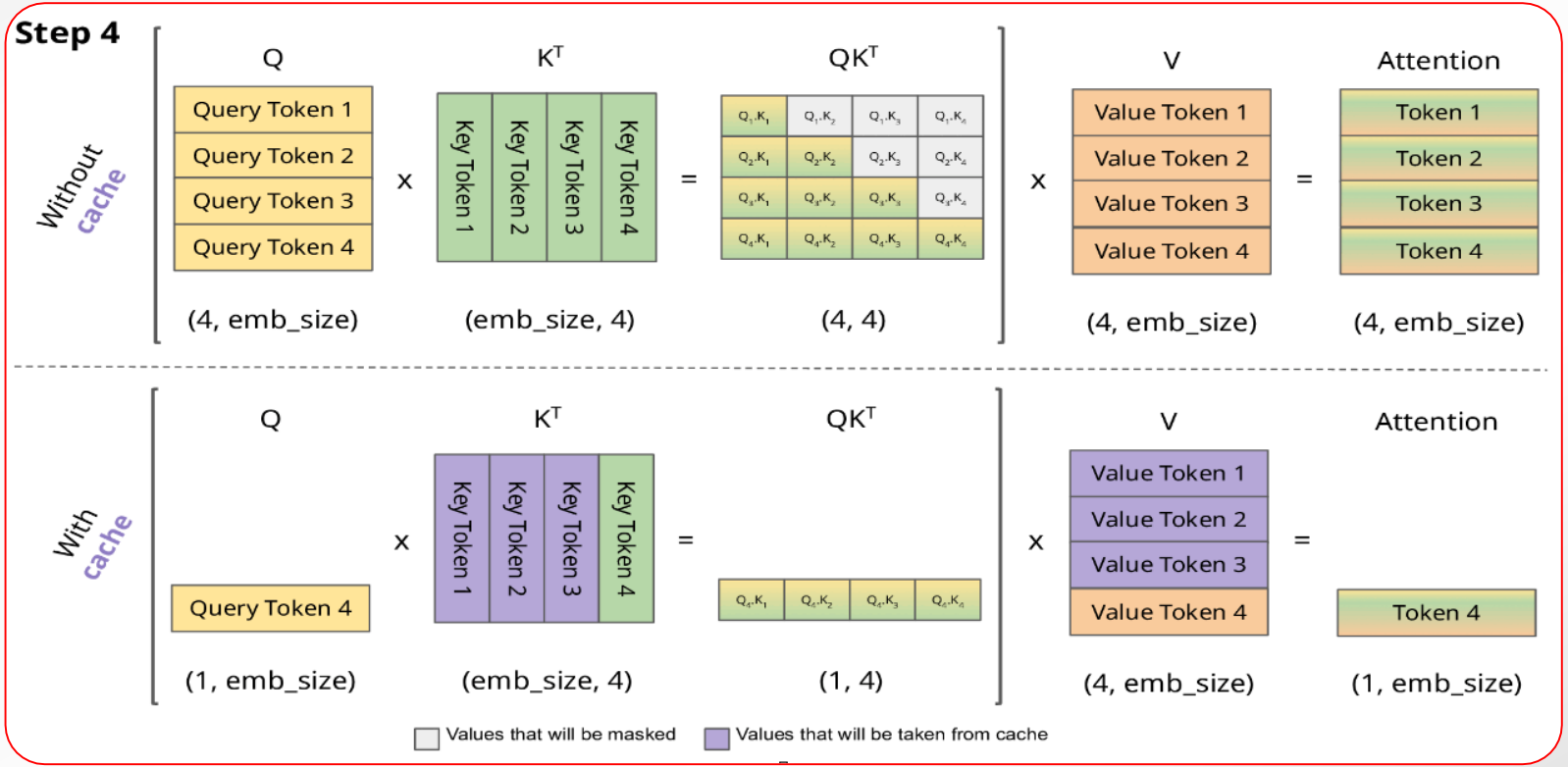
- Блокирующая генерация: только один запрос обслуживается в момент времени
- Нет динамического батчинга: GPU простаивает между запросами
- Высокие задержки при каждом новом запросе
- Python-оверехед и GIL мешают масштабировать службу
- Память резервируется под максимальную последовательность, фрагментируется



Сравнение движков для инференса

Бэкенд	Класс	Плюсы	Минусы	Где хорош
VLLM	Промышленный	PagedAttention, Continuous Batching, OpenAI API	Требует GPU, сложнее, чем Ollama	Продакшн с высокой нагрузкой
SGLang	Промышленный	Новые оптимизации, Python	Менее зрелый	Эксперименты, спец. модели
TensorRT-LLM	Промышленный	Макс. скорость, CUDA-оптимизация	Сложная сборка, закрытые части	Короткие запросы, low-latency
Ollama	Лёгкий	Простота, кроссплатформа	Нет оптимизаций, низкая нагрузка	Прототипы, локальные тесты
Llama.cpp	Лёгкий	Лёгкий, кроссплатформенный, квантование	Нет батчинга, медленнее на GPU	Edge, слабое железо
LM Studio	Лёгкий	GUI, API-режим	Ограничен Llama.cpp	Тесты, разработка, обучение

