SmartData 2026

LLM Ops: оптимизация инференса и ML-serving

в реальном production-кластере · что бы я делал по-другому, если бы знал, что знаю сейчас

Дмитрий Ибрагимов · Лемана Про · SmartData 2026

963 пользователя · ~252 000 запросов/день · 96 млрд токенов/мес · GPUStack

ОБО МНЕ

Дмитрий Ибрагимов

Директор платформы данных · Лемана Про (бывш. Леруа Мерлен)

· Строю и сопровождаю корпоративную AI-платформу
· 963 пользователя, рост +20% в месяц
· В докладе — 4 мои роли в платформе и то, что я сделал бы по-другому

Доклад — не про выбор GPU, инференса или модели, а про построение платформы и 90% работы вокруг неё.

(Data → Embedding → RAG ← Model → Harness) × FinOps = Observability

Платформа данных даёт фундамент, а языковые модели реализуют новый интерфейс доступа к данным. И это новый вызов для нас, как это сделать убедившись, что: модель запущена, каждый токен посчитан, каждая метрика наблюдаема.

Observability

трейсы, метрики, алерты

× FinOps — учёт и тарифы

Data

DataLake, DLH, DWH, DQ/DG — данные качественные и у каждой цифры есть владелец

Embedding

Извлечение знаний — чанкование и загрузка в VDB

RAG

200+ баз знаний
Управление знаниями

Model

GPUStack, 4 бэкенда

Harness

gateway, квоты и лимиты

SmartData 2026

Все сидят на персональных подписках. Это проблема?

НЕУПРАВЛЯЕМОСТЬ

· ИИ используется вслепую — нет наблюдаемости и мониторинга качества ответов
· Токены не видны по продуктам и пользователям
· Нет стандартизированного сервиса для разработки AI-продуктов

БЕЗОПАСНОСТЬ И КОМПЛАЕНС

· Нет единого управления доступом и политиками безопасности
· Нет фильтрации входа и выхода LLM для чувствительных данных
· Ключи и доступ к LLM не централизованы

ФРАГМЕНТАЦИЯ И ПОТЕРЯ ГИБКОСТИ

· Каждая команда выбирает свои LLM, инструменты, подходы к работе с AI
· Отсутствие единых стандартов и инструментов для работы с AI
· Архитектура не готова к оркестрации AI-агентов
· Нет механизма обкатки новых LLM на подготовленных проверенных данных

РЕГУЛЯТОРНЫЙ КОНТЕКСТ (РФ)

152-ФЗ: утечка ПДн — до 15 млн ₽ первично, 1–3% выручки повторно.
Без фильтрации и аудита соблюдение 152-ФЗ и 243-ФЗ гарантировать нельзя.
243-ФЗ — рамочный закон о развитии ИИ: задаёт рамку, а не штрафы.

Четыре пути компаний к LLM

Подход	Плюсы	Минусы
Персональные подписки	Нулевой порог входа	Нет учёта и безопасности, запрещать бессмысленно
API-only (внешние модели)	Топовые модели без железа	Для РФ неприменимо: нужен роутер, оплата из РФ проблемна
Cloud only (облачные LLM)	Yandex Cloud, Cloud.ru, Selectel — инфраструктура на вендоре	Данные у провайдера, зависимость от его цен и лимитов
Open-source on-prem	Данные внутри, предсказуемая цена (capex)	Нужны инфраструктура и ops-команда

Запретить подписки нельзя — можно дать платформу, которая честнее по цене и безопаснее по данным.

Мы выбрали строить AI платформу.

Единая инфраструктура и набор сервисов для управляемого использования ИИ во всех продуктах компании.
Без единого шлюза — теневые внедрения и риски DLP. Внешние модели — только через наш AI Gateway и guardrails.

