DevOops

Корпоративный GPT

Как внедрить так, чтобы все остались довольны



MAGNIT TECH



Унтура Михаил

Главный MLOps-инженер
MAGNIT TECH

MagnitGPT

Корпоративный ассистент для всего Магнита



20.6 K

пользователей

MagnitGPT

Корпоративный ассистент для всего Магнита



20.6 K

пользователей

~2 K

DAU

MagnitGPT

Корпоративный ассистент для всего Магнита



20.6 K

пользователей

~2 K

DAU

263 K

чатов

MagnitGPT

Корпоративный ассистент для всего Магнита



20.6 K

пользователей

~2 K

DAU

263 K

чатов

~2 B

токенов ежедневно

Зачем усложнять?



Чат

Ну конечно же Open WebUI на ВМ
в Docker!

Зачем усложнять?



Чат

Ну конечно же Open WebUI на ВМ в Docker!

Инференс

Ollama подойдет, одна команда и модель запущена!

Зачем усложнять?



Чат

Ну конечно же Open WebUI на ВМ в Docker!

Инференс

Ollama подойдет, одна команда и модель запущена!



Обвязка

Бизнес просит добавить новые фичи и еще 20 интеграций...

Зачем усложнять?



Чат

Ну конечно же Open WebUI на ВМ в Docker!

Инференс

Ollama подойдет, одна команда и модель запущена!



Обвязка

Бизнес просит добавить новые фичи и еще 20 интеграций...

Стабильность

Ой-ой, к нам пришло 20 тысяч пользователей...

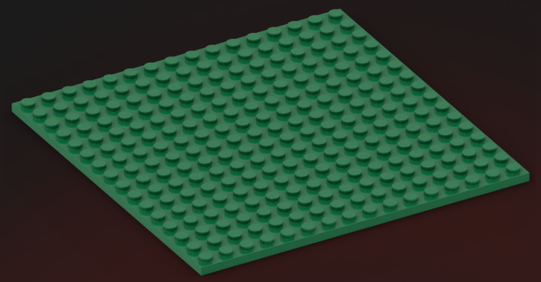
На чем мы живем?

Вся инфраструктура в облаке

Основная нагрузка в k8s, из облака тянем S3, Managed Posgtres и Redis

88	нод
1543	пода
16	GPU H200
7	GPU H100
3	GPU A100
1	GPU T4

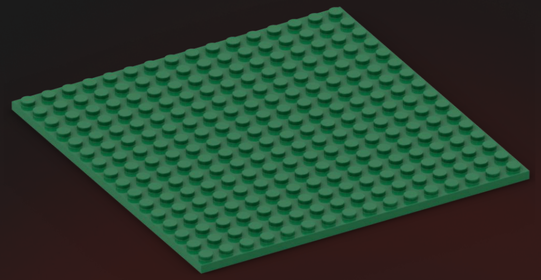
Собираем конструктор



Kubernetes

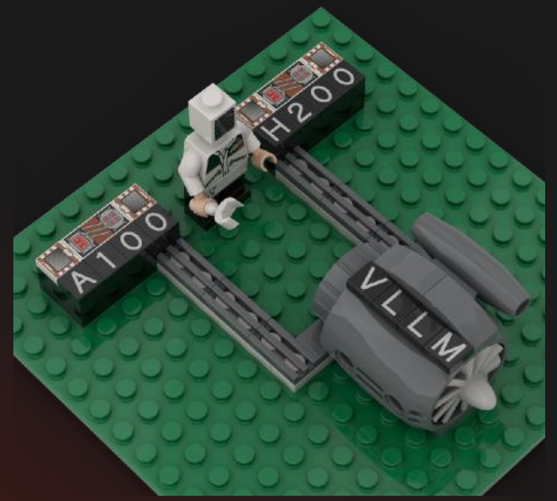
Высокие нагрузки, масштабируемость, удобство и экспертиза

Собираем конструктор



Kubernetes

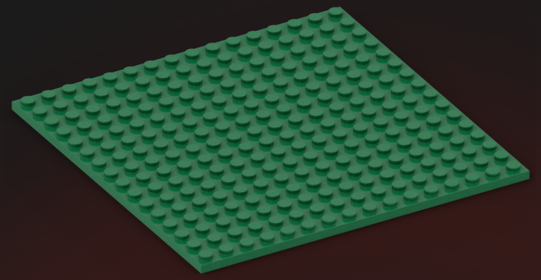
Высокие нагрузки, масштабируемость, удобство и экспертиза



GPU Operator

Управление и проброс GPU, нарезка на MIGи

Собираем конструктор



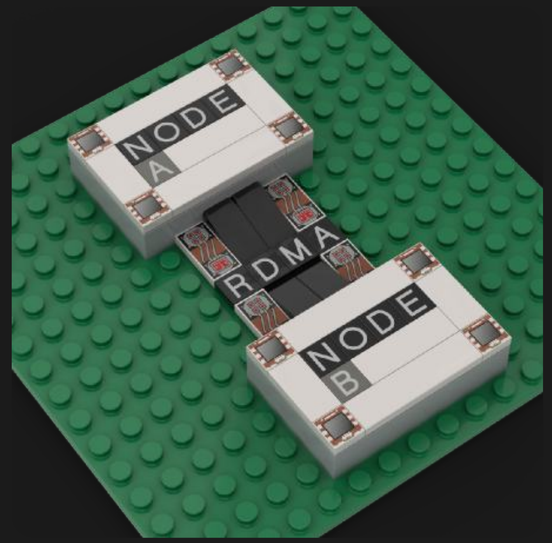
Kubernetes

Высокие нагрузки, масштабируемость, удобство и экспертиза



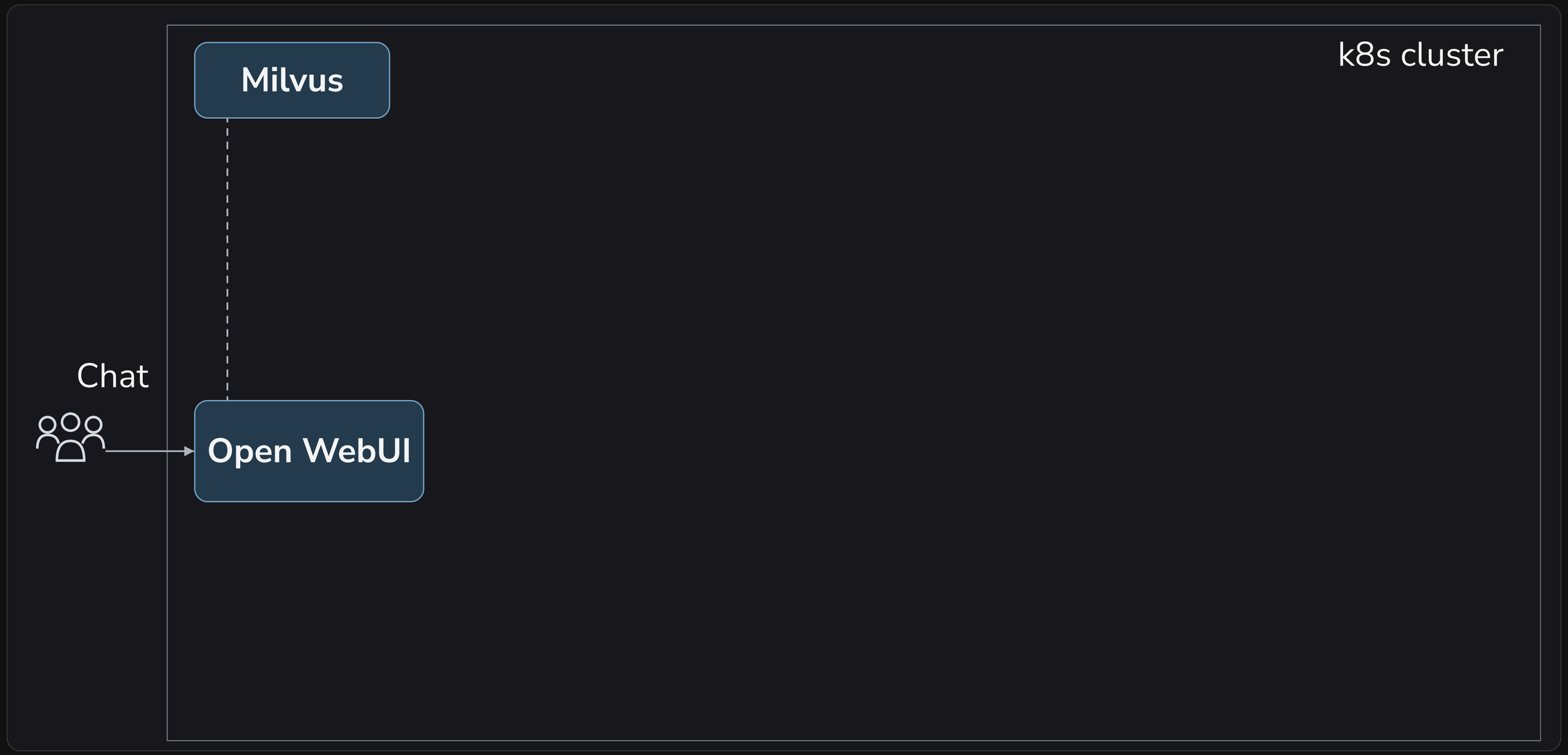
GPU Operator

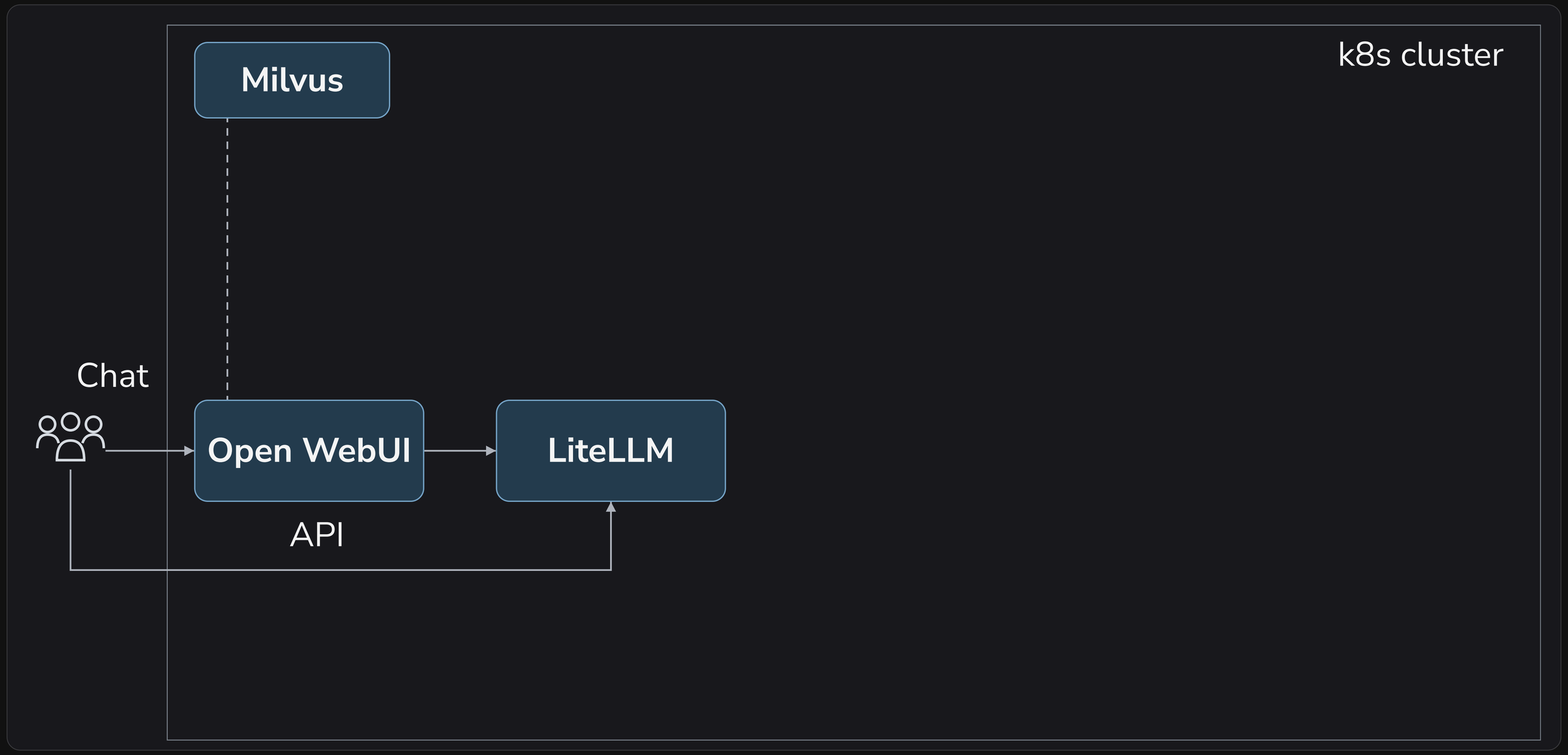
Управление и проброс GPU, нарезка на MIGи