PagedAttention: эффективность памяти



vLLM: Continous batching

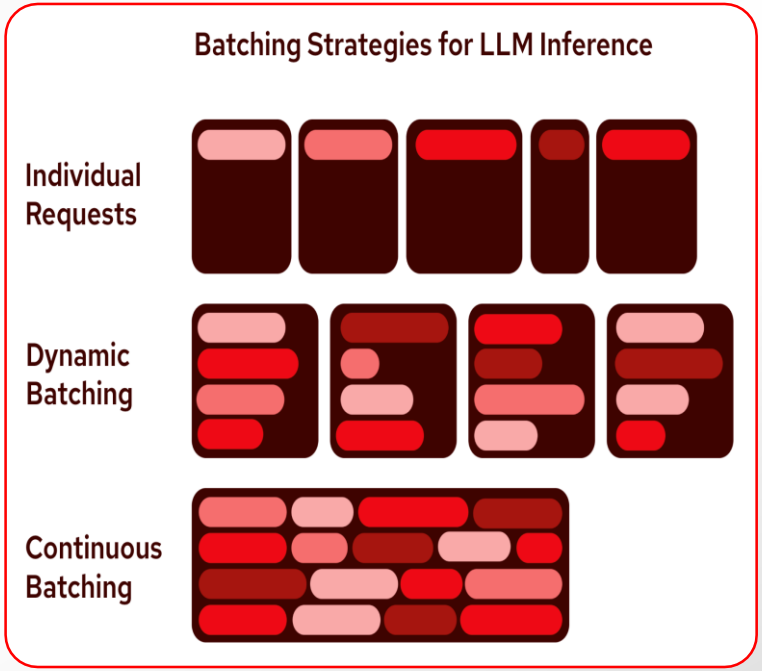
Идея: батч формируется динамически на каждом шаге декодирования — новые запросы добавляются, завершившиеся сразу уходят.

Зачем: убирает «head-of-line blocking», снижает пустой паддинг и держит GPU постоянно загруженной.
Как работает: разделяет этапы prefill и decode, интерливит их по шагам; слоты в батче мгновенно переиспользуются.

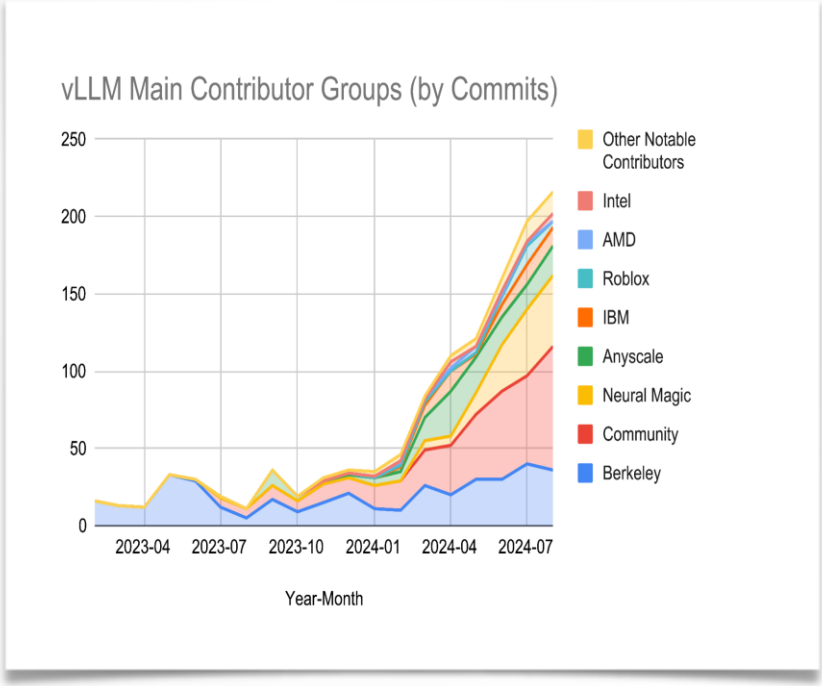
Эффект: выше throughput (tokens/s) при стабильном TTFT и предсказуемой p95/p99-латентности даже при разных длинах запросов.

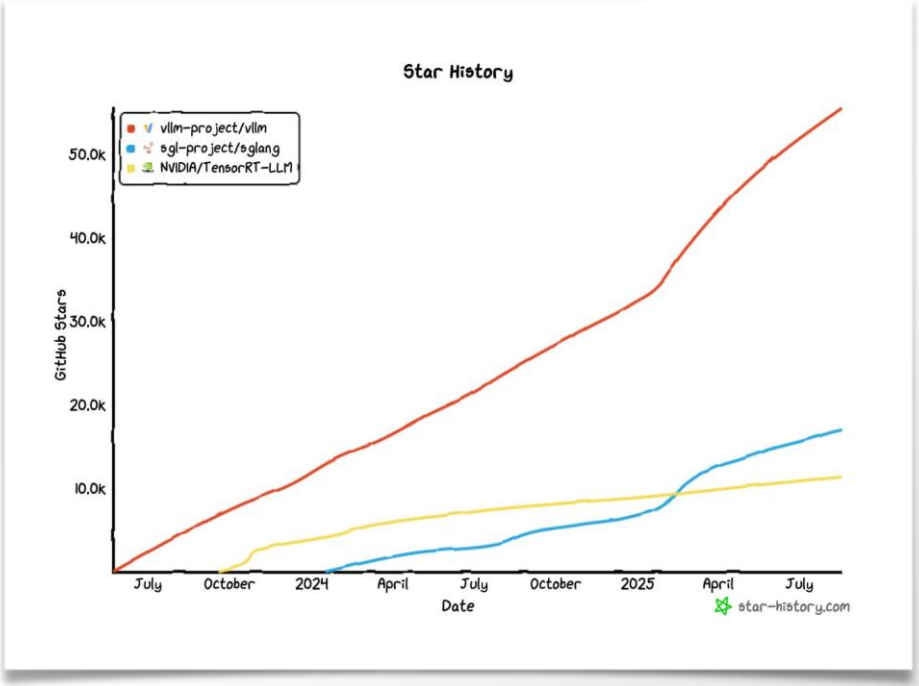
Когда особенно полезно: чат-инференс, стриминг ответов, RAG/интерактивные сценарии с «пилообразной» нагрузкой.

Итог: больше полезных токенов с каждого шага, меньше простоя и лучшая стоимость токена на проде.



vLLM: Комьюнити





Self-hosted LLM сервис. Это просто?