INTEGRATION

AI Gateway, MCP: единое взаимодействие с моделями, системами и инструментами

SECURITY & OBSERVABILITY

Политики, безопасность, аудит, мониторинг и управление затратами

SDLC

Среды и инструменты для разработки, тестирования и безопасного исполнения агентов

MODELS & ARTIFACTS

Жизненный цикл моделей, промптов и агентов: версии, метаданные, статусы

DATA & KNOWLEDGE

Сбор, хранение и подготовка данных и знаний для AI-агентов

INFRASTRUCTURE & INFERENCE

Вычислительная мощность и оркестрация ресурсов под все AI-нагрузки

Платформа, о которой пойдёт речь

963

активных пользователя,
рост +20%/мес

~252 000

запросов в день

96 млрд

токенов в месяц

12

моделей: LLM, STT,
TTS, Embedding, Vision

- RAG на Qdrant — 200+ баз знаний; PR-Agent — 500+ репозиториев
- Kubernetes в Yandex Cloud, узлы 8×H200 (141 ГБ), 4 инференс-бэкенда
- KeyCloak SSO, RBAC, 1000+ API-ключей — каждая цифра на этом слайде из боевой системы

Платформа целиком: пять слоёв

ПРОДУКТЫ

Digital Products → AI Self-service Portal

единый интерфейс платформы для сотрудников

▼

ВХОД

AI Gateway · Trust & Guardrails · RAG & Knowledge

маршрутизация и квоты · проверки промптов и ответов · знания и поиск

▼

СЕРВИСЫ

Model Registry · AI Test & Evaluation

Agent Studio + Agent Runtime · Feature Store

▼

ИНФЕРЕНС

Model Serving — KServe, vLLM

GPU Orchestration & Scheduling — Volcano, Liqo

▼

ЖЕЛЕЗО

GPU Operator — подключение GPU к Kubernetes

External LLM — внешние поставщики моделей

СКВОЗНЫЕ

AI FinOps

стоимость и распределение затрат

Data Observability

мониторинг и оповещения

Enterprise Context

контекст компании для AI

Упрощённая карта платформы; API/MCP Gateway, TechGate и песочницы — уровень деталей.

Первым делом — граница владения

АРЕНДОВАТЬ

· модели как сервис

· внешние API для burst

· новые модели на пробу

ДОРАБАТЫВАТЬ

· роутинг и квоты (LiteLLM)

· портал и интеграции

· RAG поверх моделей

ВЛАДЕТЬ

· GPU и нарезка

· инференс-бэкенды

· мониторинг и экономика

Ответ «где наша зона» дешевле любых архитектурных споров.

Владение — это 4 роли

1

GPUOps

оптимизация инфры

2

LLMOps

оптимизация моделей

3

FinOps

оптимизация потребления

4

SRE

повышение надёжности

В каждой роли — наша история провала и то, что я сделал бы по-другому.

01

GPUOps

Провал роли: Fabric Manager не настроили — OOM в RAM, полдня без моделей.

LLM не нужна целая карта.

- 27B-модели помещаются в 105 ГБ — это 0.75 карты H200
- Embedding и Whisper живут на слайсах в 10–20 ГБ
- Отдавать целую GPU «команде на попробовать» — дорого и расточительно
- Поэтому GPU режем — и это самое недооценённое умение владельца платформы

Три способа нарезки GPU

TIME-SLICING

GPU делится по времени. Просто, но задержка вывода (TTFT, TPOT, Latency) плавает

MIG

Multi-Instance GPU: VRAM режется на аппаратные части — модели живут параллельно и изолированно

ДИНАМИЧЕСКИЙ

Project HAMi — пока наблюдаем

GPU Operator делает это автоматически

AUTOMATES SETUP

· Driver container — сам ставит драйвер на ноду
· Container toolkit и device plugin без ручных действий
· Ресурс nvidia.com/gpu в Kubernetes

PARTITIONING

· MIG Manager и time-slicing из коробки
· Физический GPU режется на слайсы для нескольких LLM-реплик
· DRA: MIG-устройства через ResourceClaim (v26.7)

TELEMETRY

· DCGM Exporter — утилизация, память, health
· Метрики в Prometheus + Grafana
· Связка с HPA + cluster autoscaler

Ручная настройка GPU — цепочка жёстких зависимостей: CUDA ↔ драйвер ↔ ядро ОС ↔ архитектура GPU.
Реальная боль: драйвер 520 «отваливался» после обновления ядра Ubuntu — с Operator-ом драйвер ставится декларативно.