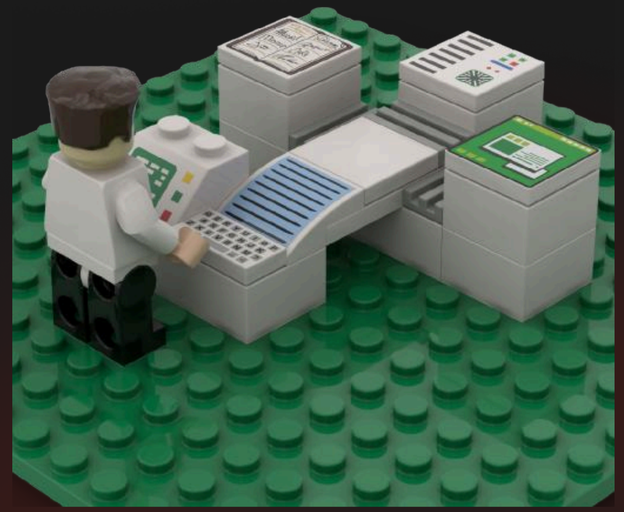
Network Operator

Управление сетевыми сущностями и драйверами, проброс rdma_cm


k8s cluster
Milvus
Chat
Open WebUI


k8s cluster
Milvus
Chat
Open WebUI
LiteLLM
API

Собираем конструктор



OpenWebUI

Частые обновления, много интеграций, модульность и привычный интерфейс

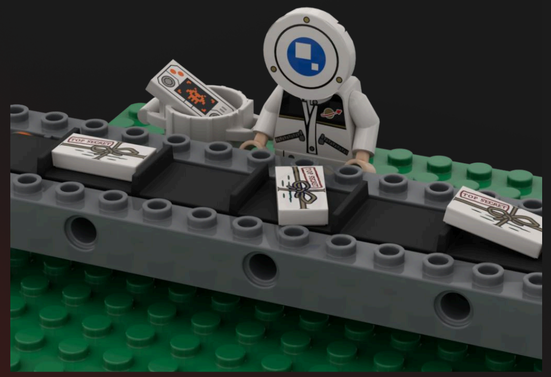
MAGNIT TECH

Собираем конструктор



OpenWebUI

Частые обновления, много интеграций, модульность и привычный интерфейс



LiteLLM

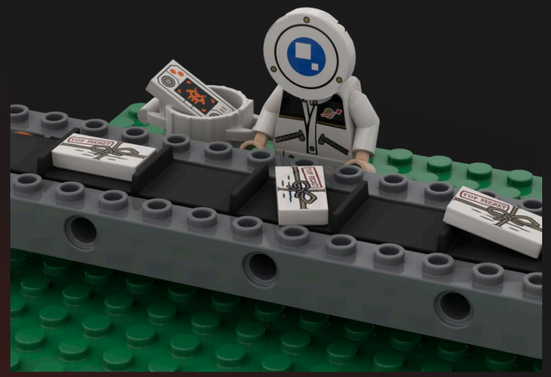
Выбор провайдеров, расширяемость, удобная ролевая модель

Собираем конструктор



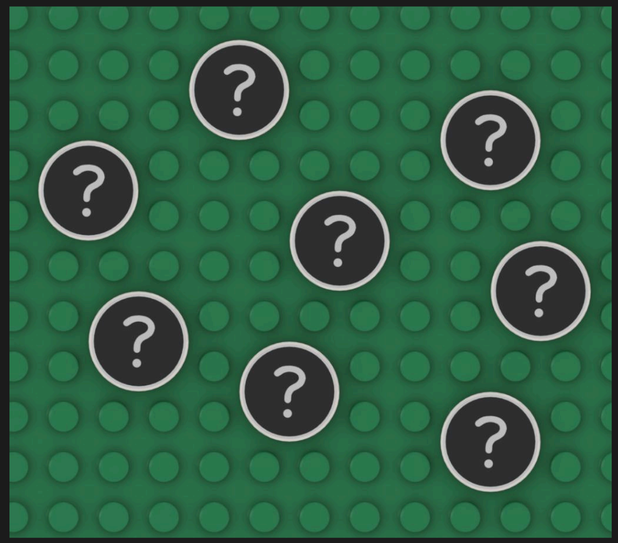
OpenWebUI

Частые обновления, много интеграций, модульность и привычный интерфейс



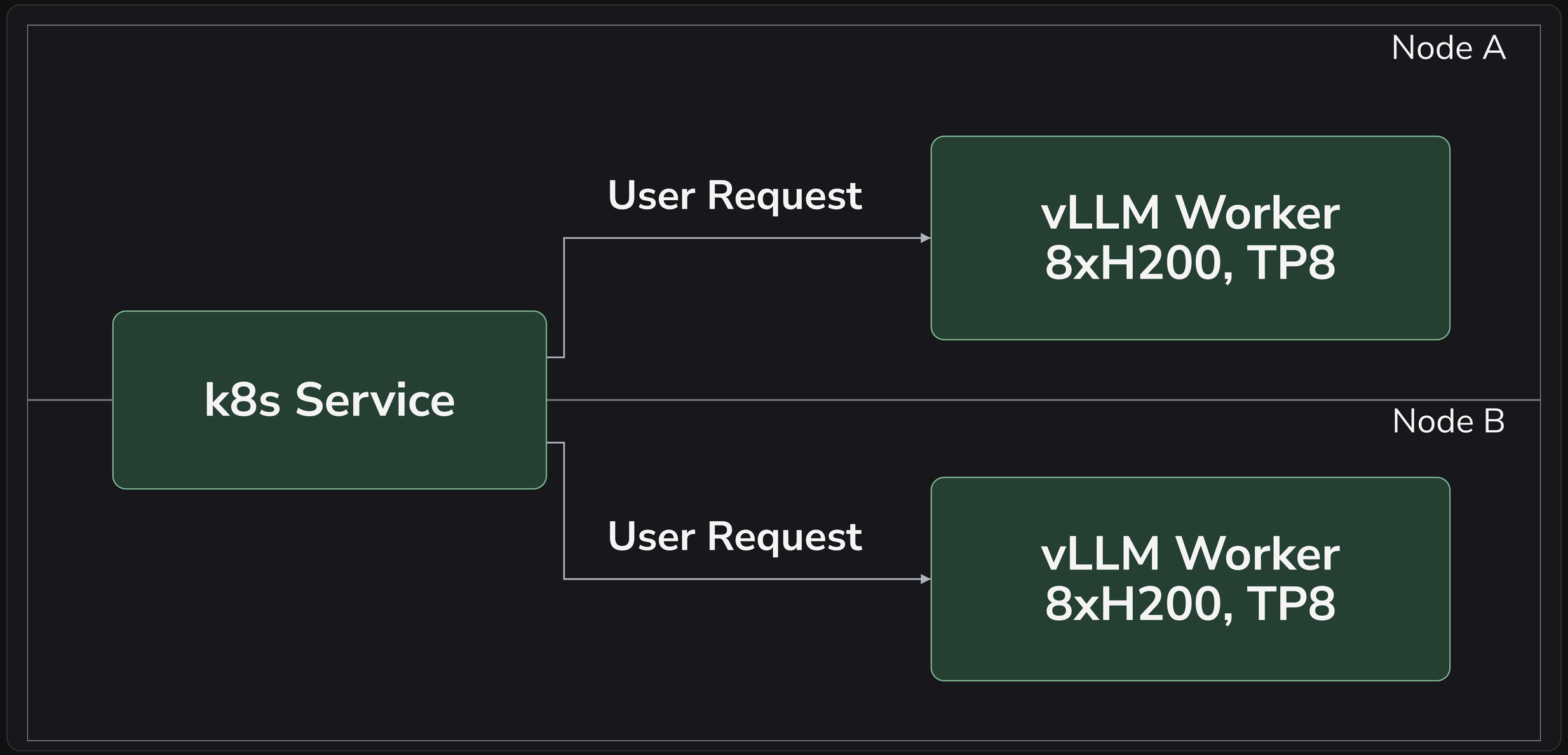
LiteLLM

Выбор провайдеров, расширяемость, удобная ролевая модель



Инференс

А вот тут все уже не так однозначно


Node A
User Request
vLLM Worker
8xH200, TP8
k8s Service
Node B
User Request
vLLM Worker
8xH200, TP8

vLLM

vllm-project/vllm

Зрелость

Стандарт отрасли с большим коммьюнити и экосистемой

Обновления

Постоянные обновления и поддержка самых новых моделей

Экосистема

LMCache для KV-кеша,
vLLM Production Stack, llm-d, Ray

Универсальность

Широкий выбор моделей, оборудования и способов деплоя

SGLang

sgl-project/sglang

Производительность

Оптимизации под высокую пропускную способность и низкие задержки

Технологичность

RadixCache, Zero-Overhead Batch Scheduler

Интеграции

HiCache для многоуровнего кеша, Mooncake и DeepSeek 3FS

Не для всех :(

Некоторые модели показывают худший результат не смотря на оптимизации

NVIDIA Dynamo

ai-dynamo/dynamo

Платформенность

Единый формат манифестов и гибкость в конфигурации

Гибкость

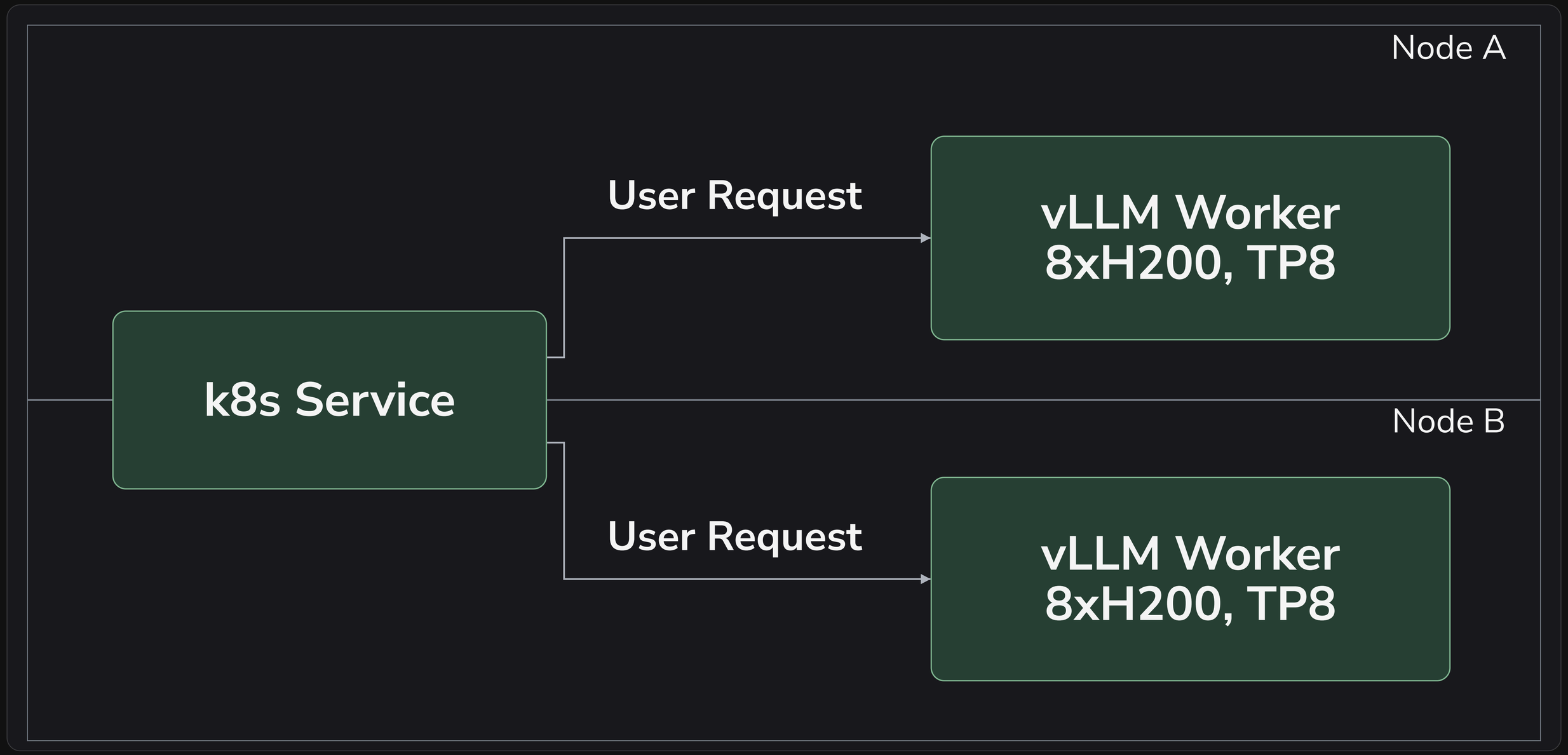
Поддержка vLLM, SGLang и Tensor-RT в качестве бекендов

