

LLM Gateway

наводим порядок в работе
с языковыми моделями



Роман Песков
старший инженер-разработчик
ML-платформа, Циан



Обо мне

9 лет разработки на Python

В Циане начинал как python developer, реализовывал backend:

- ml-ранжирования
- движка рекомендаций
- платформы a/b-экспериментов

4 года в ML-платформе Циана

- инструменты и сервисы MLOps-платформы Циан*
- занимаюсь внедрением и масштабированием AI-решений



*Эволюция MLOps в Циан



Стэк

- Python/Pytorch
- k8s, Spark, Hadoop
- MLflow, Airflow, z2jh
- Kafka, Redis, PG, Qdrant

30 в ML департаменте

- в основном DS/MLE
- в ML платформе 4 инженера

100+ различных моделей

- поиск и рекомендации
- модерация, анти-фрод
- анализ речи и изображений



Направления использования LLM в Циане

Новые продукты

- диалоговые рекомендательные системы
- автоматизированная модерация
- агенты поддержки клиентов

Автоматизация бизнеса

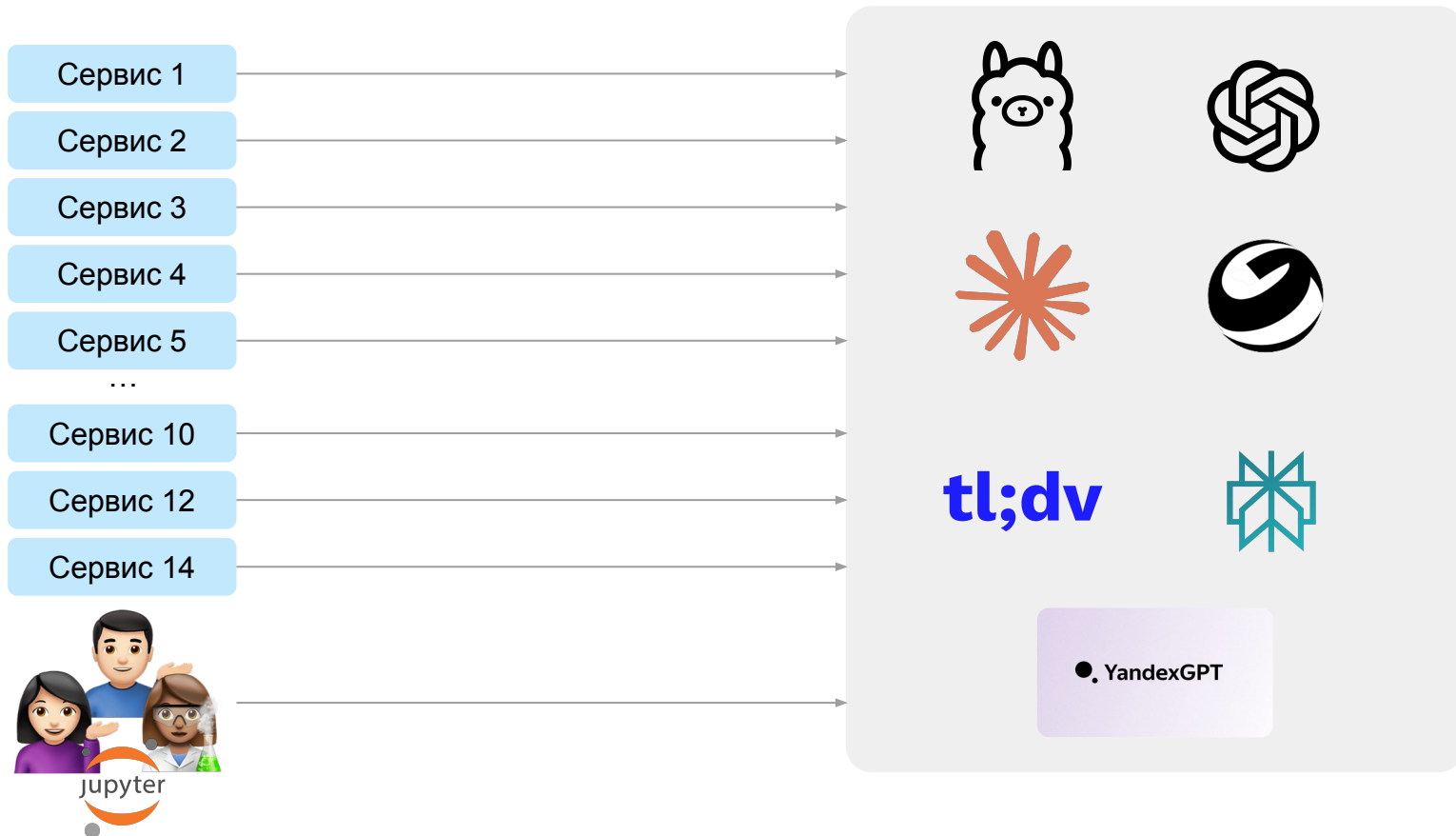
- автоматизация и анализ продаж
- генерация отчетов, документов, поиск рисков
- работа с резюме и откликами