Как модели лежат на узле 8×H200

Модель	Назначение	Карточек	VRAM
glm-5.3-flash (флагман)	сложная логика, архитектура, vision-анализ	4×H200	511 ГБ
deepseek-v4-flash	массовые агенты: код, ревью, большой контекст	2×H200	247 ГБ
qwen3.8-27b / qwen3.6-27b	UI-генерация и массовый параллельный инференс	по 0.75×H200	105 ГБ
embedding-модель	вектора для RAG-поиска	слайс	20 ГБ
Whisper (STT)	распознавание речи	слайс	~10 ГБ

Один узел несёт флагман, две средние модели, embedding и речь. Так выглядит «плотный» прод.

Провал: Fabric Manager

ЧТО СЛУЧИЛОСЬ

На серверном железе не настроили Fabric Manager: OOM в RAM, инстансы моделей падали одна за другой, платформа лежала почти полдня.

ЧТО БЫ Я ДЕЛАЛ ПО-ДРУГОМУ

Железо до моделей: NVLink/Fabric-конфигурация — часть чек-листа деплоя, а не «мелочь для сисадминов». Проверять до загрузки первого весового файла.

Нарезал бы GPU с первого дня, управлял бы через GPU Operator

Недоиспользованная целая карта дороже инженерного времени на настройку слайсинга.

- Начните с time-slicing и MIG: 27B-модель в 105 ГБ — это 0.75 карты, а не целая H200
- Плотное размещение — тот же GPU-бюджет несёт вдвое больше моделей
- Динамический слайсинг подключите, когда статичной нарезки перестанет хватать
- GPU Operator — сразу: после обычного apt install драйвер на хосте слетел, и весь кластер встал

02

LLMOps

Провал роли: начали с сырой версии GPUStack — и заплатили стабильностью.

Свежий vLLM каждые месяц-два

- Новые модели (Qwen 3.6, Minimax) требуют свежих версий инференс-бэкендов
- Day-0 поддержка в GPUStack: модель подключается в день релиза
- Это не разовая настройка, а постоянный конвейер обновлений

Каталог: 12 моделей в production — LLM (glm-5.3-flash, deepseek-v4-flash, qwen3.8/3.6-27b), embedding qwen3-vl-embedding-8b, Whisper STT, TTS, Vision. Каждая — свой бэкенд и своя версия.

Четыре бэкенда за одним фасадом

- **vLLM** — высокий throughput и эффективность памяти — дефолт для LLM
- **SGLang** — LLM и vision-language модели; image-модели (Qwen-Image, Flux)
- **MindIE** — инференс на Huawei Ascend NPU — не-NVIDIA альтернатива
- **VoxBox** — TTS/STT: faster-whisper, qwen3-asr/tts

В проде — два из четырёх: vLLM для LLM и embedding, VoxBox для STT-моделей

Единый OpenAI-совместимый API: Chat Completions, Embeddings, Rerank, Audio — интеграция без переписывания кода

Compatibility check экономит ночи

~95%

ошибок деплоя отсекается до запуска

- Проверка архитектуры модели, ОС, VRAM и путей до первого запуска
- Подбор бэкенда и версии под модель автоматом
- Ночь не уходит на падающий инстанс с неверно оценённой памятью

Всё тестируем на test-окружении перед деплоем в прод

Параллелизм и планировщик

- Tensor Parallelism: TP2 даёт до +26.8% TPS — слои модели по картам
- Expert Parallelism для MoE-моделей — параметры подобрали замерами
- Шедулер GPUStack: единая GPU → несколько GPU (TP) → Ray → CPU-фолбэк
- Размещение: binpack — плотность для dev, spread — отказоустойчивость для prod