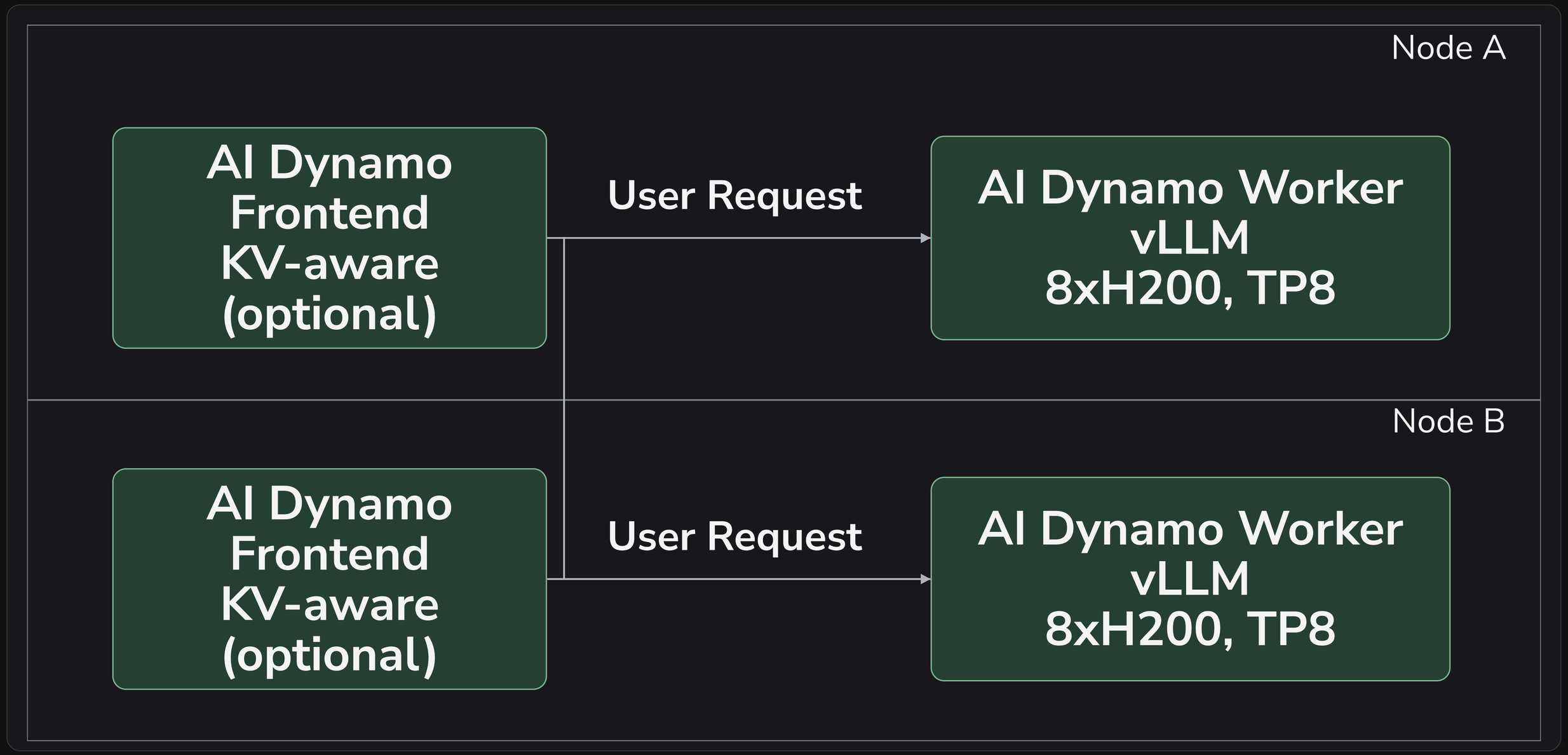
Интеграции

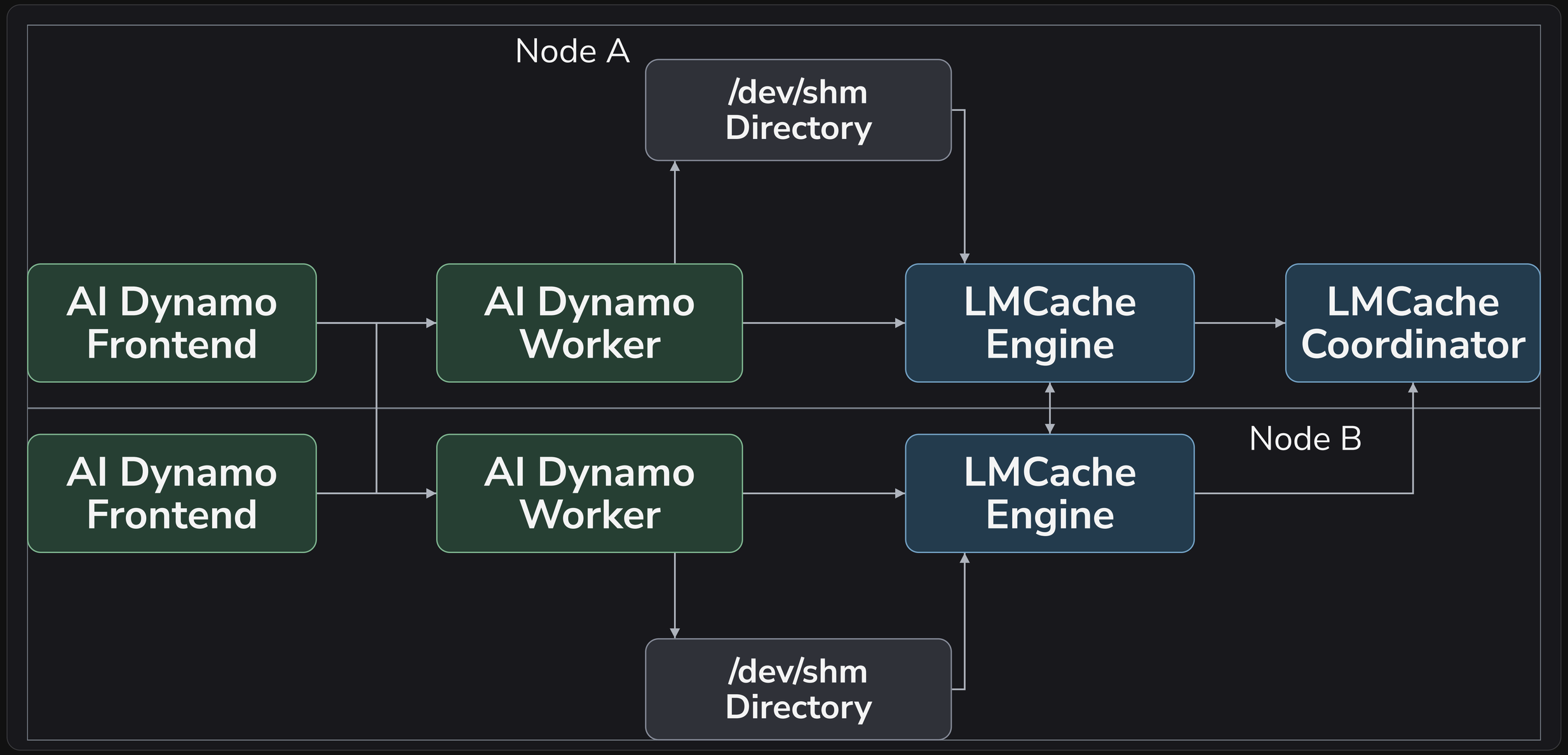
KV-Aware Routing, Cache Offloading, Disaggregated Serving