Личная продуктивность

- поддержка в коде, письме
- создание контента
- суммаризаторы встреч, статей



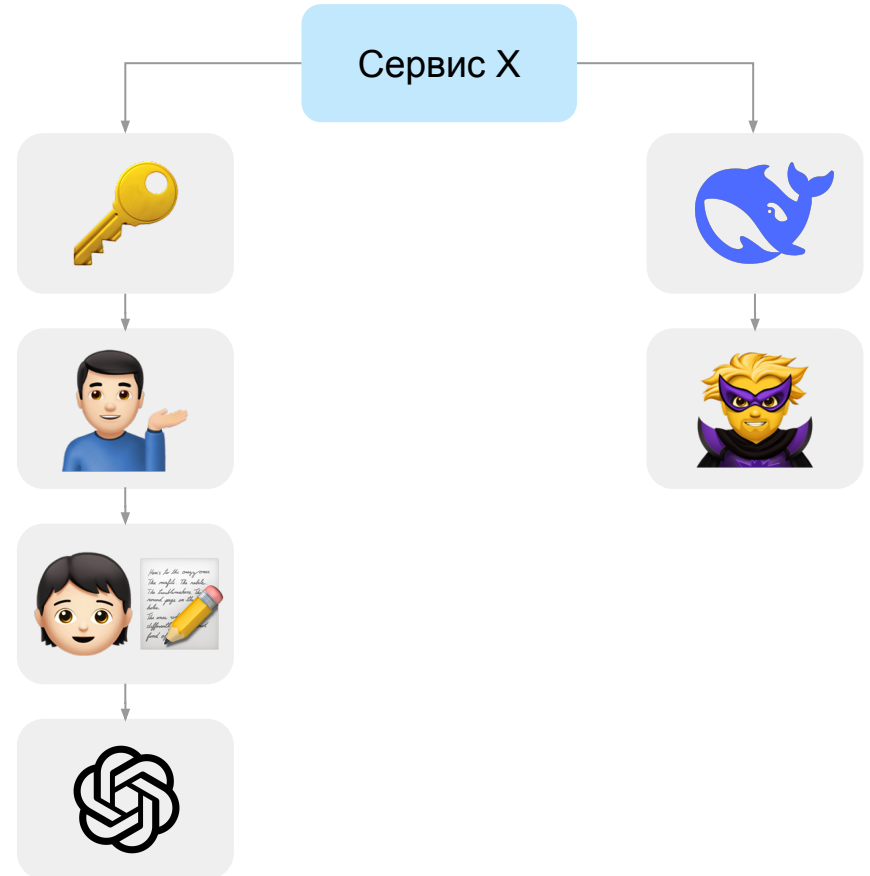
Взаимодействие с LLM



Безопасность



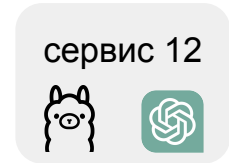
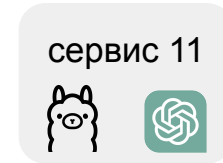
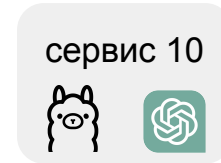
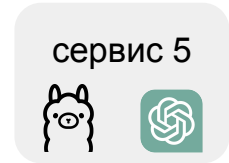
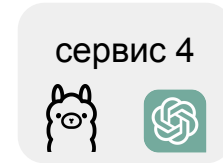
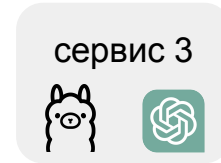
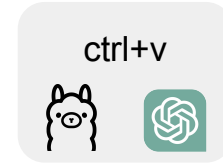
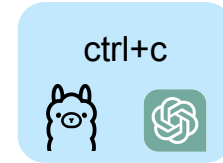
- Нет контроля что передаем
- Используются небезопасные провайдеры
- Шаринг ключей
- Несанкционированный доступ к LLM



Расширяемость



- дублирование логики и кода
- сложности с добавлением функционала
- различия в подходах и реализации