Оптимизации, которые мы реально используем

Техника	Эффект	Где применяем
Speculative decoding	2.4–2.9 токена за шаг	MTP×3 у glm и qwen, dspark×5 у deepseek
KV-cache в FP8, block 256	1M контекста на 2×H200	deepseek-v4-flash
Chunked prefill 8–16k	не блокирует декодинг	агентные промпты 16k (glm 8k, qwen 16k)
Distributed cache (PD)	префилл и декод раздельно	shared KV-cache между ними; на 27B — EFAULT (ниже)
Шедулинг по замерам	max-num-seqs 64 → пик c64	glm-5.3-flash

ИНСАЙТЫ ИЗ СТРЕСС-ТЕСТОВ

Acceptance спекуляции меряем на своём трафике: qwen3.6 — 63%, glm — 47–54%, qwen3.8 — 45%, deepseek — 29%; черновики №4–5 dspark принимаются в 16% и 12% случаев — сокращаем до трёх.

Дефолтный reasoning_effort xhigh у qwen3.8: в 2.5–3 раза больше reasoning-токенов и 6.3% doom loops при качестве ниже, чем у low/medium. Снижение дефолта срезало нагрузку на GPU почти на 20%.

Пробуем PD на 27B: offload KV-cache 128 GiB на CPU → madvise(MADV_POPULATE_WRITE) пре-фолтит mmap в /dev/shm пода → EFAULT (SIGBUS), не «bad address». Shared KVCache в GPUStack — только 3 недели.

Замеры: vllm bench serve, concurrency 1–64, профиль 2k/256 и agentic 16k/512, seed 42, 0 ошибок

Красивый бенчмарк не доказывает результат

ЛАБОРАТОРИЯ

Extended KV Cache (RAM):
+355% input TPS
−58.8% TTFT

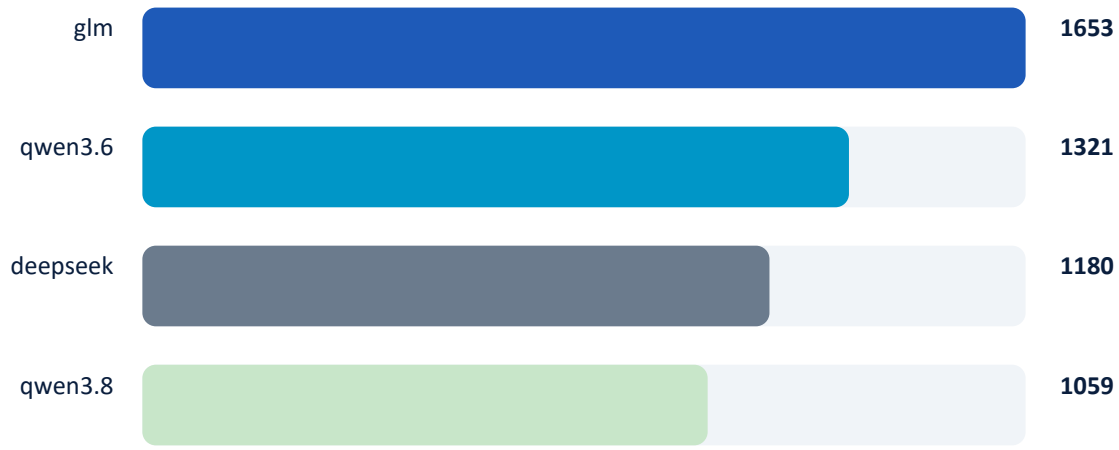
Цифры из Inference Performance Lab — впечатляют.

НАШ PROD

В проде НЕ включаем.
Prefix cache и так покрывает потребности:
hit rate 86–98%,
99% prompt-токенов читаются из кэша.

Решение не «передовое/отсталое», а измеренное на своём трафике.