План действий:

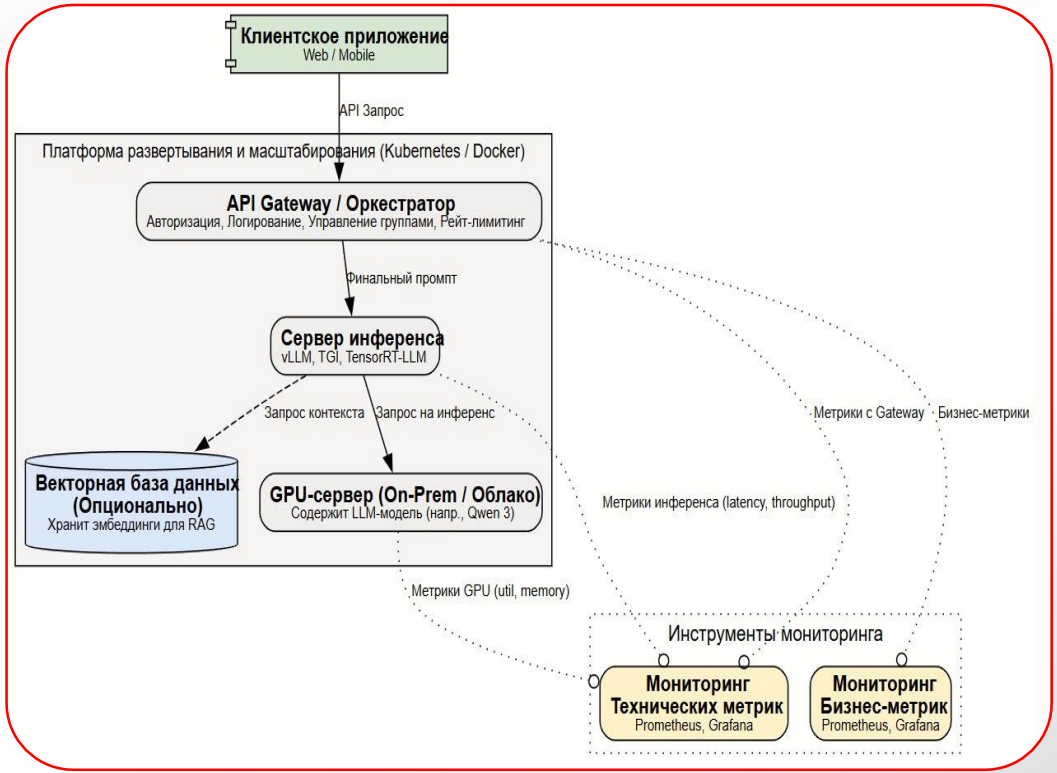
Шаг 1: Качаем модель с hugging face - `huggingface-cli download {model_name}`

Шаг 2: Разворачиваем - `vllm serve {model_name}`

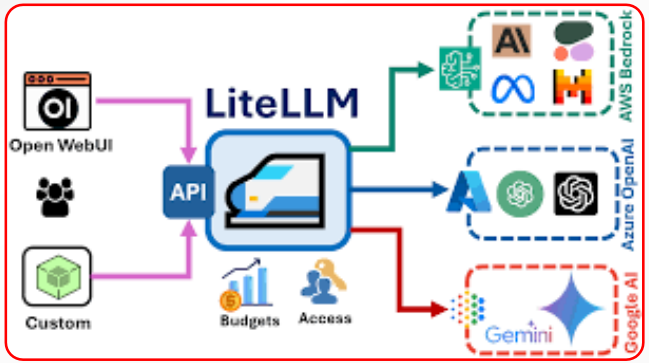
Шаг 3: Profit???

LLM Gateway / Прокси

- Единая точка доступа к нескольким моделям
- Балансировка нагрузки и fallback при ошибках
- Авторизация, квоты и учёт токенов
- Логирование и мониторинг запросов



Примеры LLM Gateway



ИЛИ

Kong

Кастомный LLM-прокси

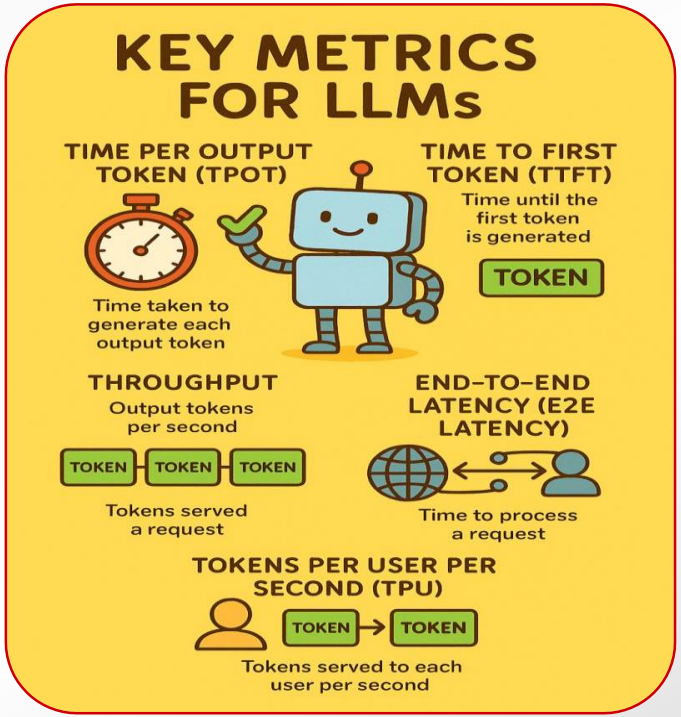
А также:

PortKey AI Gateway
MLFlow AI Gateway
Bifrost
...