LLM Gateway

1 Централизация и унификация

единая точка доступа к различным языковым моделям

2 Мониторинг и аналитика

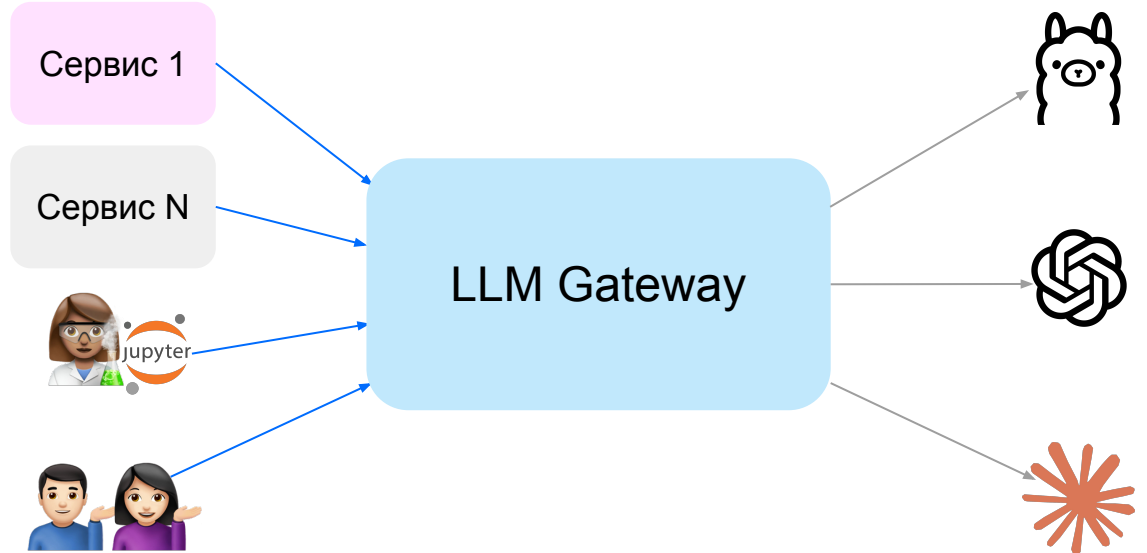
метрики и логи

3 Безопасность и контроль доступа

управление ключами и доступом к моделям

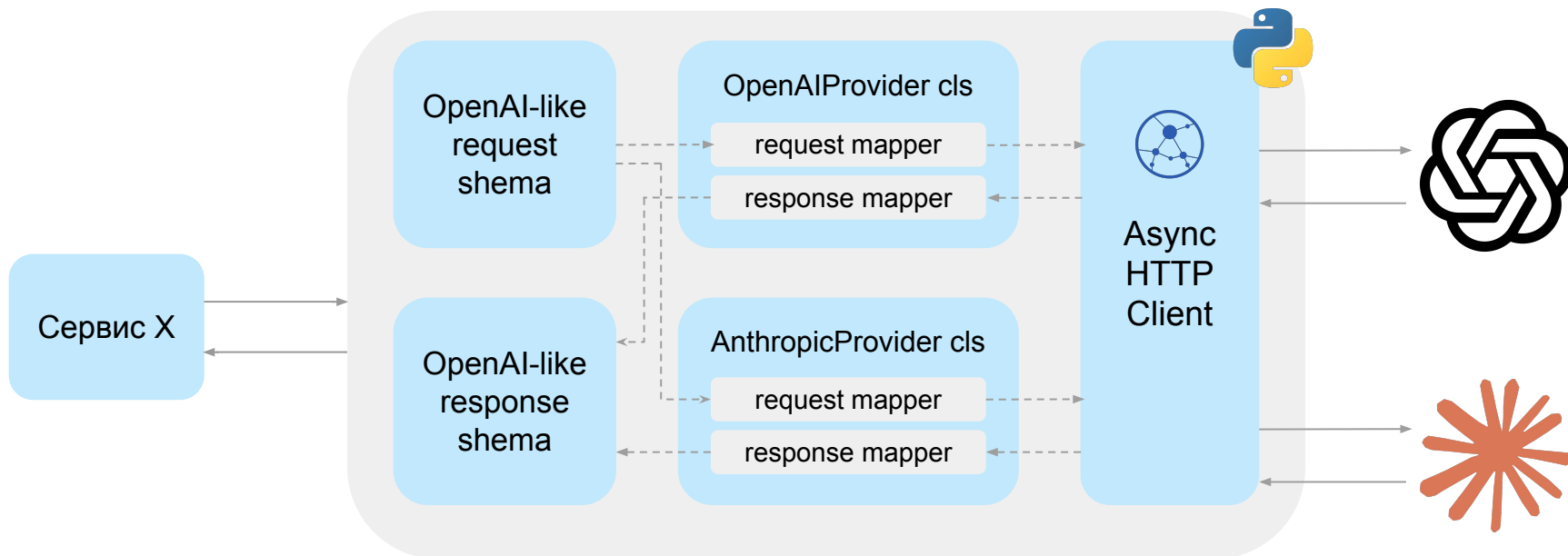
4 Гибкость и расширяемость

новые фичи без доработки клиентов



Централизация и унификация

- обычный микросервис на Python
- стандартизированный REST API для работы с разными LLM
- простое добавление новых провайдеров LLM



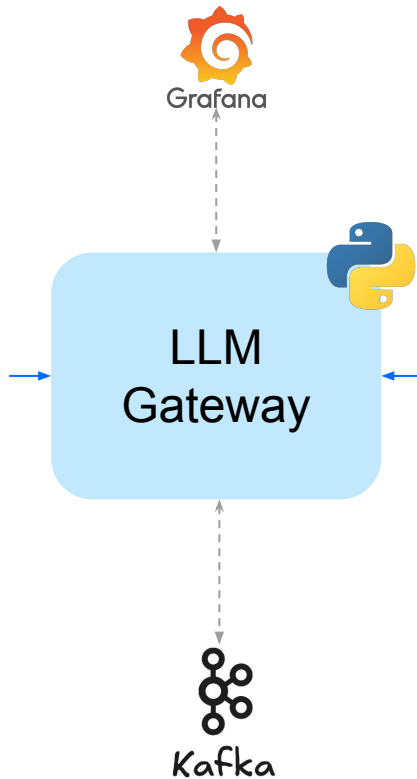
Мониторинг и аналитика

Метрики

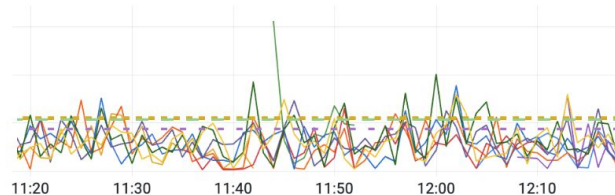
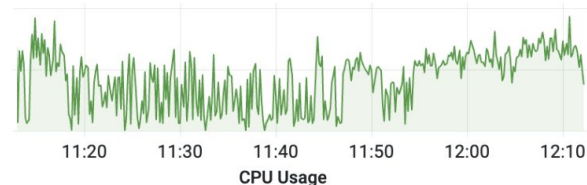
- rps, response time, latency
- ресурсы: инстансы, cpu, mem
- коды ошибок

Логи

- запрос: источник, параметры, тело сообщения
- ответ: тело ответа, токены
- ошибки