Пик ≠ стабильность



Пиковый output TPS, direct: glm-5.3-flash · qwen3.6-27b · deepseek-v4-flash · qwen3.8-27b

- Через шлюз glm теряет 12% (1653 → 1459) — у прокси есть цена
- deepseek на c=64: TPS деградирует, TTFT p99 7.6c
- qwen3.8: немонотонная кривая, TTFT p99 15.5c, KV cache 93.2%
- Agentic-профиль glm: TTFT p99 до 40c — смотрите p99, не среднее

Методика: промпт ~2k токенов, ответ 256, до 64 клиентов, 1920 запросов, 0 ошибок; контроль по пределу H200 4.8 ТБ/с.

Свои бенчмарки — с первого дня

26,7% команд вообще не тестируют модель перед продом. Публичные лидерборды — фильтр шорт-листа, а не релизный гейт.

- Агентные бенчмарки (deepSWE, Agents' Last Exam, MERA) не знают ваших данных, инструментов и SLA
- Заявлениям вендоров не верим — цифры снимаем на своём трафике
- Golden dataset из логов: smoke 20–50 трейсов, основной сет — сотни; синтетика — лишь редкие классы
- Раскладываем по типам: частые сценарии, редкие критичные, граничные случаи, запросы-отказы
- Оффлайн-эвалы перед релизом + онлайн-эвалы на проде — против дрейфа
- LLM as Judge на трейсы и логи — с чёткими критериями, сверкой с людьми и зафиксированной версией судьи

research.llmarena.ru (LLM Arena, 2025) · langchain.com/resources/llm-evaluation-benchmarks

Провал: мы начали с сырой версии

ЧТО СЛУЧИЛОСЬ

GPUStack 0.7.1 — сырой продукт: нестыковки шедулера, неожиданные рестарты, часовые отладки.
Мы начинали на 0.x: 1.x у проекта не было вовсе, после 0.x сразу вышли 2.x — обновление стоило нервов.

ЧТО БЫ Я ДЕЛАЛ ПО-ДРУГОМУ

Обновлял бы инференс-стек регулярно, а не «проектом» раз в полгода: выкат на тестовый контур, затем в окно обслуживания — с проверенным откатом. Сырую 0.x в прод больше не брали бы.

03

FinOps

Провал роли: полгода не считали токены вообще — не могли ответить «сколько это стоит».

Цена токена — от capacity, не от загрузки

Узел 8×H200 = 37.5 млн ₽ → амортизация 5 лет = 625 тыс. ₽/мес ≈ 20.5 тыс. ₽/сутки.
Добавление нового железа не удорожает токен — экономика предсказуема.

- Ёмкость узла: 3.7–4.1 млрд токенов/сутки
- Платформа (4 узла): 15–16 млрд/сутки
- С чтениями из промпт-кэша (×2.7–4.1): 40–70 млрд тарифицируемых/сутки

Тарифы внутри платформы

Модель	Вход	Выход
glm-5.3-flash	4.47	35.92
deepseek-v4-flash	3.14	25.18
qwen3.8-27b	1.28	10.52
qwen3.6-27b	1.05	8.45

Выход ~в 8 раз дороже входа.

Чтение из промпт-кэша — 15–30% цены полного входа.

SmartData 2026

313 тысяч против 934 тысяч

313 232 ₽

вся платформа за месяц
≈ 7 600 ₽/рабочий день

~934 тыс. ₽

те же объёмы по стрит-ценам
внешних API

≈ в 3 раза

дешевле API

- Разбивка: glm 178 878 ₽ · deepseek 73 306 ₽ · qwen3.6 54 059 ₽ · qwen3.8 5 173 ₽
- Загруженность 2–4% от потолка — запас под рост +20% пользователей в месяц
- Полная приватность данных входит в эту цену

SmartData 2026

Один пользователь — 15.8% потребления

15.8%

потребления — один пользователь
(~4.58 млрд токенов, ~140 тыс. ₽)

117

пользователей потратили
≥1 тыс. ₽ за месяц

~200

пользователей — меньше 100 ₽
(длинный хвост)

- С топ-1 разговаривают про квоты и сценарии — это нормально
- Для каждого считается экономия: сколько токенов НЕ улетело в платный API
- Chargeback по командам — следующий шаг, расчёт готов

Считал бы токены с первого дня

Учёт и тарифы — не «фича на потом», а фундамент: без цифр платформа защищается эмоциями.