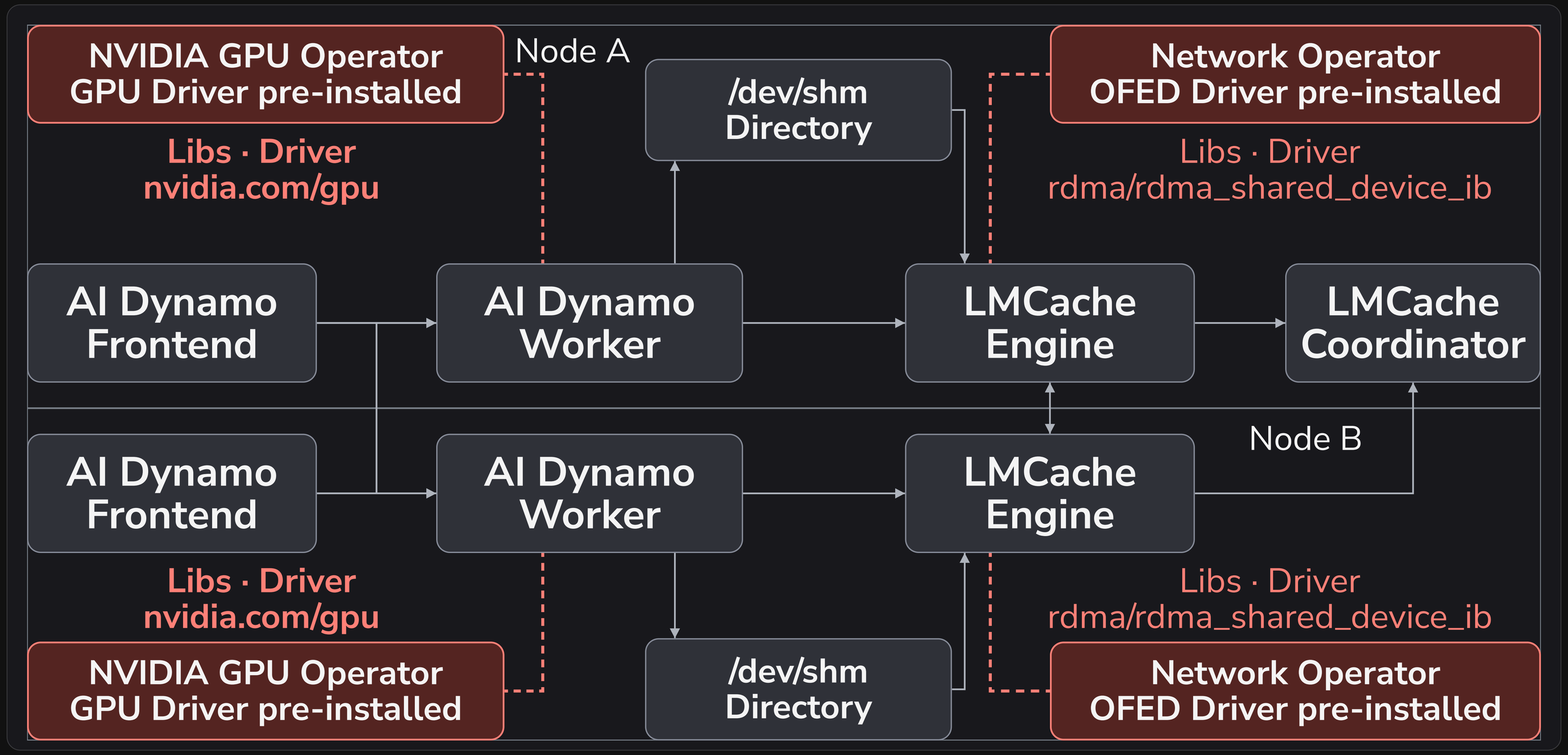
Надежность

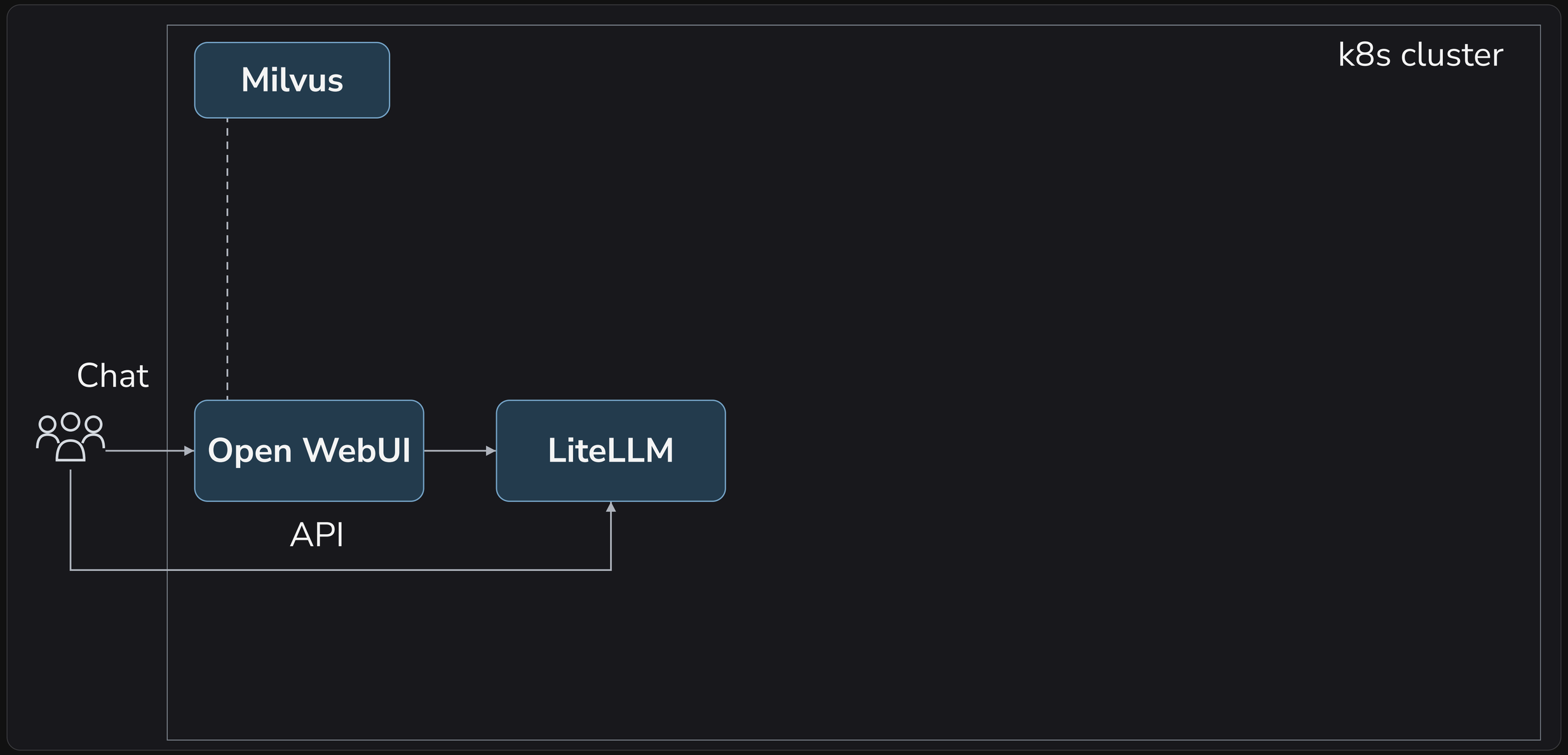
Рантайм образы с проверенной совместимостью сабкомпонентов

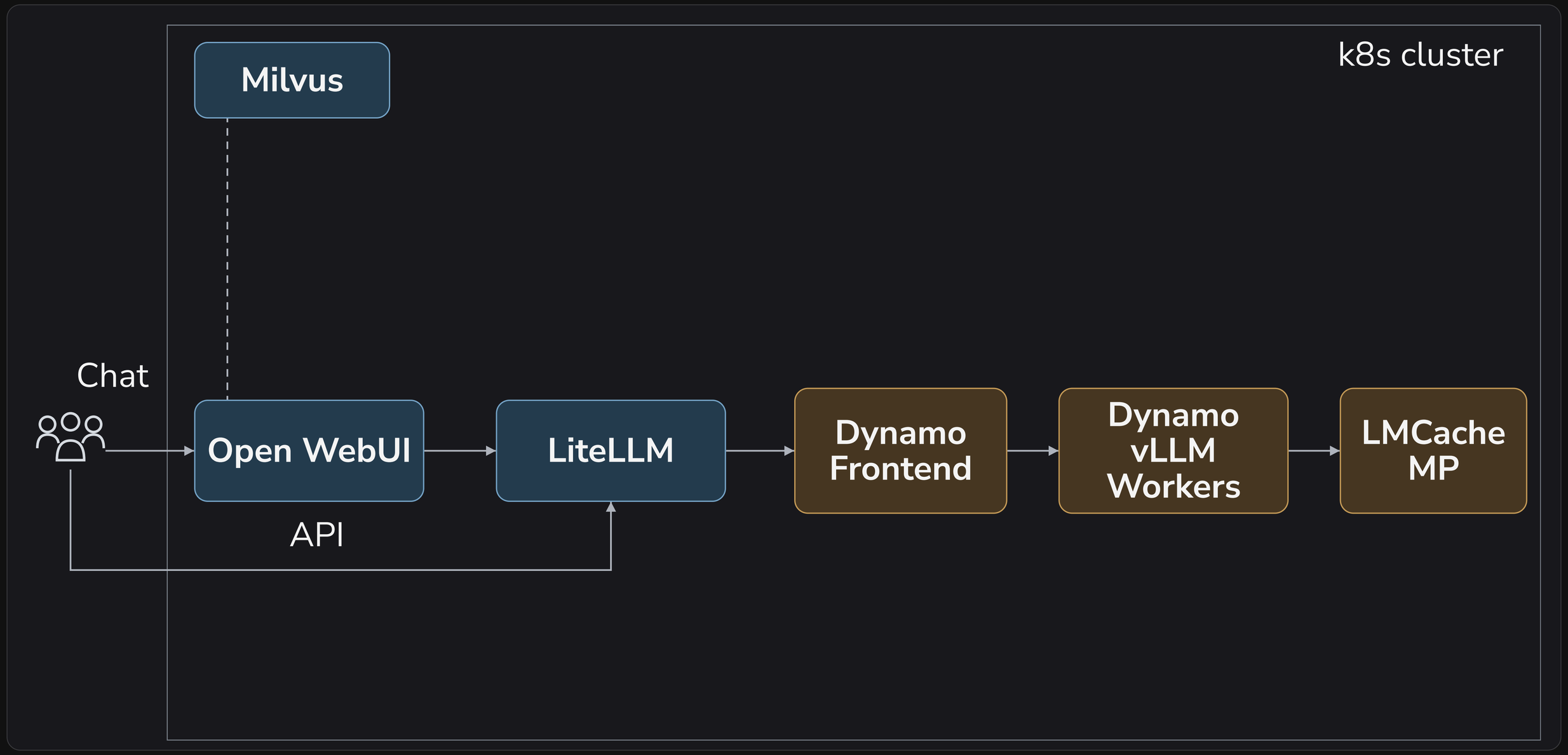

Node A
User Request
vLLM Worker
8xH200, TP8
k8s Service
Node B
User Request
vLLM Worker
8xH200, TP8

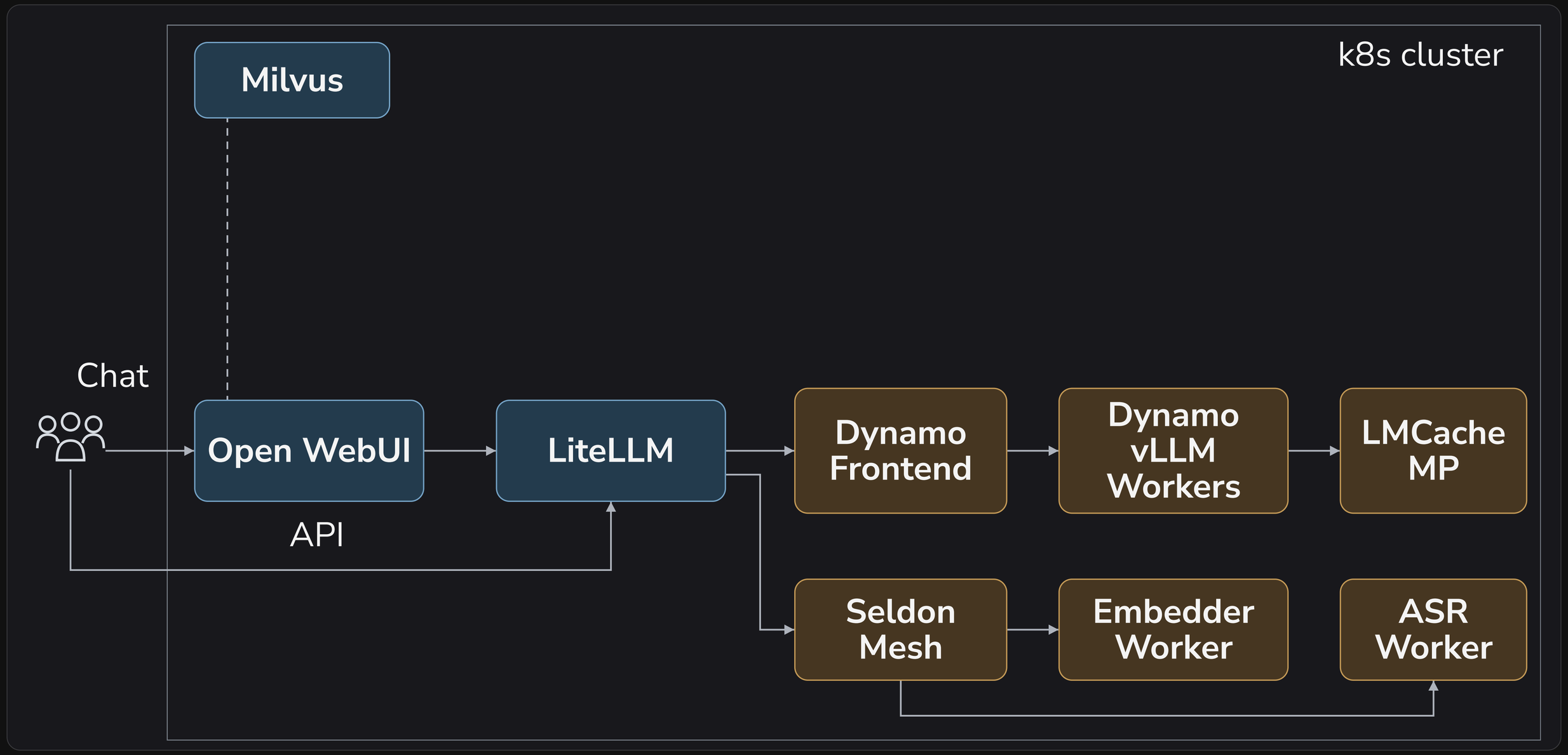

Node A
AI Dynamo Frontend KV-aware (optional)
User Request
AI Dynamo Worker vLLM 8xH200, TP8
Node B
AI Dynamo Frontend KV-aware (optional)
User Request
AI Dynamo Worker vLLM 8xH200, TP8

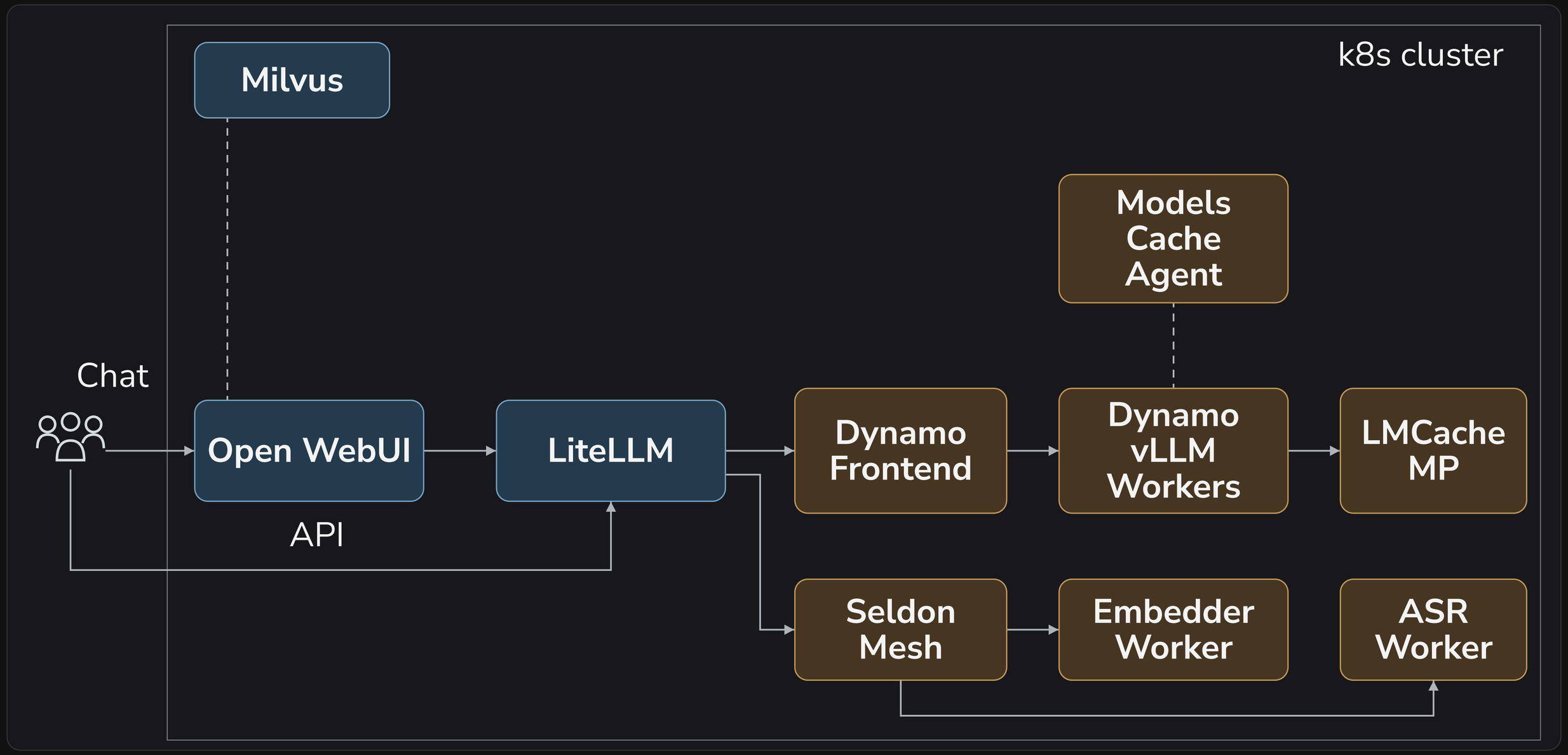

Node A
/dev/shm Directory
AI Dynamo Frontend
AI Dynamo Worker
LMCache Engine
LMCache Coordinator
Node B
AI Dynamo Frontend
AI Dynamo Worker
LMCache Engine
/dev/shm Directory