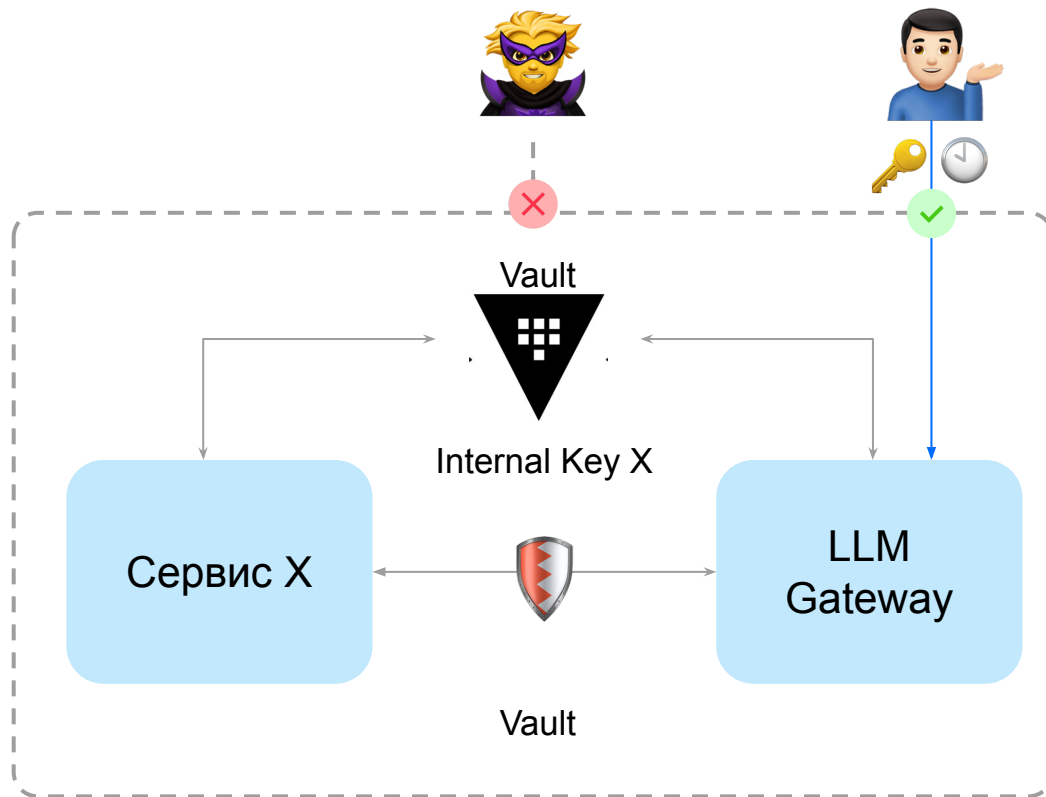


Время выполнения запросов по версиям (ELK, p90)



Безопасность и контроль доступа

- ограничение на сетевом уровне
- внутренние ключи
- безопасное хранение в Vault
- межсервисная аутентификация



LLM Gateway

Централизация
и унификация

①

единая точка доступа к различным
языковым моделям

Мониторинг и аналитика

②

метрики и логи

Безопасность и контроль
доступа

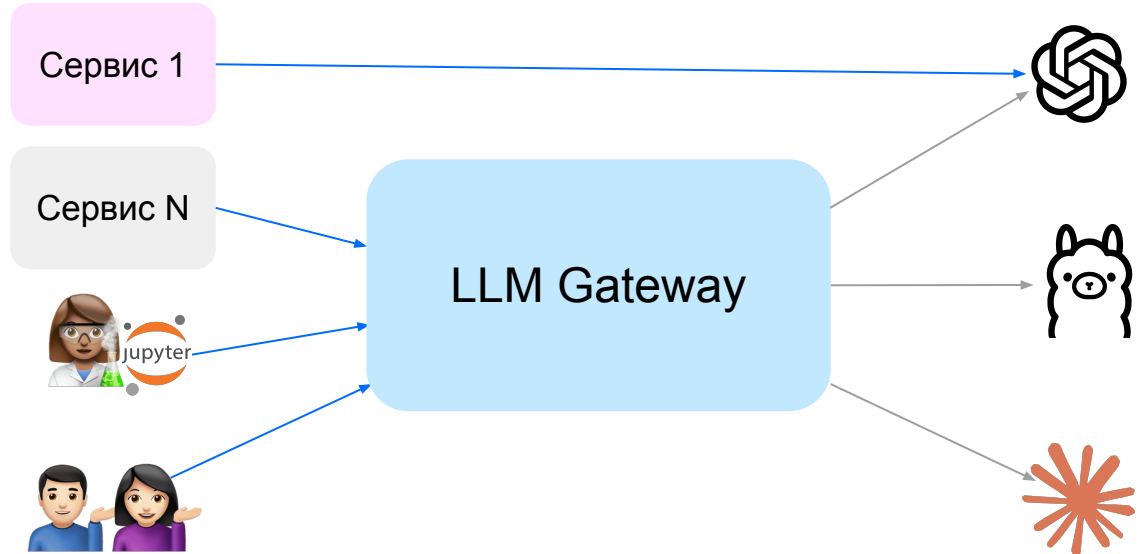
③

управление ключами и
доступом к моделям

Гибкость и расширяемость

④

новые фичи без доработки
клиентов



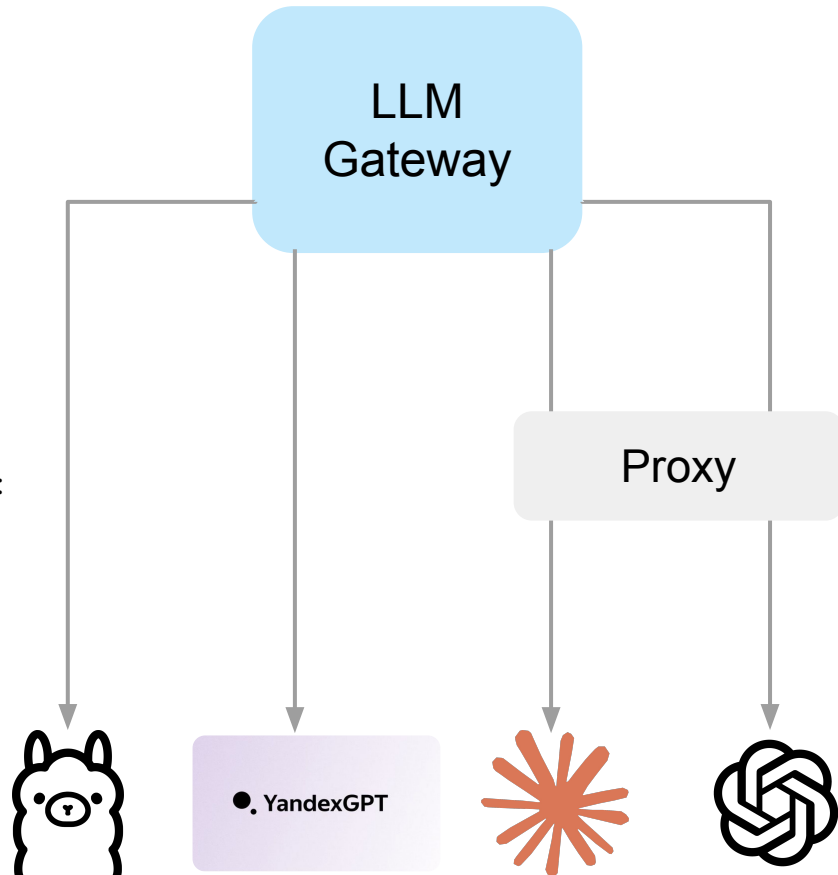
Интеграция с Proxy

❌

```
{
  "message": "Country, region, or territory
             not supported",
  ...
}
```