Нагрузочное тестирование и оценка

Ключевые метрики:

- **TTFT** — время до первого токена; это ощущаемая «быстрота ответа».
- **Inter-Token Latency / TPOT** — среднее время между токенами; показывает, насколько плавно «пишет» модель.
- **Request latency (p95/p99)** — длительность всего запроса; следим за «хвостами», а не только за средним.
- **Throughput** — сколько запросов/токенов в секунду выдаёт сервис; это реальная пропускная способность.
- **Success/Error/Incomplete** — доли успешных, ошибочных и оборванных запросов; это про надёжность.

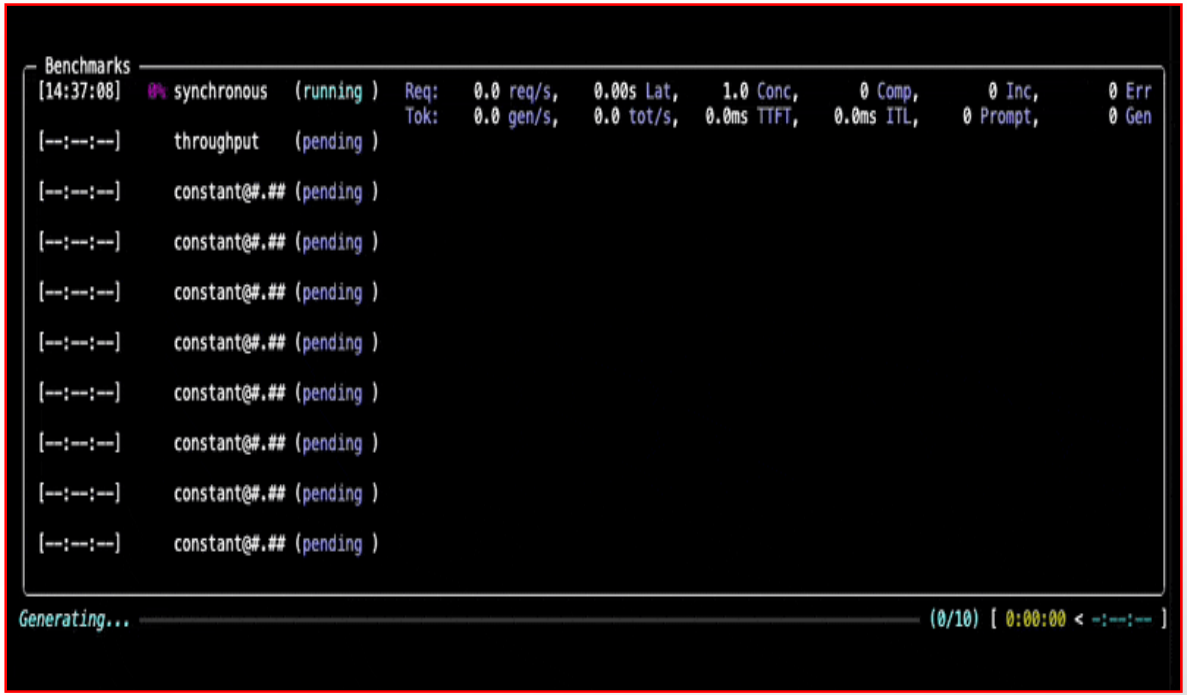


Сбор данных для рабочей нагрузки

- **Цель:** смоделировать трафик «как в проде», а не абстрактные запросы.
- **Источники:** FAQ/шаблоны саппорта, синтетика, «адверсариалка» (намеренно сложные промпты).
- **Сценарии:**
  - Короткие чат-вопросы,
  - Длинные контексты (документы/RAG),
  - Инструментальные вызовы/функции (tool calling)
  - Генерация и анализ кода

Процесс тестирования

- Поднимаем сервер инференса
- Задаём профиль нагрузки и данные
- Инструмент собирает отчёт
- Сверяемся с SLO
- Сохраняем отчет



Выводы из тестирования

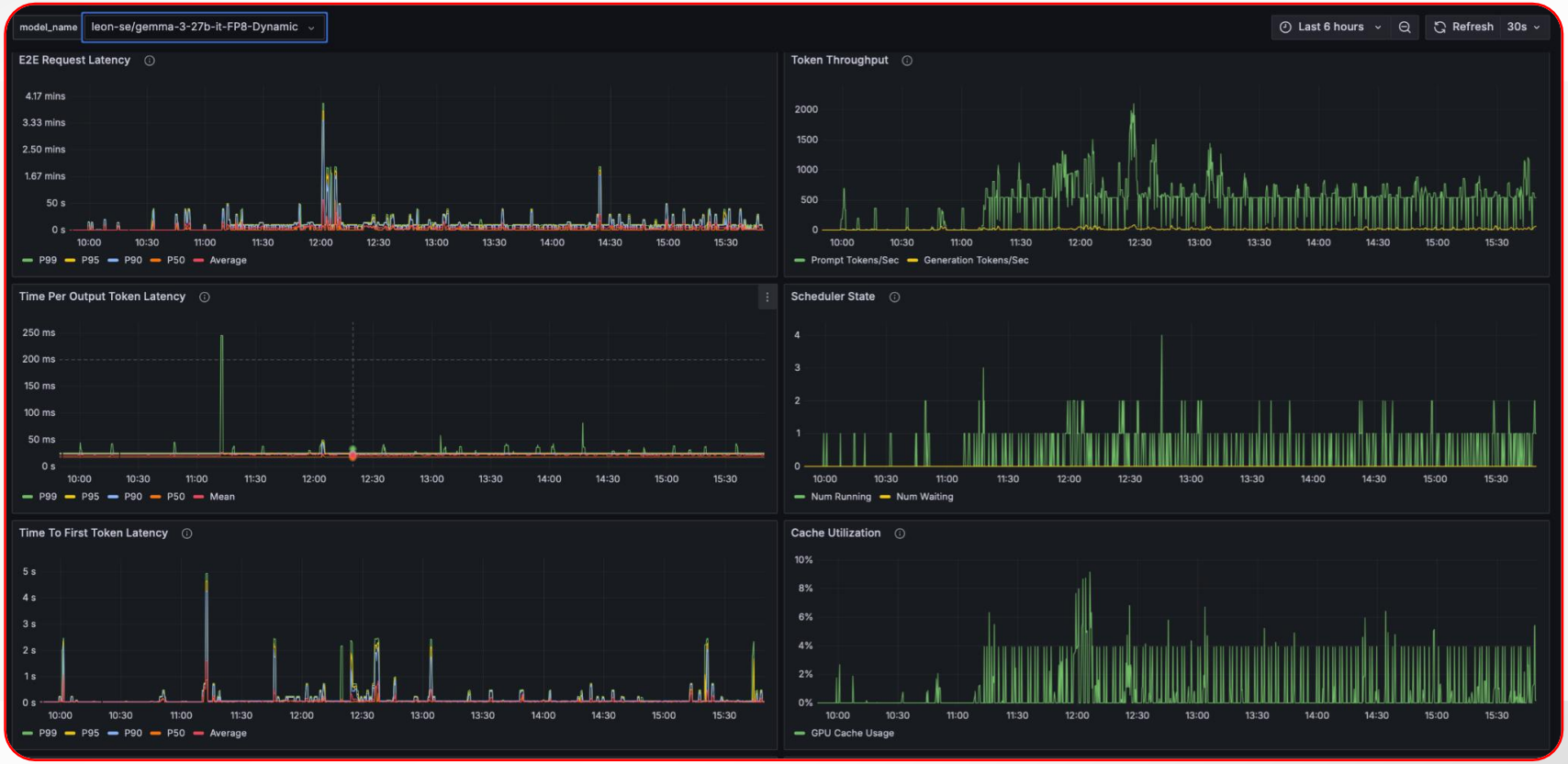
Benchmarks Info:

Metadata				Requests Made			Prompt Tok/Req			Output Tok/Req			Prompt Tok Total			Output Tok Total		
Benchmark	Start Time	End Time	Duration (s)	Comp	Inc	Err	Comp	Inc	Err	Comp	Inc	Err	Comp	Inc	Err	Comp	Inc	Err
synchronous	09:21:59	09:24:36	157.2	99	0	1	512.1	0.0	512.0	256.0	0.0	164.0	50701	0	512	25344	0	164
throughput	09:24:36	09:24:42	5.5	97	0	3	512.1	0.0	512.0	256.0	0.0	15.0	49677	0	1536	24832	0	45
constant@6.32	09:24:42	09:25:00	17.2	97	0	3	512.1	0.0	512.0	256.0	0.0	62.7	49676	0	1536	24832	0	188
constant@12.01	09:25:00	09:25:10	10.2	100	0	0	512.1	0.0	0.0	256.0	0.0	0.0	51213	0	0	25600	0	0
constant@17.69	09:25:11	09:25:19	8.0	98	0	2	512.1	0.0	512.0	256.0	0.0	121.0	50189	0	1024	25088	0	242

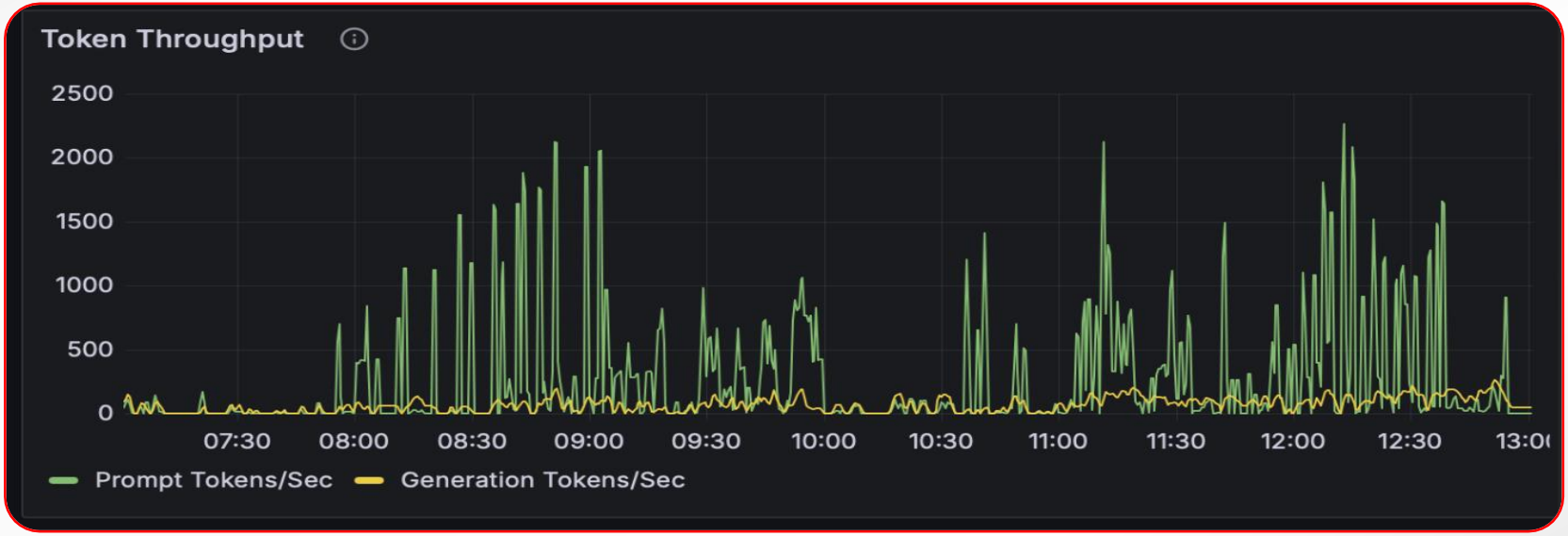
Benchmarks Stats:

Metadata	Request Stats		Out Tok/sec	Tot Tok/sec	Req Latency (ms)			TTFT (ms)			ITL (ms)			TPOT (ms)		
Benchmark	Per Second	Concurrency	mean	mean	mean	median	p99	mean	median	p99	mean	median	p99	mean	median	p99
synchronous	0.63	0.99	161.3	805.9	1.57	1.41	2.27	119.9	115.7	286.4	5.7	5.1	7.8	5.7	5.0	7.8
throughput	17.69	95.73	4529.7	22636.3	5.41	5.41	5.43	234.5	237.3	296.8	20.3	20.3	20.6	20.2	20.3	20.5
constant@6.32	5.63	15.56	1442.2	7207.2	2.76	2.76	2.96	63.8	62.7	84.2	10.6	10.6	11.4	10.5	10.5	11.3
constant@12.01	9.82	33.98	2512.6	12556.2	3.46	3.50	4.03	98.2	67.3	506.5	13.2	13.4	14.8	13.1	13.4	14.7
constant@17.69	12.30	48.86	3148.3	15733.0	3.97	3.95	4.27	84.0	77.3	137.2	15.3	15.2	16.5	15.2	15.2	16.4

Мониторинг: метрики и инструменты

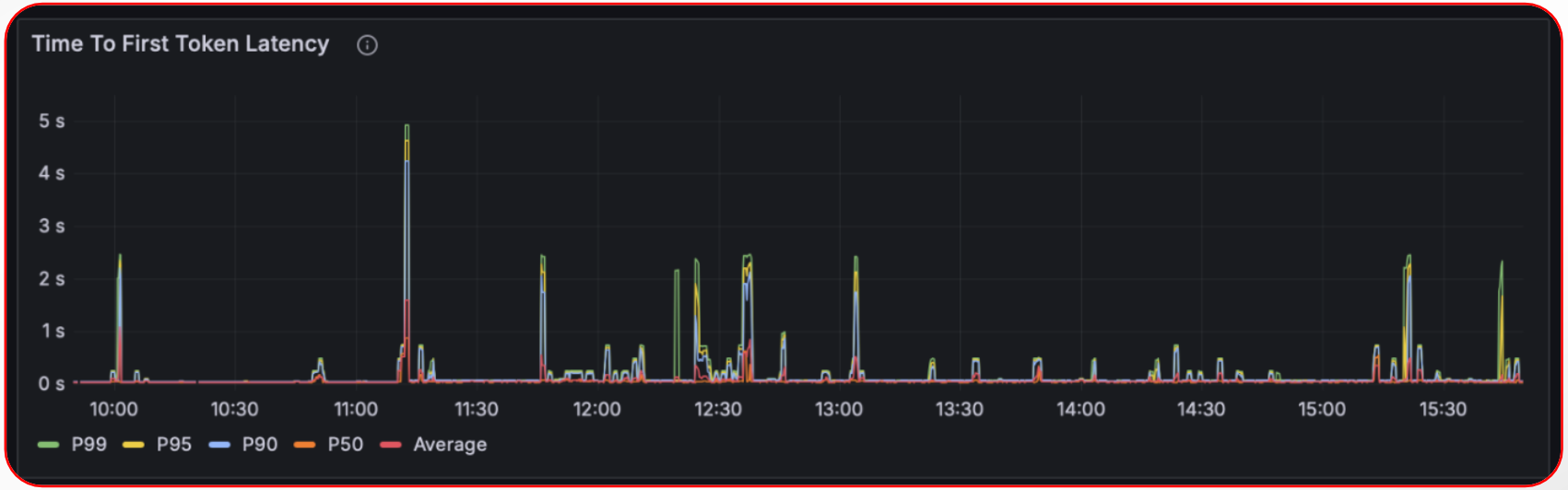


Token Throughput (Prompt vs Generation)



Интерпретация: зелёная линия — скорость «проглатывания» входных токенов (префилл), жёлтая — скорость генерации.

Время до первого токена (хорошо)