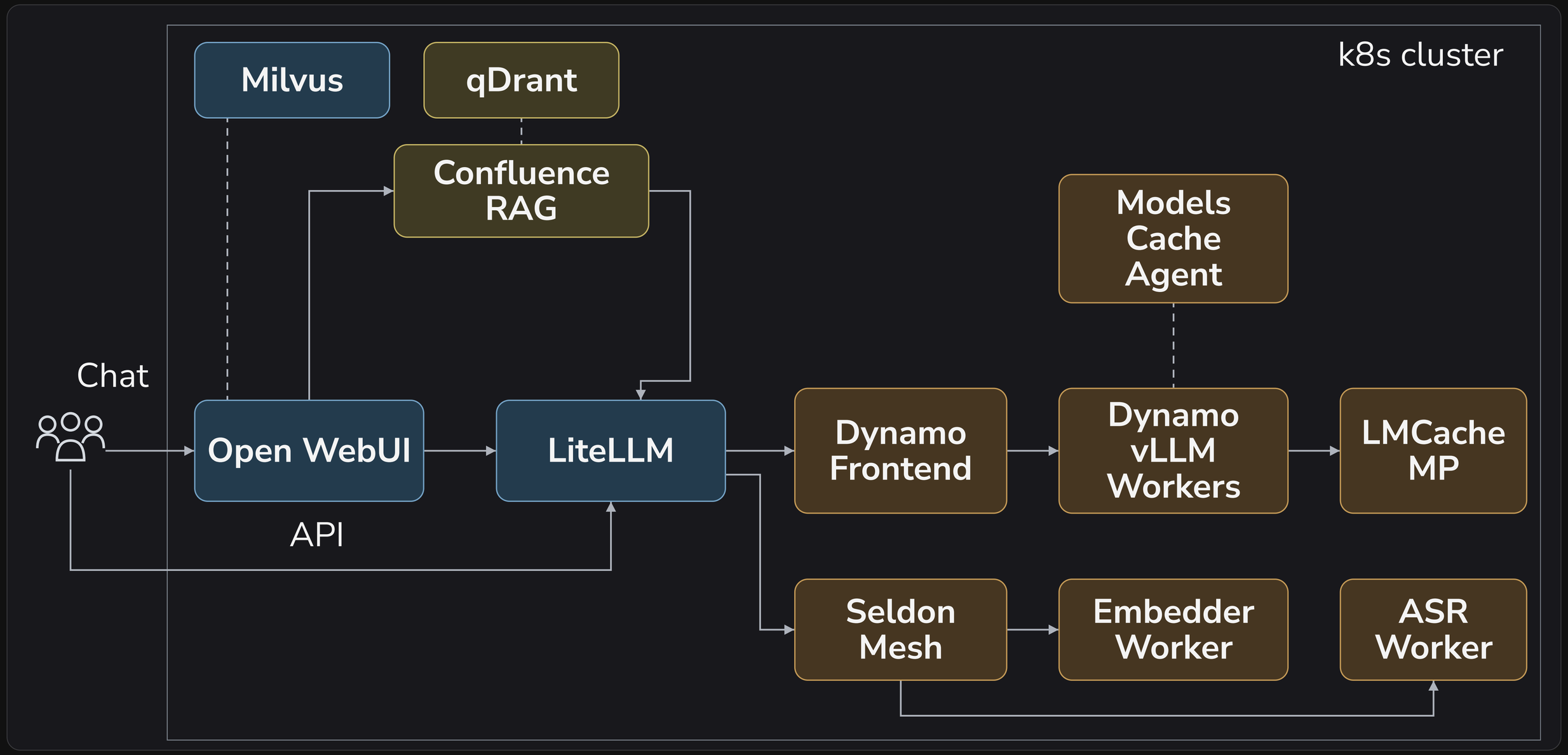

NVIDIA GPU Operator
GPU Driver pre-installed
Node A
/dev/shm
Directory
Network Operator
OFED Driver pre-installed
Libs · Driver
nvidia.com/gpu
Libs · Driver
rdma/rdma_shared_device_ib
AI Dynamo
Frontend
AI Dynamo
Worker
LMCache
Engine
LMCache
Coordinator
Node B
AI Dynamo
Frontend
AI Dynamo
Worker
LMCache
Engine
Libs · Driver
nvidia.com/gpu
Libs · Driver
rdma/rdma_shared_device_ib
NVIDIA GPU Operator
GPU Driver pre-installed
/dev/shm
Directory
Network Operator
OFED Driver pre-installed

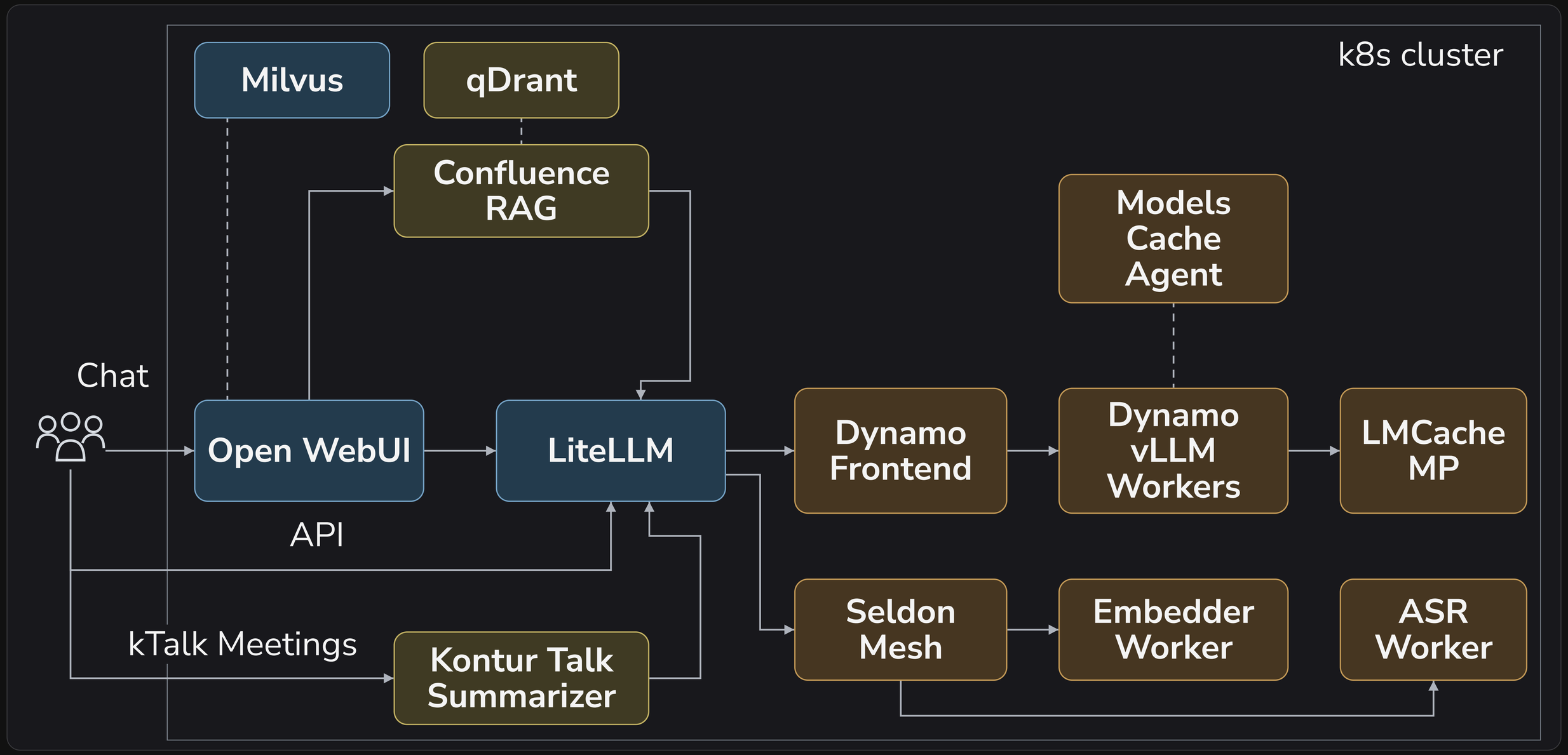

k8s cluster
Milvus
Chat
Open WebUI
LiteLLM
API

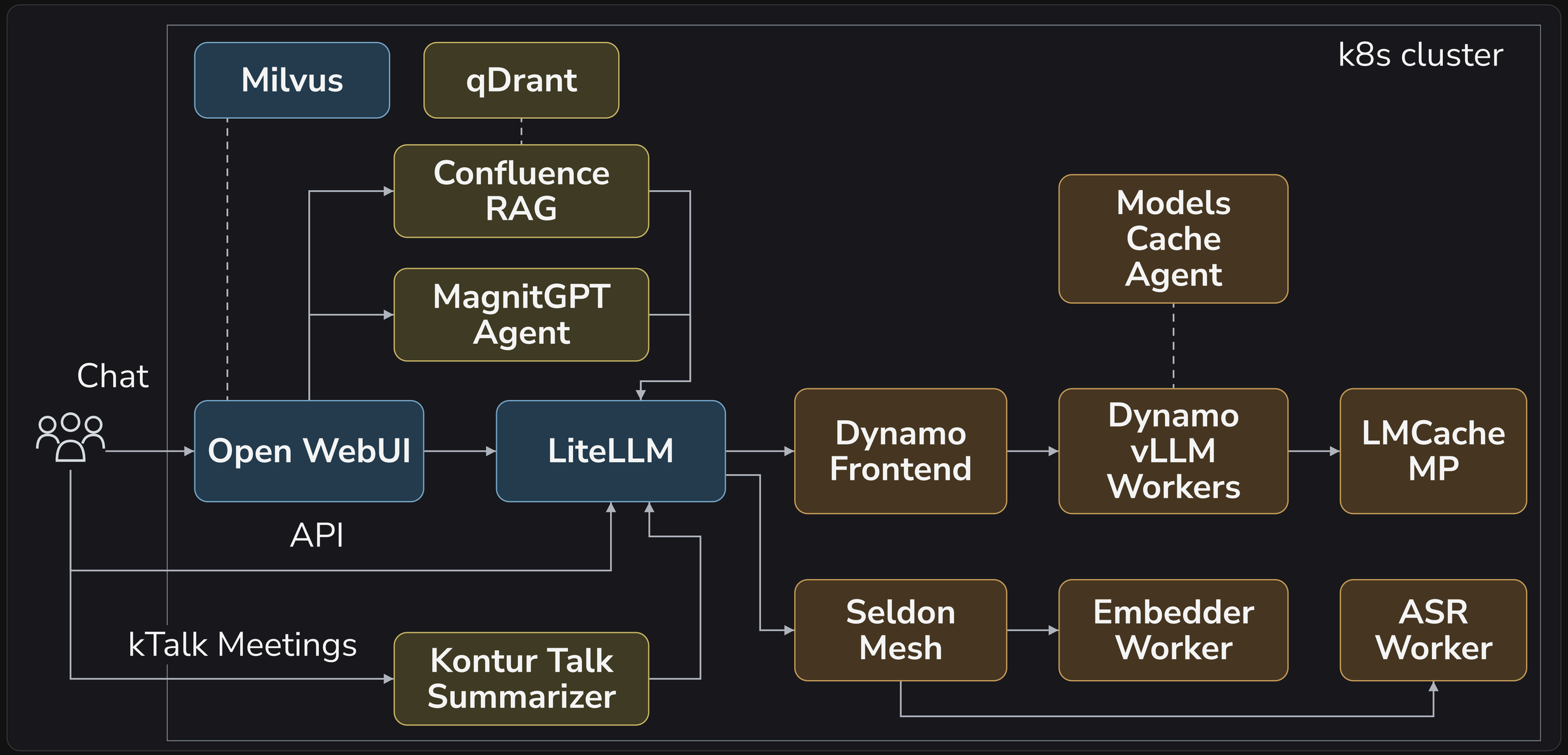

k8s cluster
Milvus
Chat
Open WebUI
LiteLLM
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
API


k8s cluster
Milvus
Chat
Open WebUI
LiteLLM
API
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
Seldon Mesh
Embedder Worker
ASR Worker


k8s cluster
Milvus
Models
Cache
Agent
Chat
Open WebUI
LiteLLM
Dynamo
Frontend
Dynamo
vLLM
Workers
LMCache
MP
API
Seldon
Mesh
Embedder
Worker
ASR
Worker


k8s cluster
Milvus
qDrant
Confluence RAG
Models Cache Agent
Chat
Open WebUI
LiteLLM
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
API
Seldon Mesh
Embedder Worker
ASR Worker


k8s cluster
Milvus
qDrant
Confluence RAG
Models Cache Agent
Chat
Open WebUI
LiteLLM
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
API
Seldon Mesh
Embedder Worker
ASR Worker
kTalk Meetings
Kontur Talk Summarizer


k8s cluster
Milvus
qDrant
Confluence RAG
MagnitGPT Agent
Models Cache Agent
Chat
Open WebUI
LiteLLM
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
API
kTalk Meetings
Kontur Talk Summarizer
Seldon Mesh
Embedder Worker
ASR Worker