✅

- async with aiohttp.ClientSession() as session:
 - async with session.get(
 - url="https://api.openai.com/",
 - proxy="<http://proxy.com>",
 -) as resp:
 - ...



ЛИМИТЫ И КВОТЫ

```
{  
  "type": "insufficient_quota",  
  "message": "You exceeded your  
             current quota...",  
}
```

✓ **SRE Pager** APP
[ALERTING] ML Platform SLO / LLM Gateway:
утилизация квоты
Grafana
SRE Pager
Команда: ml / ml-platform
Сервис достиг использования 90% квоты.
ml-platform-mcs



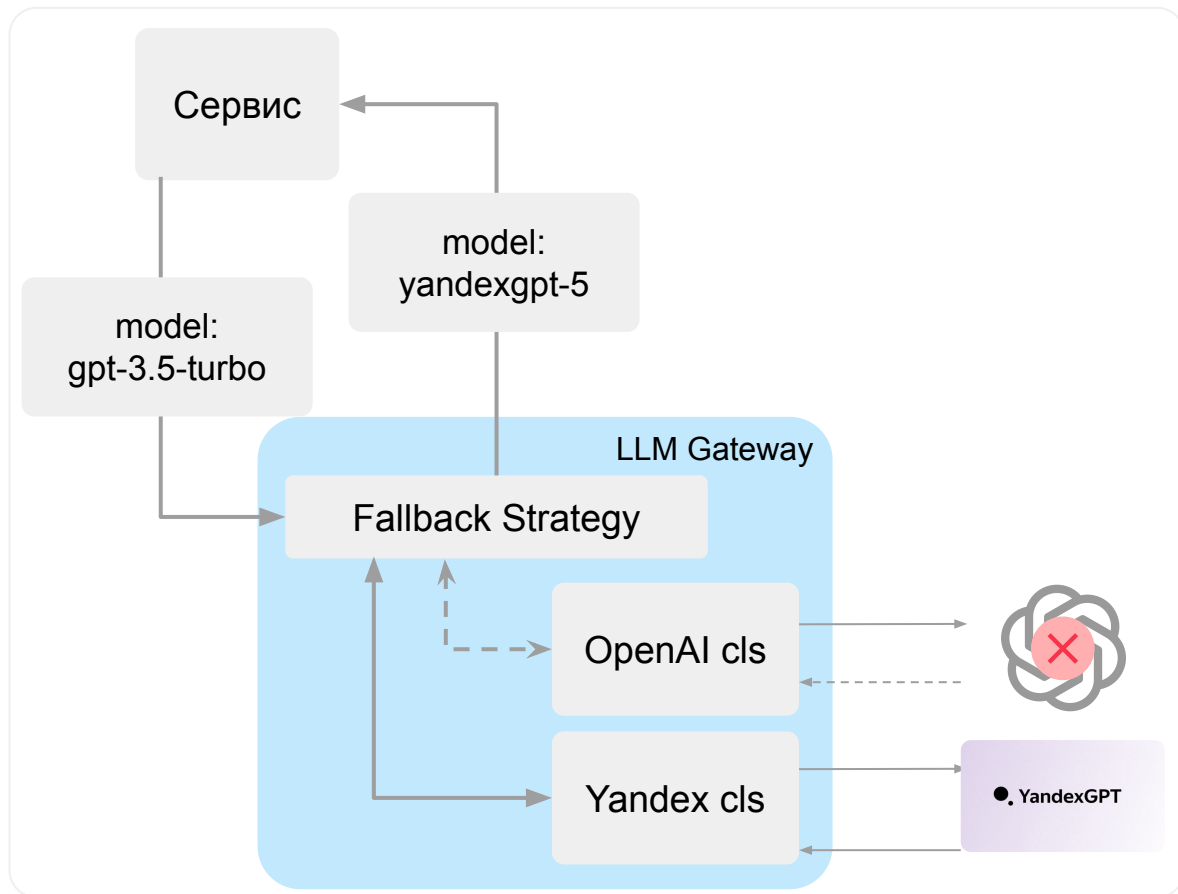
Fallback



- список моделей в конфигурации
- стратегии обхода



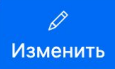
- снижение качества
- МОЖЕТ ЧТО-ТО СЛОМАТЬ
- промпт не подходит






Prompt Management


- CRUD
- версионирование
- разделение на команды
- возможность указать в запросе

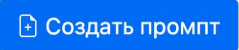
```
• {  
  "meta": {  
    "promptName": str | None  
    ...  
  },  
  ...  
}
```

tickets_reply_dial_v1  Изменить

Команда online-marketing  

Название * Ответственная команда *  Создать

Текст промпта *  Отмена

 Создать промпт

Рабочее место Prompt-инженера

Prompt system

Ты очень плохой эксперт в географии. Но я дам тебе 100\$ за правильный ответ. Это очень важно!

Prompt user

Какая столица страны {country}?