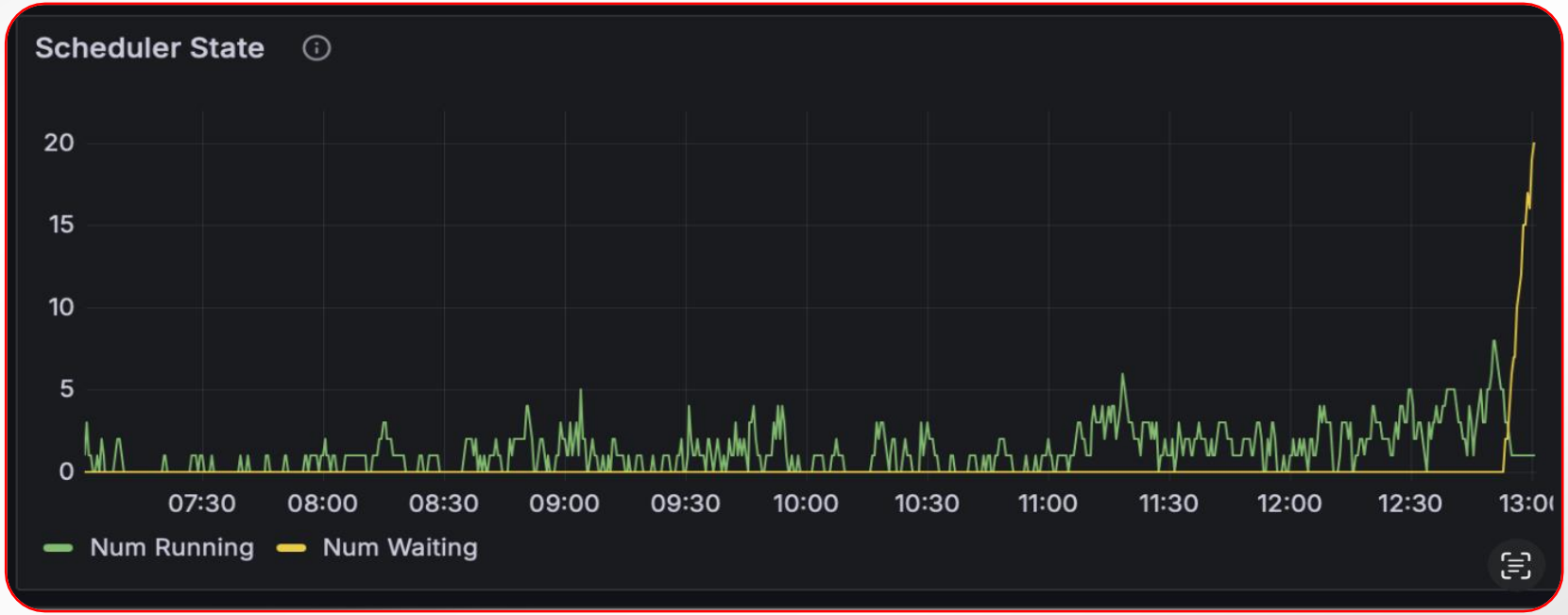


Интерпретация: пики — это или очередь у планировщика, или «дорогой» префилл (длинные промпты, RAG/инструменты), иногда — холодные старты.

Время до первого токена (плохо)

Пример критической ситуации, где возникают алерты.

Scheduler State (Num Running / Num Waiting)

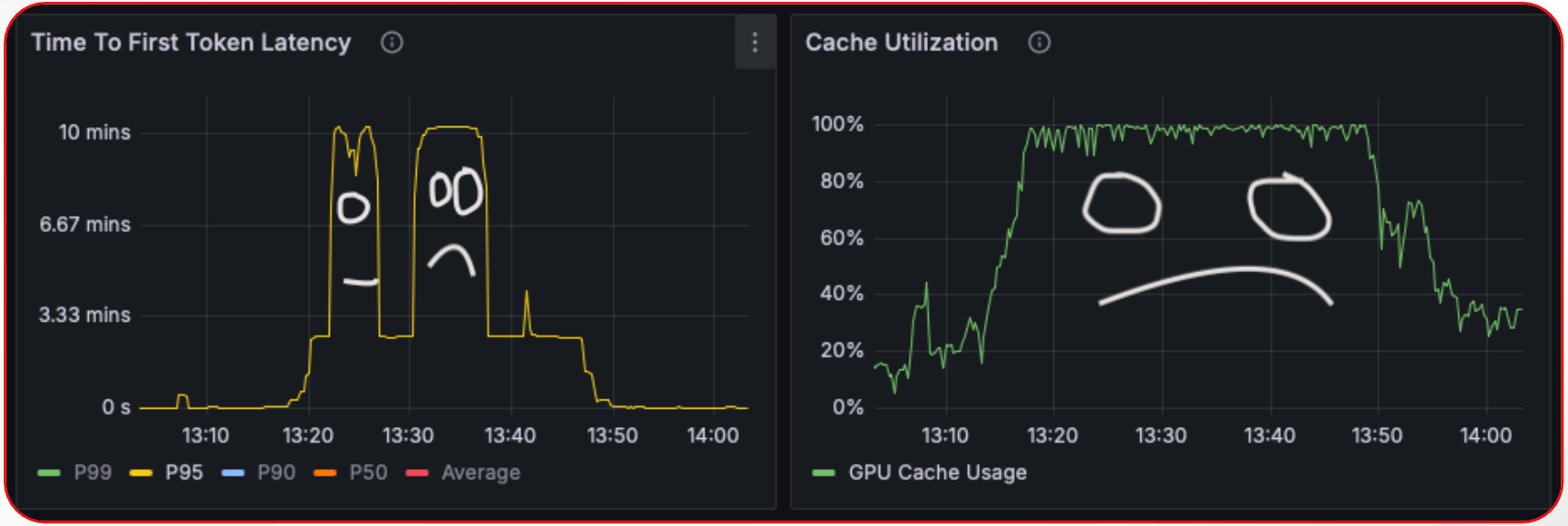


Интерпретация: зелёная линия — активные запросы, жёлтая — те, что стоят в очереди. Всплеск «waiting» к 13:00 — явная перегрузка.

GPU cache Utilization (KV-кэш моделей)

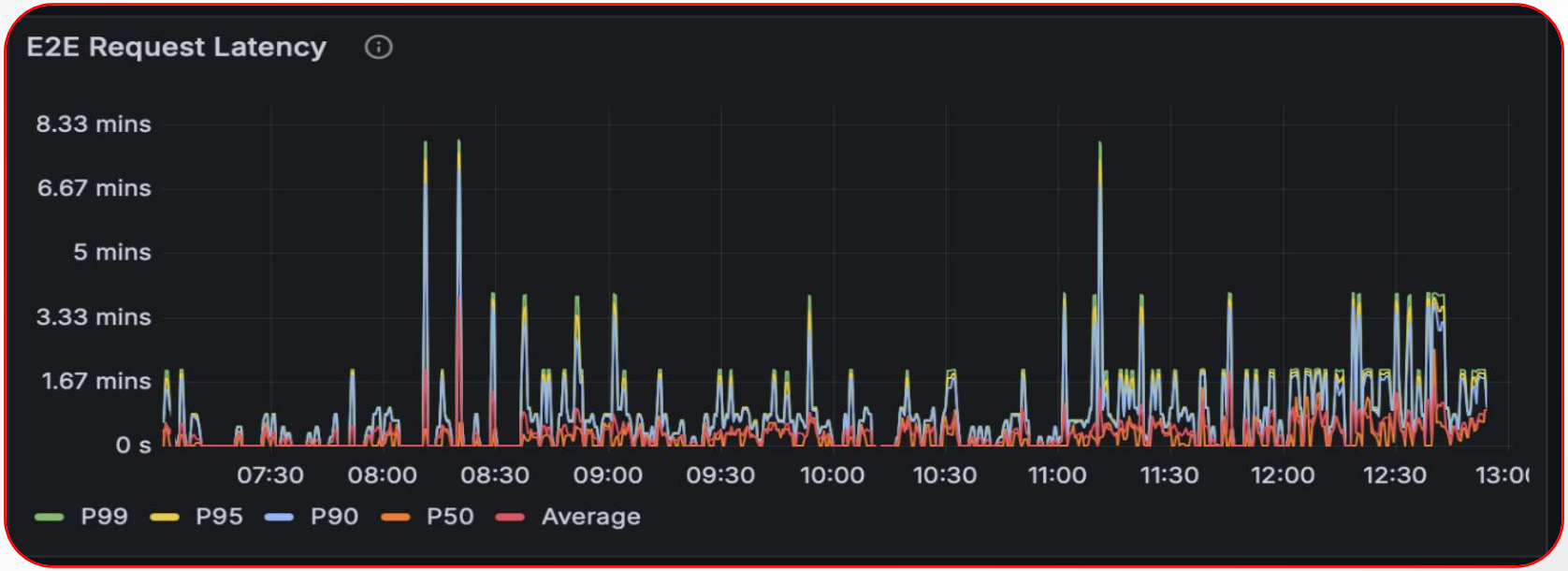
Интерпретация: показывает загрузку KV-кэша на GPU. Рост к ~100% означает близкую фрагментацию/эвикты и потенциальные скачки задержек.