MagnitGPT Agent

01

Задачи

- Анализ данных
- Визуализации
- Документы
- Таблицы
- БД
- Произвольный код

MAGNIT TECH

MagnitGPT Agent

01

Задачи

- Анализ данных
- Визуализации
- Документы
- Таблицы
- БД
- Произвольный код

02

Среда

- Поды по запросу
- Быстрая готовность
- Предустановленные либы
- Жизненный цикл

MagnitGPT Agent

01

Задачи

- Анализ данных
- Визуализации
- Документы
- Таблицы
- БД
- Произвольный код

02

Среда

- Поды по запросу
- Быстрая готовность
- Предустановленные либы
- Жизненный цикл

03

Интеграция

- Работа через OWUI
- Синхронизация файлов
- Сохранение артефактов

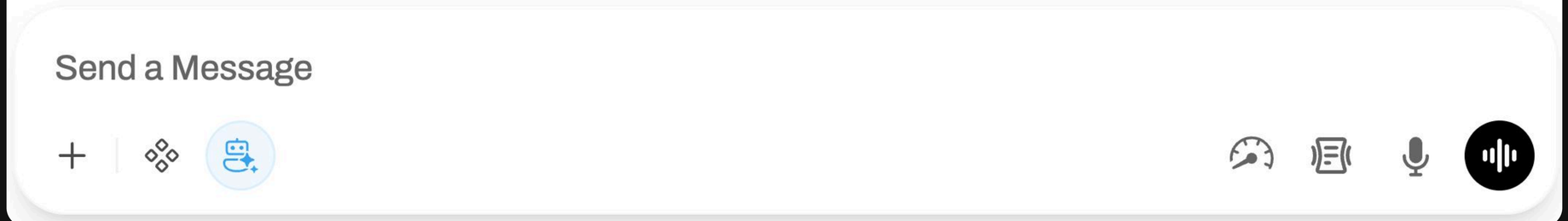
MagnitGPT Agent – Open WebUI

Метод: итерация $z \leftarrow z^2 + c$ с векторизацией через NumPy — на каждом шаге обновляются только ещё не «вышедшие» точки ($|z| \le 2$), что заметно ускоряет расчёт.

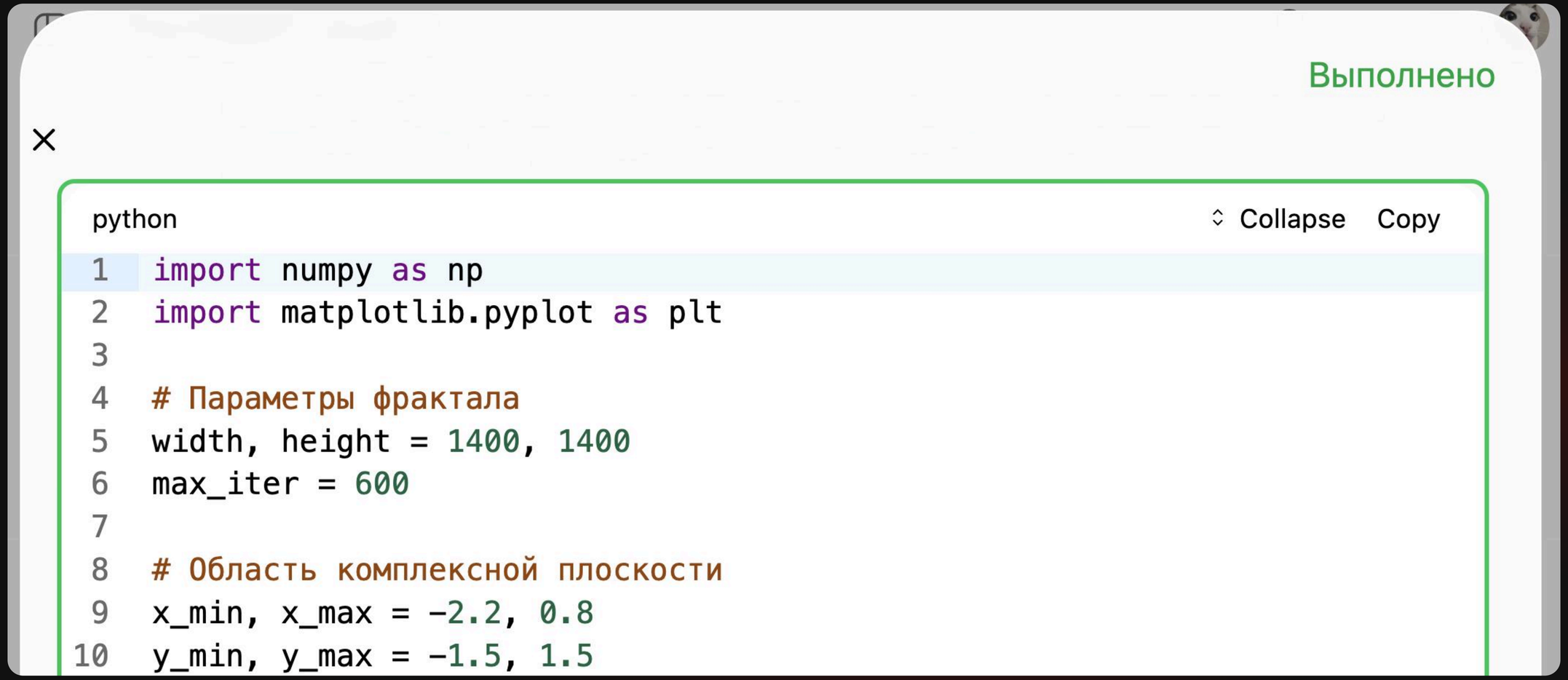
Скачать файл

Если хочешь, могу сделать зум в интересную область (например, «спираль» у (–0.75, 0.1) или «ракушку» у (0.285, 0.01)), другое разрешение или другую палитру.

Выполнено Выполнено



MagnitGPT Agent – Open WebUI



Jupyter Enterprise Gateway

jupyter-server/enterprise_gateway

01

Кастомные кернелы

Зашиваем необходимые зависимости в образ, отдаём JEG спецификацию, и он поднимает поды за

1–1,5 секунды

MAGNIT TECH

Jupyter Enterprise Gateway

jupyter-server/enterprise_gateway

01

Кастомные кернелы

Зашиваем необходимые зависимости в образ, отдаём JEG спецификацию, и он поднимает поды за

1–1,5 секунды

02

Жизненный цикл

JEG следит за состоянием подов и процессов внутри, неактивные удаляются по таймауту

Jupyter Enterprise Gateway

jupyter-server/enterprise_gateway

01

Кастомные кернелы

Зашиваем необходимые зависимости в образ, отдаём JEG спецификацию, и он поднимает поды за

1–1,5 секунды

02

Жизненный цикл

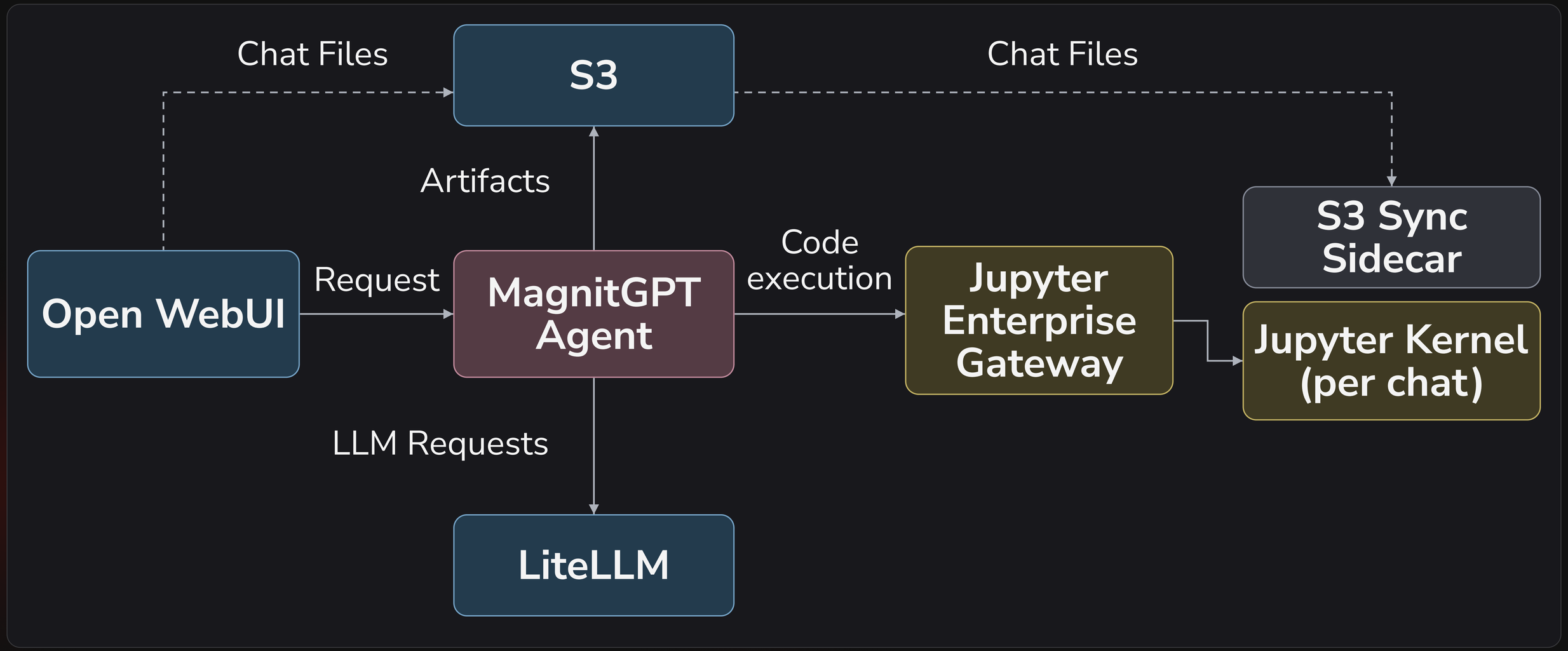
JEG следит за состоянием подов и процессов внутри, неактивные удаляются по таймауту