Model Parameters

Temperature 1

Model llama

Max tokens 10

Force json

2. Preview and test

Load Test sets

Run all

Input parameters

Нарния

Results

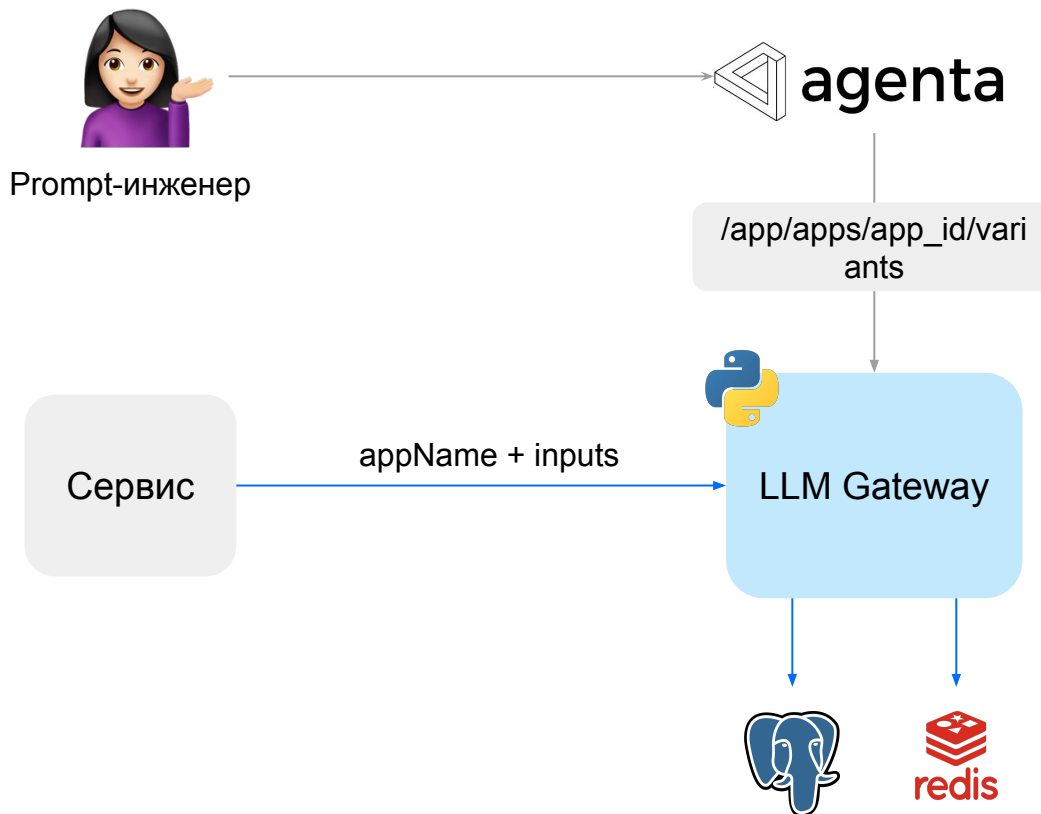
+ Add to Test Set

▶ Run Test

Не существует столицы Нарнии

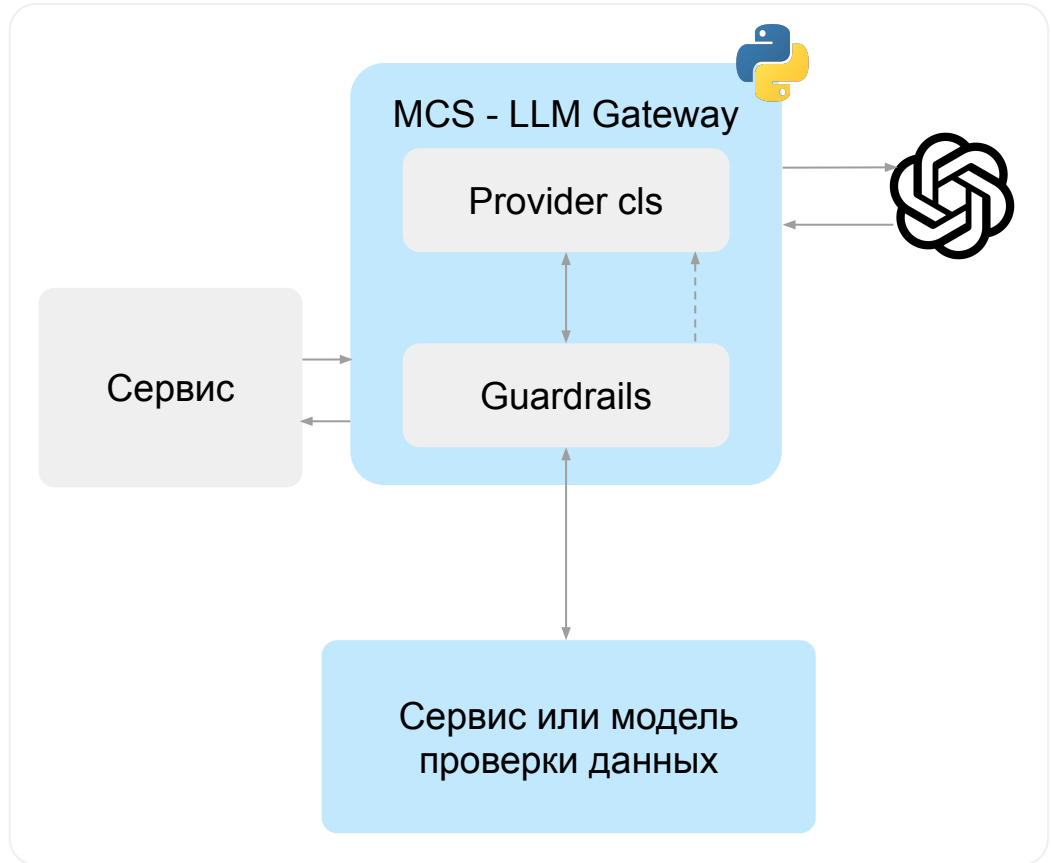
Tokens: 57 Cost: \$0.000684 Latency: 629ms