Критическая ситуация



Интерпретация: TTFT взлетает до минут ровно в тот период, когда GPU-KV-кэш забит почти на 100%.

E2E Request Latency



Интерпретация: если E2E растёт, а TPOT нормальный — проблема в **TTFT/очереди** (длинные промпты, загрузка шедулера, внешние I/O). Если растут **и** TPOT — упёрлись в GPU/кэш/батчинг.

Бизнес метрики. DAU

DAU — верхушка айсберга

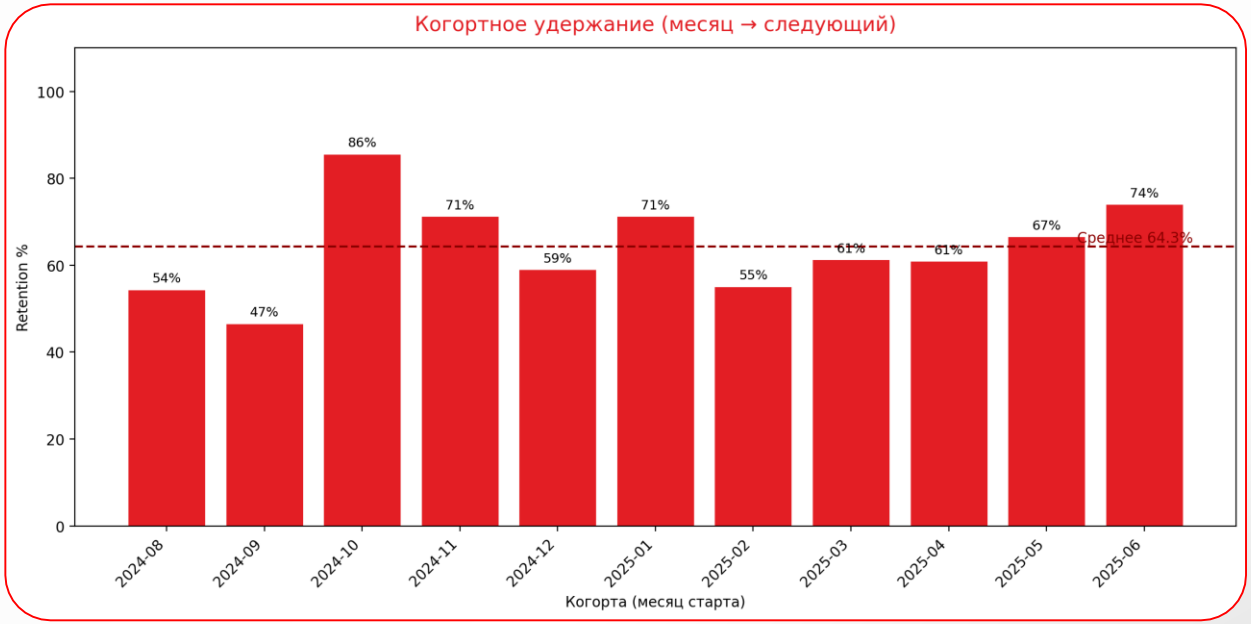
1 - релиз новых моделей + железо
2 - релиз нового UI
3 - переезд на новый модели + железо

На что еще можно смотреть:

- DAU / MAU: Показывает, насколько регулярно пользователи возвращаются. Если он высокий (> 0.5), значит, сервис входит в привычку
- Активность по отделам / департаментам

Бизнес метрики. Retention

Retention m2m — это процент сотрудников, которые продолжают использовать сервис месяц к месяцу для решения своих рабочих задач.
По сути одна из главных метрик для сервиса «общего назначения».



Бизнес метрики. Специфика

Категория 1: Повышение продуктивности и эффективности

- ❑ Скорость выполнения задач
- ❑ First response time
- ❑ Time-to-market
- ❑ Скорость онбординга новых сотрудников

Категория 2: Экономический импакт

- ❑ Public API cost vs Self-hosted
- ❑ LLM Feature / Product profit

Категория 3. Удовлетворенность пользователей

- ❑ Обратная связь от пользователей

Спасибо за внимание!

Вебинар: “Остается ли в тайне ваш диалог с LLM? Как использовать большие языковые модели безопасно!”

Вебинар: “ Не словом, а кодом. Как организовать защиту LLM без потерь производительности?”

Telegram канал ML-команды PT