03

Синхронизация файлов

Самописный sidecar-сервис, скачивающий файлы конкретного пользователя из S3

MagnitGPT Agent – Jupyter Enterprise Gateway



Agent Sandbox

kubernetes-sigs/agent-sandbox

01

Оператор

Создание ресурсов через CDR, отлично ложится в IaC-паттерны

Agent Sandbox

kubernetes-sigs/agent-sandbox

01

Оператор

Создание ресурсов через CDR, отлично ложится в IaC-паттерны

02

Роутер

Позволяет отделить data-plane от уровня управления ресурсами

Agent Sandbox

kubernetes-sigs/agent-sandbox

01

Оператор

Создание ресурсов через CDR, отлично ложится в IaC-паттерны

02

Роутер

Позволяет отделить data-plane от уровня управления ресурсами

03

Скорость

WarmPool’ы моментально отдают предсозданные поды, ресурс выделяется за

200 мс

Agent Sandbox

kubernetes-sigs/agent-sandbox

01

Оператор

Создание ресурсов через CDR, отлично ложится в IaC-паттерны

02

Роутер

Позволяет отделить data-plane от уровня управления ресурсами

03

Скорость

WarmPool’ы моментально отдают предсозданные поды, ресурс выделяется за

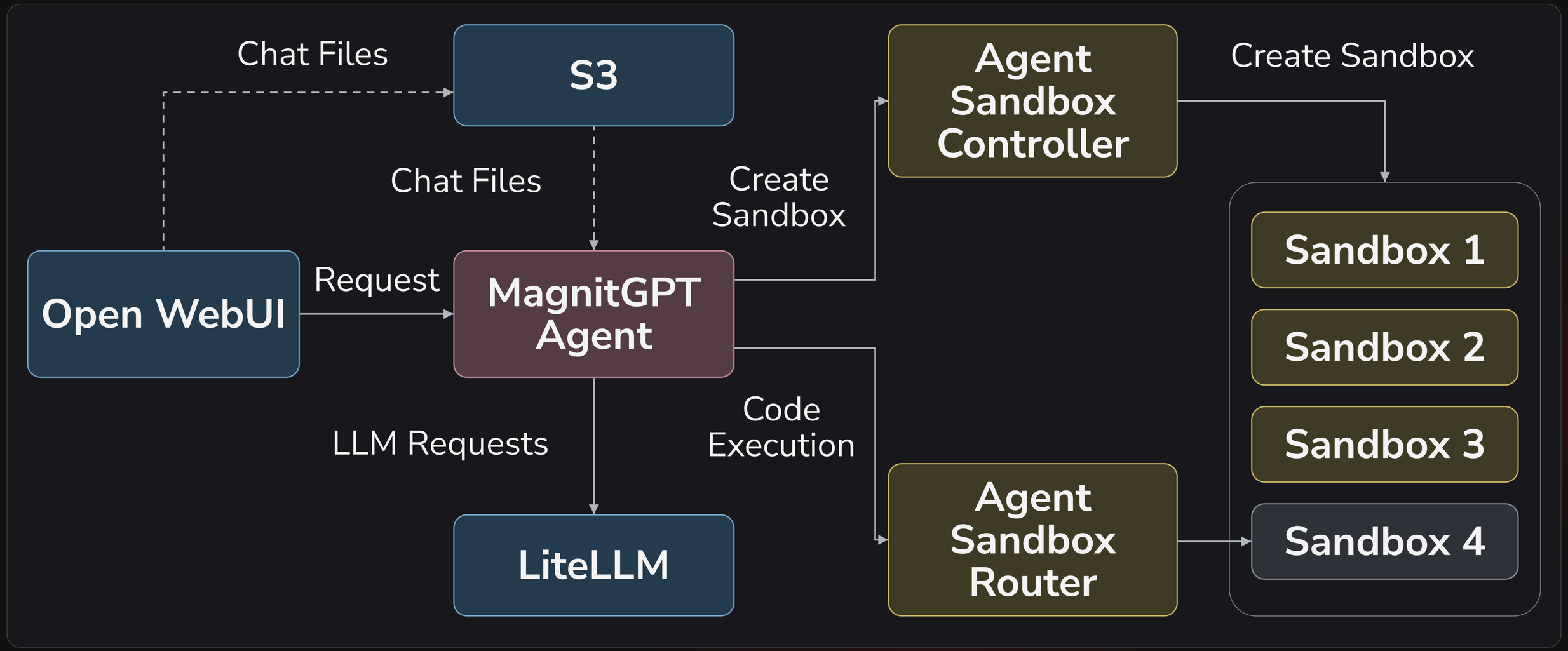
200 мс

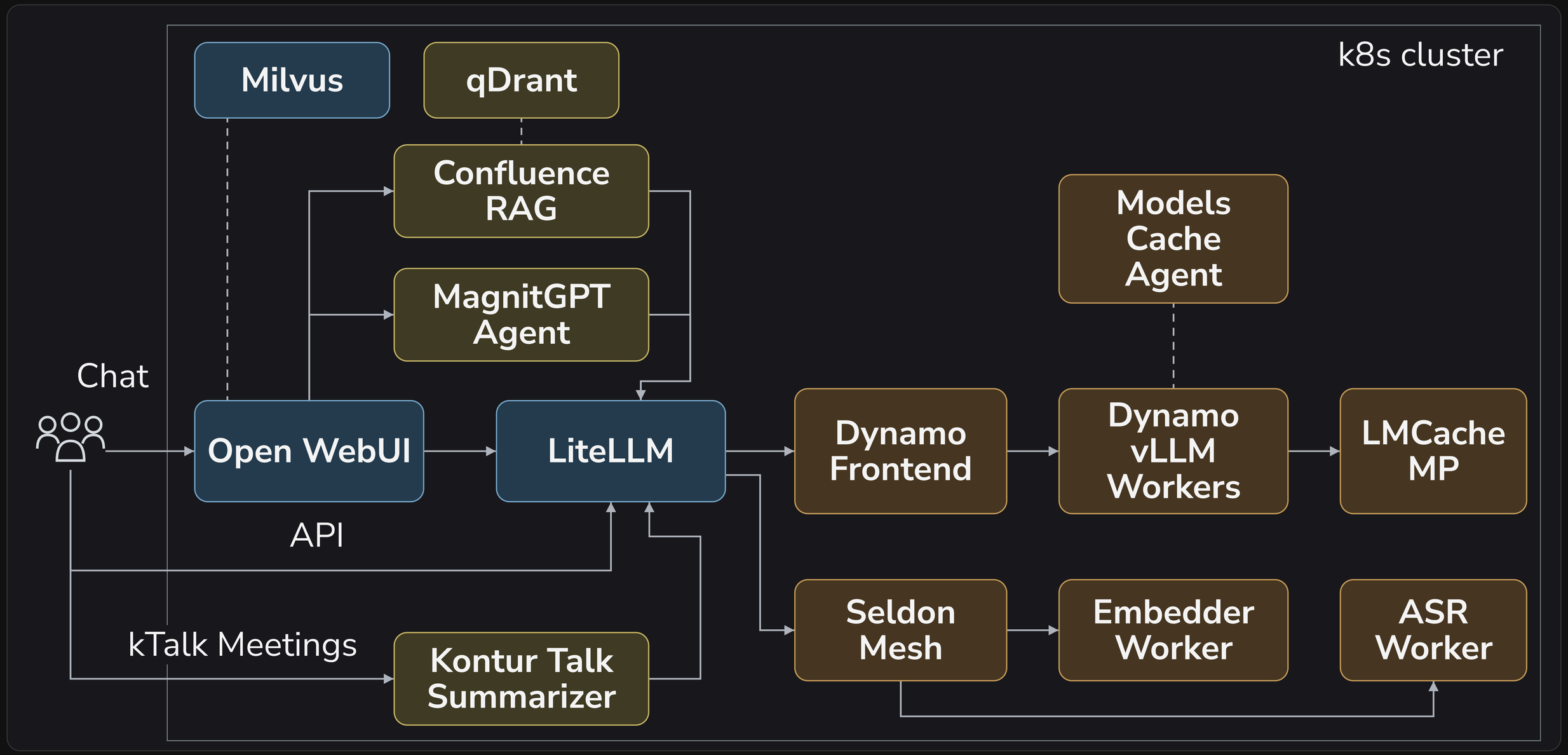
04

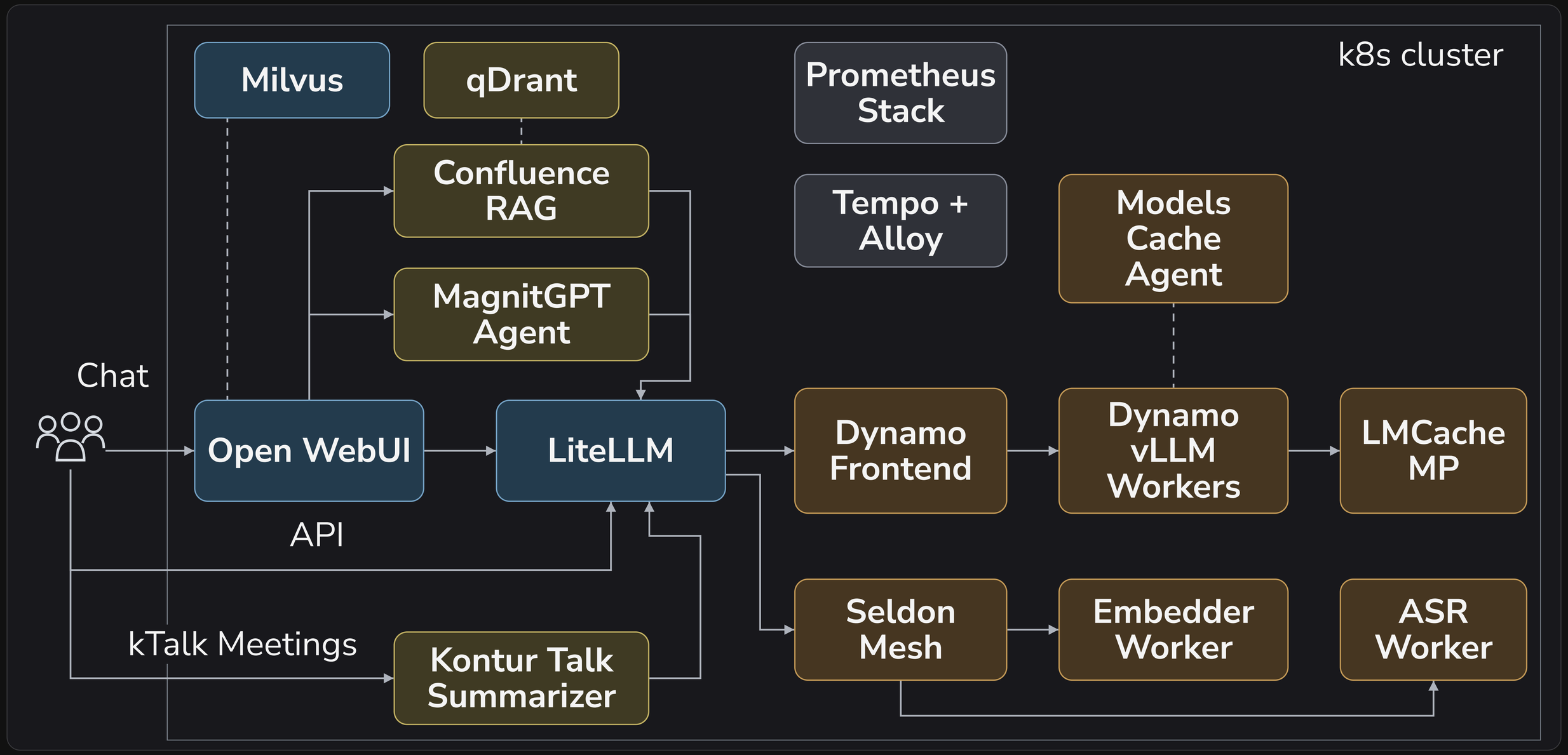
Файлы

Синхронизация файлов через SDK, сайдкары больше не нужны

MagnitGPT Agent – Agent Sandbox




k8s cluster
Milvus
qDrant
Confluence RAG
MagnitGPT Agent
Models Cache Agent
Chat
Open WebUI
LiteLLM
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
API
kTalk Meetings
Kontur Talk Summarizer
Seldon Mesh
Embedder Worker
ASR Worker


k8s cluster
Milvus
qDrant
Prometheus Stack
Confluence RAG
Tempo + Alloy
Models Cache Agent
MagnitGPT Agent
Chat
Open WebUI
LiteLLM
Dynamo Frontend
Dynamo vLLM Workers
LMCache MP
API
Seldon Mesh
Embedder Worker
ASR Worker
kTalk Meetings
Kontur Talk Summarizer

Open Source Moment

Давай завтра	NVIDIA Dynamo обновляется гораздо реже vLLM, SGLang и прочих дочерних компонентов