- Per-user учёт с первого дня: кто, сколько, какой моделью
- Тарифы от capacity — экономика предсказуема, добавление железа не удорожает токен
- Квоты и rate limiting (LiteLLM) — в первый месяц, а не «когда-нибудь»

04

SRE

Провал роли: дашборды зелёные, а пользователи уже час пишут «не отвечает».

Мониторинг уже в комплекте

LLM RUNTIME	requests running/waiting, prefix_cache_hit_rate, kv_cache_usage, TTFT, e2e
WORKER NODE	GPU utilization, temperature, VRAM, CPU, memory, filesystem
MODEL / CLUSTER	model_desired_instances, model_running_instances, model_instance_status
LLM / TOOL / AGENT CALLS ДОБАВЛЯЕМ САМИ	Latency / Rate / Errors — по каждому типу вызовов; TTFT · tokens/sec · agent error rate

Метрики GPUStack меняются на лету через REST: GET/POST /v2/metrics/config; свой слой — в Prometheus + Grafana.

Инфраструктура зелёная, а продукт молчит

- Опсовых алертов по качеству нет: только инфраструктурные (K8s, GPU, железо)
- Никто не алертит на деградацию TTFT p99 или рост ошибок агентов
- Ночью и в выходные использование падает в ноль — это норма, но её надо видеть на графике
- Если токены «горят» ночью — повод поговорить с командой (или завести алерт на выгорание)

Платформа лечится сама

- Auto-restart инстансов с exponential backoff (максимум 5 минут)
- Status scorer: при scale-down первыми уходят unhealthy-реплики
- Offload layer scorer: аккуратно удаляет частично offloaded GGUF-инстансы
- Compatibility check до деплоя — большая часть аварий просто не случается

Метрика зрелости: платформа vs подписки

Доля работы через платформу против персональных подписок — в динамике.

- Когда доля растёт — shadow AI умирает сам, без запретов и приказов
- У нас: 963 активных пользователя, рост +20% в месяц
- За день — ~252 000 запросов, и каждый посчитан

Опсы и SLA с первого дня

Алерт на TTFT p99 и ошибкам агентов — в первый месяц, а не «когда-нибудь».

- Real-time и batch — разные инстансы: p99 ниже в 2.3 раза
- SLA-таргеты фиксируйте письменно: Gateway 99.5%/P99 15ms, Serving 99%/P99 700ms
- Cached tokens на дашборде — первый индикатор деградации перед падением качества

05

Выводы

Ключевые выводы из четырёх ролей: что бы я делал по-другому, если бы знал, что знаю сейчас.

Ключевые выводы

01 Владейте сопровождением

02 Считайте токены с первого дня

03 Пик ≠ стабильность: смотрите TTFT p99

04 Нарезайте GPU до того, как прижмёт

05 Мониторьте продукт, а не только GPU

Спасибо за внимание!

github.com/gpustack/gpustack	GPU cluster manager designed for efficient AI model deployment. It configures and orchestrates inference engines — vLLM, SGLang, or your own — to optimize performance across GPU clusters.
github.com/project-hami/hami	разделяемые GPU в K8s: несколько команд на одной карте
docs.litellm.ai	LLM Gateway: роутинг, лимиты, фолбэки

Вопросы: Telegram @diarworld · SmartData 2026 · smartdataconf.ru

SmartData 2026

Приложение: как это устроено сегодня

ПОЛЬЗОВАТЕЛИ И АГЕНТЫ

Open WebUI — чат · внутренние data-продукты · агенты: OpenCode, ZooCode, Hermes, OpenClaw, Kilo Code, MiMo