Интеграция с LLM Gateway



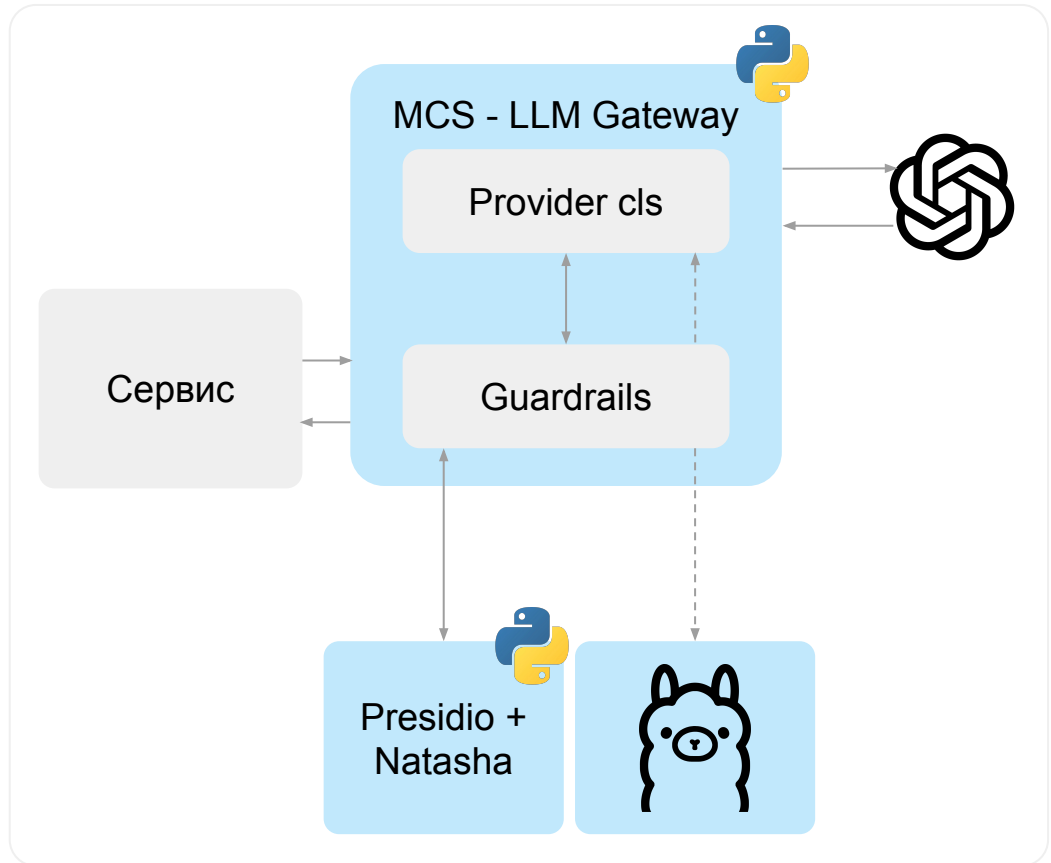
Безопасность данных

- определение: номер телефона, email, секреты
- маскирование, анонимизация, блокирование
- примитивно: на регулярках



Безопасность данных

- NER, регулярки
- встроенный анонимизатор
- ограниченное покрытие сущностей и языков "из коробки"
- зависимость от внешних NLP-библиотек
- требует кастомизации



Self-hosted LLM

Манифест

```
apiVersion: apps/v1
```

```
kind: Deployment
```

```
...
```

```
spec:
```

```
  nodeSelector:
```

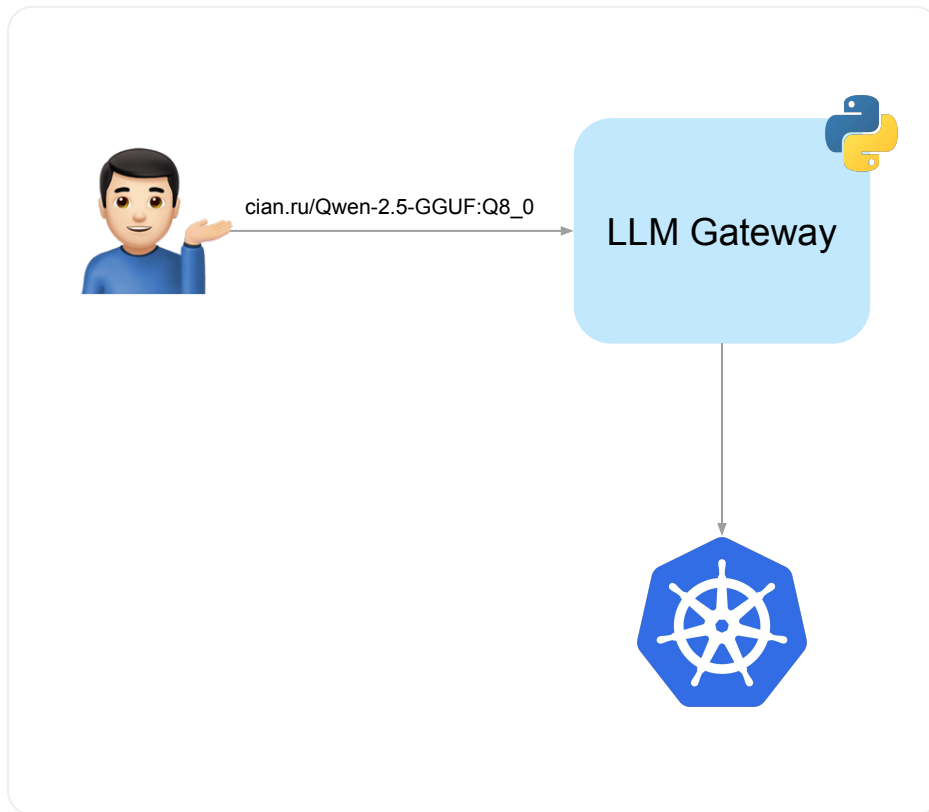
```
    gpuId: gpu-a100-rsch4
```

```
    ...
```