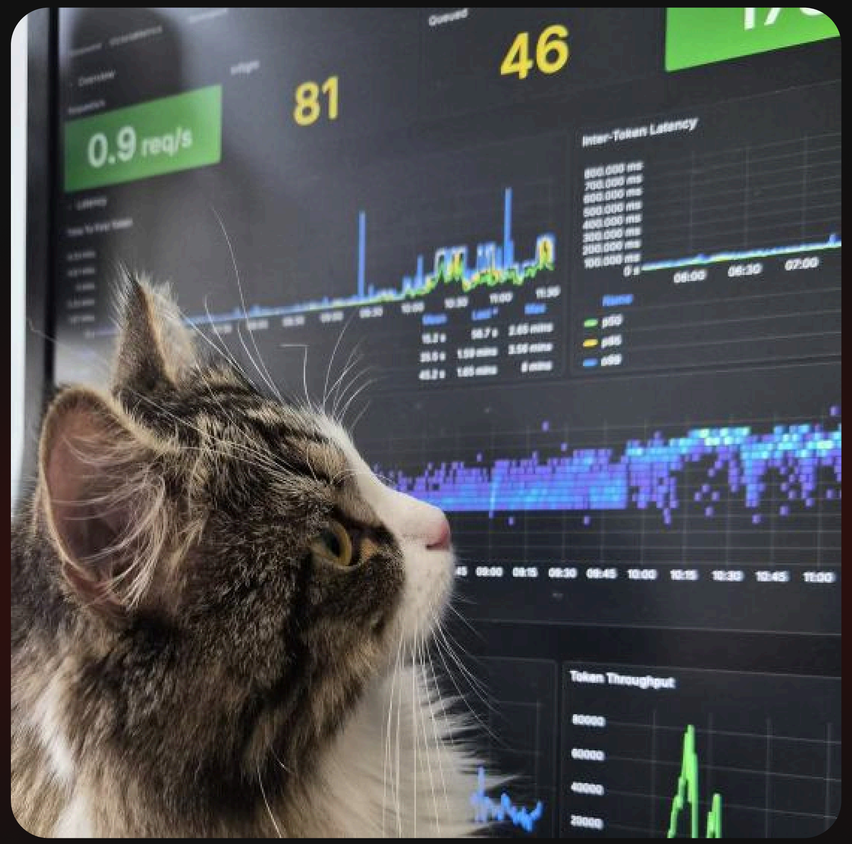
Open Source Moment

Давай завтра	NVIDIA Dynamo обновляется гораздо реже vLLM, SGLang и прочих дочерних компонентов
Нас не затронет	Больше компонентов – больше потенциальных багов. Что-то можно исправить самим, а на что-то уже не хватает ресурса

Open Source Moment

Давай завтра	NVIDIA Dynamo обновляется гораздо реже vLLM, SGLang и прочих дочерних компонентов
Нас не затронет	Больше компонентов – больше потенциальных багов. Что-то можно исправить самим, а на что-то уже не хватает ресурса
Хотим больше фичей	Не существует платформы, способной закрыть все потребности пользователей

SLO



Это Кики!

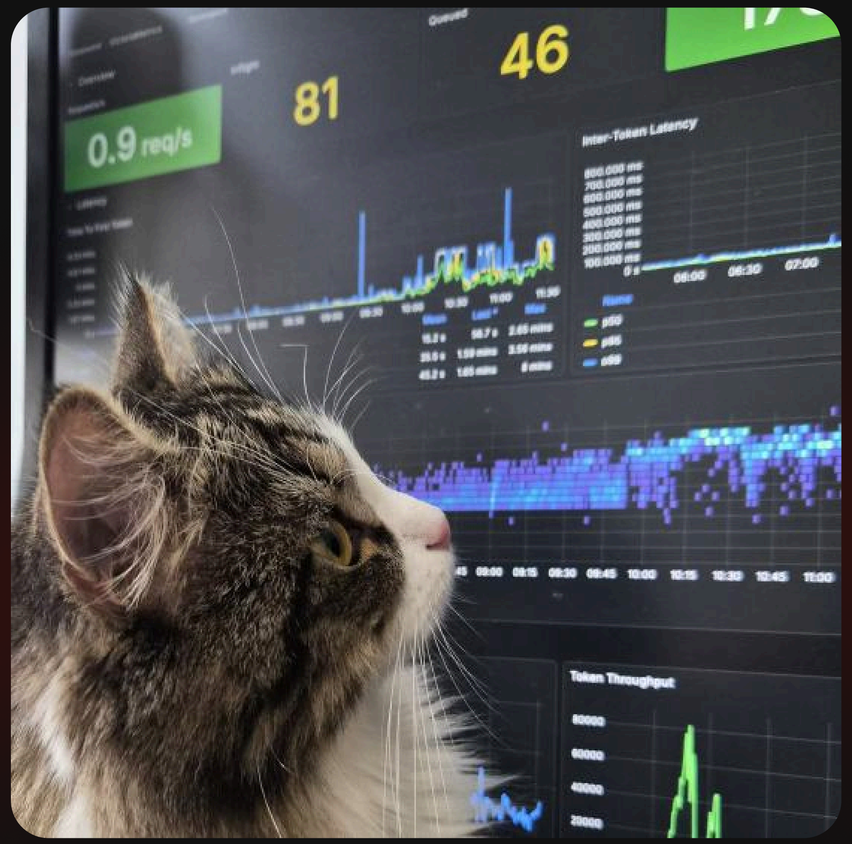
И она в шоке с метрик MagnitGPT

“Откуда столько народу??”

“Почему TTFT 8 минут, а E2E 26?!”

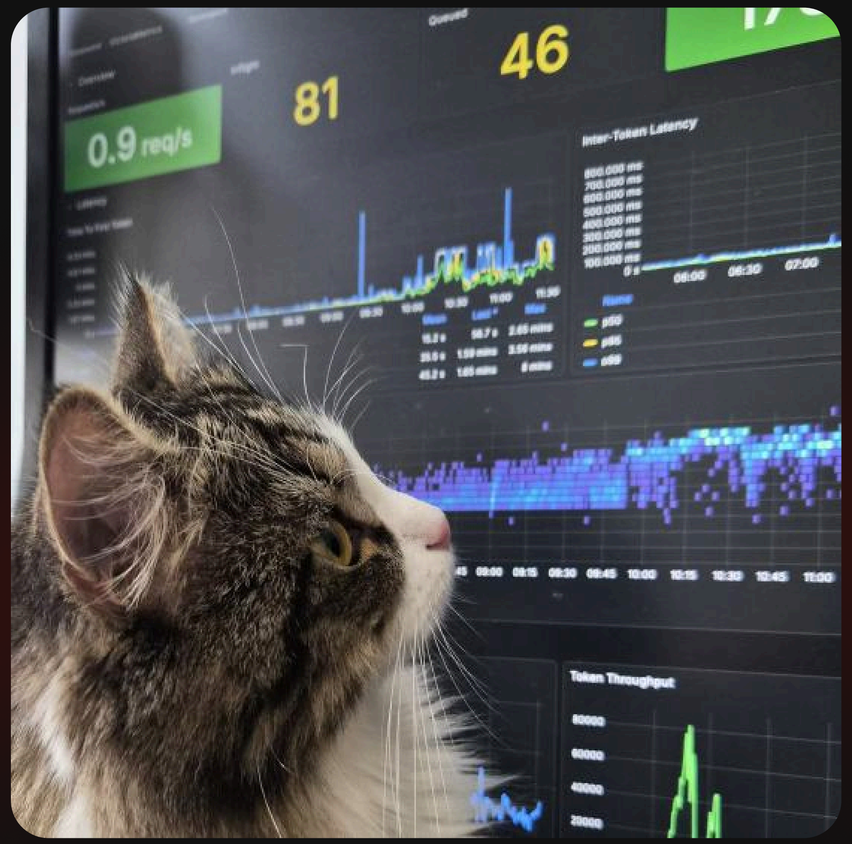
“Хочу бегать за плюшевой мышью, а не вот это все...”

SLO



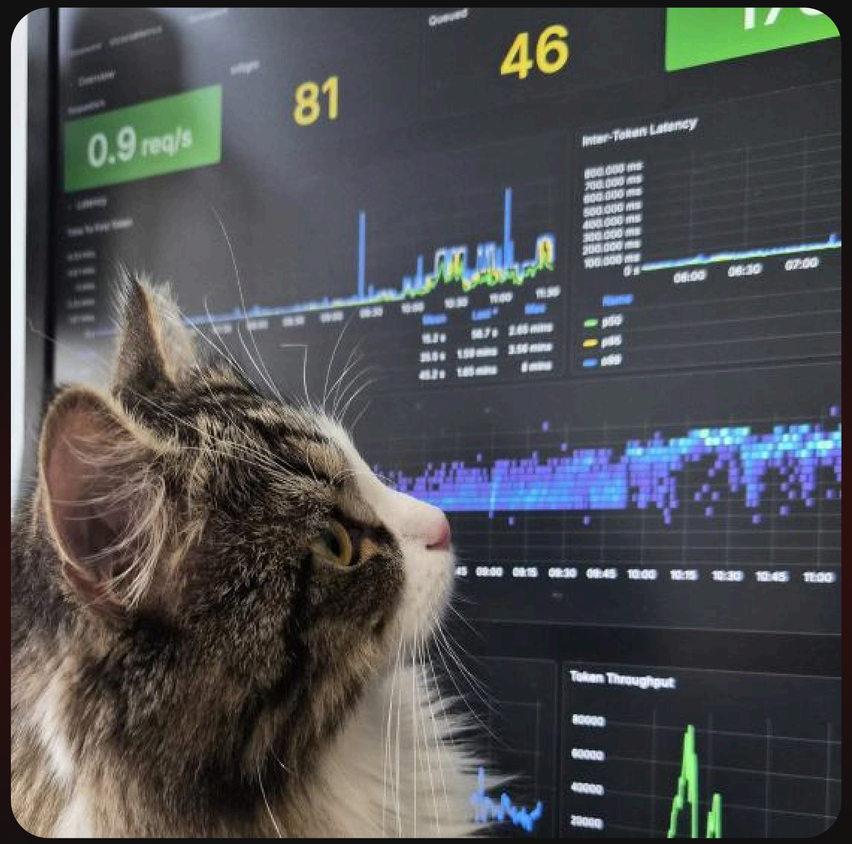
Rate Limits	Общие или индивидуальные в зависимости от кейса

SLO



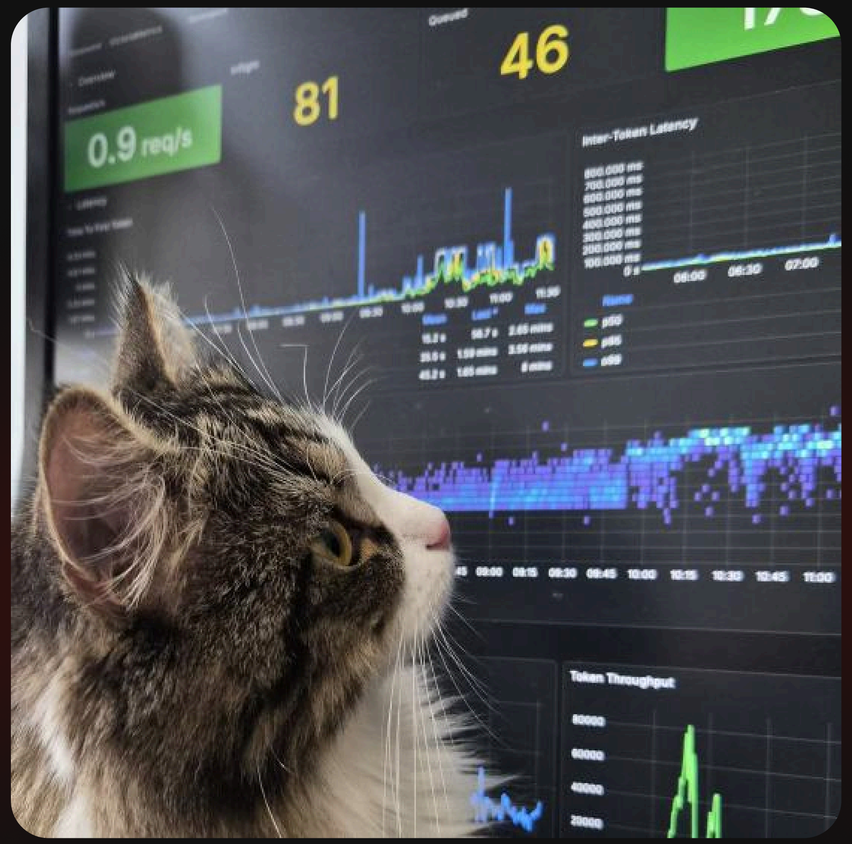
Rate Limits	Общие или индивидуальные в зависимости от кейса
Тестирование	Нагрузочное тестирование моделей с учетом TTFT и E2E

SLO



Rate Limits	Общие или индивидуальные в зависимости от кейса
Тестирование	Нагрузочное тестирование моделей с учетом TTFT и E2E
Мониторинг	Выявление аномального потребления токенов, автоотчетность

SLO



Rate Limits	Общие или индивидуальные в зависимости от кейса
Тестирование	Нагрузочное тестирование моделей с учетом TTFT и E2E
Мониторинг	Выявление аномального потребления токенов, автоотчетность
Автоскейлинг	Разделение на префил-декод и Dynamo SLA-Based Planner

Корпоративный GPT

Как внедрить так, чтобы все остались довольны



Унтура Михаил

Главный MLOps-инженер
MAGNIT TECH

Канал MAGNIT TECH