API GATEWAY (LemanaPro) — единый вход для интерфейсов, продуктов и агентов

LiteLLM — единая точка доступа к моделям: внутренние vLLM и внешние API · кэширование, балансировка, rate limits

GUARDRAILS

Presidio — фильтрация запросов и ответов, защита персональных данных

OBSERVABILITY

Langfuse — трейсы LLM-вызовов · OpenLLMetry — телеметрия

RAG & MEMORY

RAGFlow / LangChain · Mem0 — память · FireCrawl · LanceDB / Qdrant / Weaviate / PGVector

ИНФЕРЕНС

GPUStack — нарезка GPU, 4 бэкенда → vLLM — локальные модели

DATA LAYER

S3 Data Lake · Airflow + Unstructured.io / MarkItDown — ETL в векторную БД · Argilla / Label Studio — разметка

ВНЕШНИЕ API

GPT · Claude · Yandex · Cloud.ru

Сверху вниз: пользователи и агенты → Gateway → ядро платформы → инференс и данные.

Приложение: куда растём

AI Self-service Portal

единый интерфейс, уже работает

Model Serving & Inference

GPUStack, 4 бэкенда — уже работает

AI Gateway (LiteLLM)

роутинг внутренних и внешних LLM, квоты, фолбэк

GPU Operator / Scheduling

федерация кластеров, квоты и приоритеты

Model Registry

модели, версии, жизненный цикл

RAG & Knowledge Management

Qdrant, 50 баз — уже работает

AI Trust & Guardrails

фильтрация промптов/ответов, PII-аудит

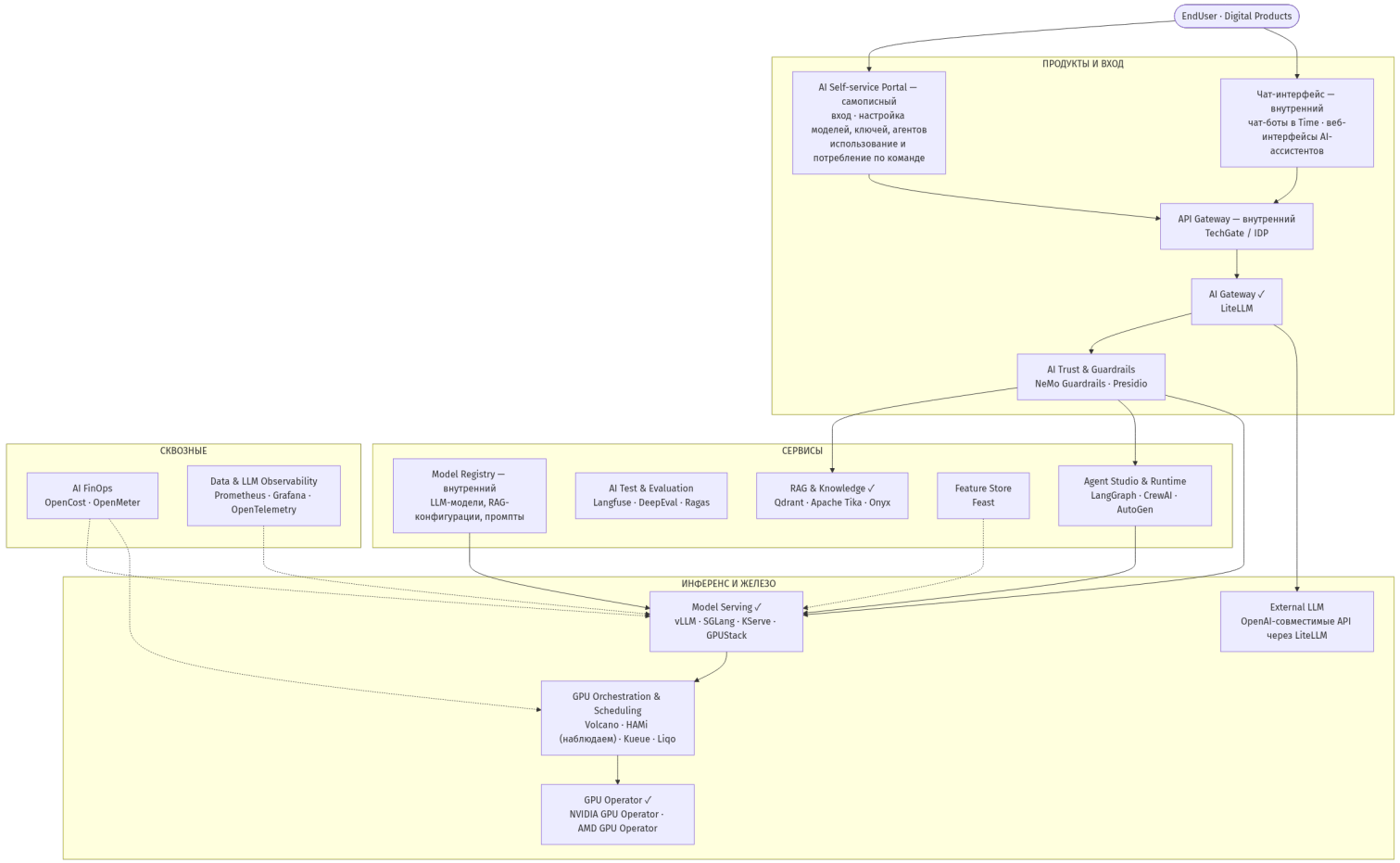
Agent Profile Registry + Runtime

~100 агентных сервисов

Цель: 100% трафика через AI Gateway · 100% моделей под единым фасадом · 100% запросов через фильтрацию

SmartData 2026

Приложение: целевая архитектура