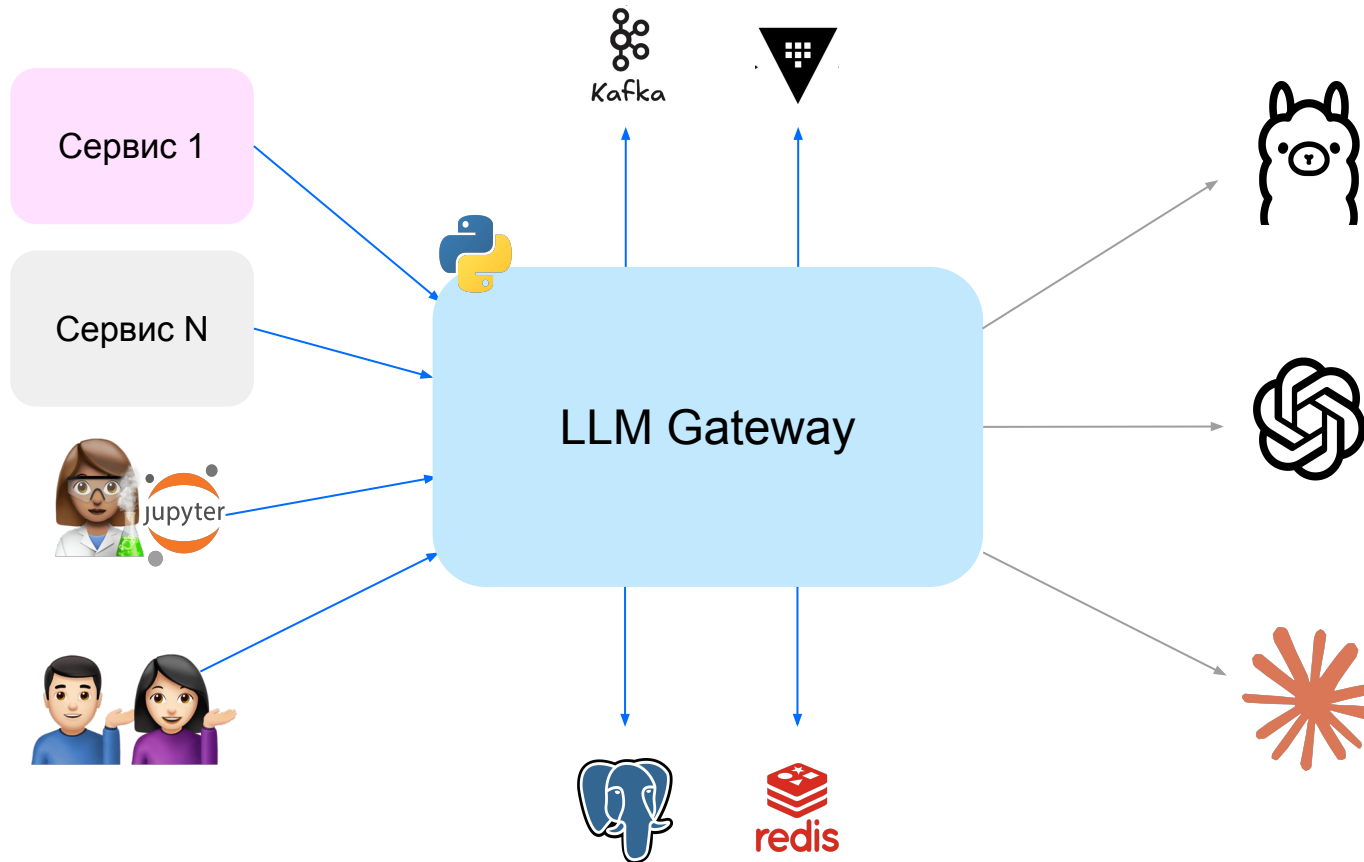
```
  env:
```

```
    - name: LLM_MODEL
```

```
      value: "cian.ru/Qwen..GGUF:Q8_0"
```



LLM Gateway



Облачные решения



- быстро
- минимум разработки
- уже есть наиболее популярные llm



- более высокая стоимость
- не для всех моделей плата за токен
- ограничены моделями и функционалом провайдера
- остаются вопросы с безопасностью и ключами

Облачные решения

Централизация
и унификация



Гибкость и расширяемость



Управление промптами



Мониторинг и аналитика



Прокси до провайдеров



Безопасность данных



Безопасность и контроль
доступа



Лимиты и квоты



Интеграция своих моделей



Open Source - LiteLLM



- free self-hosted
- много возможностей фич из коробки
- поддержка популярных llm
- низкая задержка
- поддерживается в некоторых ai-фреймворках



- часть функций только в enterprise
- нет поддержки популярных российских моделей
- сложности с масштабированием и отказоустойчивостью
- потребует частичного допиливания под вас
- функционал может быть избыточным

Open Source - LiteLLM

Централизация
и унификация



Гибкость и расширяемость



Управление промптами



Мониторинг и аналитика



Прокси до провайдеров



Безопасность данных



Безопасность и контроль
доступа



Лимиты и квоты



Интеграция своих моделей



1

Контроль над архитектурой.

Независимость от сторонних вендоров.

2

Кастомизация и гибкость без лишних и ненужных функций.

3

Масштабируемость по мере роста бизнеса.

4

Снижение затрат на доработки и кастомизацию.



Не ограничивайте себя только стандартными решениями

инвестируйте в собственный LLM Gateway
и управляйте ИИ по своим правилам



Телеграм: @romaneverlate