✓ — уже используется в платформе · «внутренний» — собственная разработка

Приложение: open-source компоненты

Функция	Продукты / OSS	Описание
AI Self-service Portal	— (самописный), прямого аналога нет; частично: Open WebUI + LiteLLM	вход, настройка моделей, ключей и агентов; потребление по команде
Чат-интерфейс	**Open WebUI ✓ · Time-боты (внутренний)**	чат-фасад для сотрудников, SSO
API Gateway	внутренний: TechGate / IDP · OSS: Kong	авторизация, rate limit на входе
AI Gateway	**LiteLLM ✓**	роутинг моделей, квоты, фолбэки
AI Trust & Guardrails	NeMo Guardrails · Presidio	фильтрация промптов и ответов, PII
Model Registry	внутренний (создаётся) · OSS: MLflow	реестр LLM-моделей, RAG-конфигураций, промптов
AI Test & Evaluation	Langfuse · DeepEval · Ragas	трейсинг, оффлайн/онлайн-эвалы, golden dataset
Agent Studio / Runtime	LangGraph · CrewAI · AutoGen	оркестрация агентов, инструменты по MCP
Feature Store	Feast	фичи для ML-сценариев
RAG & Knowledge	**Qdrant ✓ · Apache Tika · Onyx · RAGFlow**	гибридный поиск dense+sparse, ingestion документов
Model Serving	**vLLM ✓ · SGLang · KServe · GPUStack ✓**	инференс-движки + фасад управления
GPU Orchestration & Scheduling	Volcano · HAMi · Kueue · Liqo	очереди, шаринг GPU, мультикластер
GPU Operator	**NVIDIA GPU Operator ✓ · AMD GPU Operator**	драйверы и device plugins в K8s
External LLM	через LiteLLM (OpenAI-совместимые API)	API внешних провайдеров
AI FinOps	OpenCost · OpenMeter	стоимость по командам и токенам
Data & LLM Observability	Prometheus · Grafana · OpenTelemetry	метрики, дашборды, трейсинг