

ВАШЕЙ КОМПАНИИ LLM НЕ НУЖЕН?

МИЧИЛ ЕГОРОВ

РУКОВОДИТЕЛЬ РАЗРАБОТКИ, X5 TECH

MICHIL.EGOROV@X5.RU

PiterPy 2024

**AI
Run**

о себе



ЕГОРОВ МИЧИЛ
TECH LEAD

программировал в
яндексе, старлайне
руководжу в x5 tech
менторю в itmo

airun: статистика

30+ человек
состоит в команде

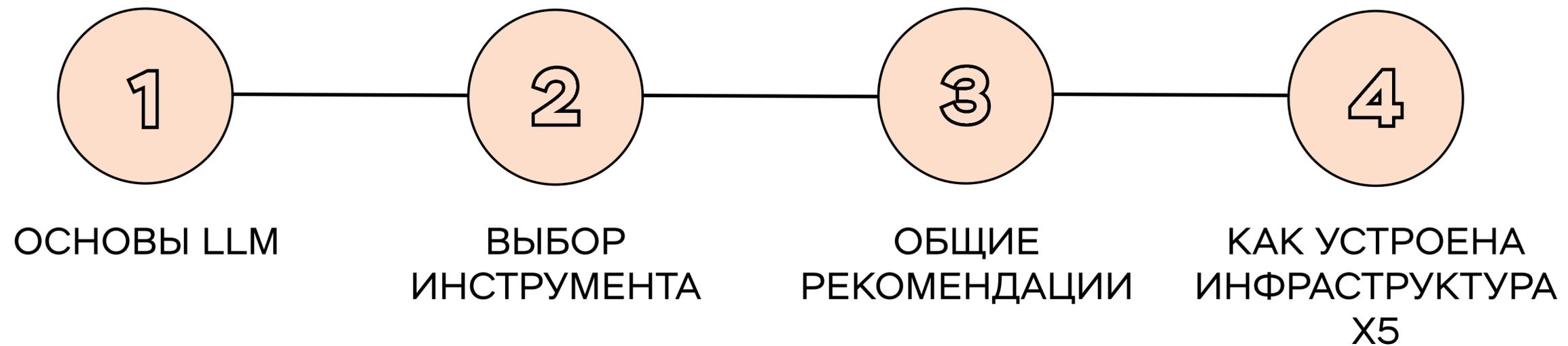
4 чат-бота
на базе LLM

5+ сервисов
на базе LLM

500k+
запросов на генерацию

10+ LLM моделей
протестировано

план презентации



1

ОСНОВЫ LLM

Что такое LLM?

ChatGPT Авто ▾ ↑ Поделиться MI

напиши мне что такое LLM

 LLM (Large Language Model) — это тип модели машинного обучения, основанный на архитектуре трансформеров, который обучается на больших объемах текстовых данных для обработки естественного языка. Эти модели способны понимать и генерировать текст, выполняя различные задачи: от ответов на вопросы и написания статей до перевода и анализа текста.

Основные характеристики LLM:

- 1. Большое количество параметров** — модели могут содержать миллиарды или триллионы параметров, что позволяет им запоминать и обрабатывать огромное количество информации.
- 2. Генерация и понимание текста** — LLM обучаются на корпусах текста с целью предсказания следующего слова или предложения, что позволяет им создавать осмысленные и связные тексты.
- 3. Применение в различных задачах** — такие модели могут быть использованы в чат-ботах, системах рекомендаций, анализе настроений, переводе текстов и других областях, связанных с языковыми данными.

Примеры LLM — это GPT (Generative Pretrained Transformer), BERT, и другие крупные модели, разработанные различными исследовательскими группами и компаниями, такими

 Сообщить ChatGPT ↑

ChatGPT может допускать ошибки. Рекомендуем проверять важную информацию. ?

что такое llm?

llm - это искусственная
нейронная сеть, обученная на
огромных текстовых данных

применение ИИ в x5

ПОИСК ПО
БАЗАМ
ЗНАНИЙ

ЧАТ-БОТЫ
ДЛЯ
ПОДДЕРЖКИ

ОБРАБОТКА
ОТЗЫВОВ
КЛИЕНТОВ

СУФЛЕРЫ
СОТРУДНИКОВ
ПОДДЕРЖКИ

структура llm



веса модели



скрипт для
инференса

закрытые решения

какие есть риски?

- ☑️ конфиденциальность данных
- ☑️ зависимость от внешних систем
- ☑️ затраты на долгосрочной перспективе
- ☑️ ограничения на использование

2

ВЫБОР ИНСТРУМЕНТА

КЛЮЧЕВЫЕ АСПЕКТЫ

- ✓ latency (tps)
- ✓ пропускная способность
- ✓ качество ответов
- ✓ потребление ресурсов
 - ✓ cpu only
 - ✓ gpu ~~poor~~ regular
 - ✓ gpu rich

постановка эксперимента

- ✓ модель: mistral-7b
- ✓ cpu only: 500ram 72 cpu
- ✓ gpu regular: rtx8000
- ✓ gpu rich: a100

бекенды инференса

LLaMA⁺⁺



text-generation-inference

vLLM

LLAMA-CPP

обзор

легковесная c++ библиотека для инференса языковых моделей

- ✓ поддержка различных архитектур (x86, arm)
- ✓ возможно работать без gpu, только на cpu
- ✓ работает только с форматом gguf

применение

конвертация в gguf

```
python convert_hf_to_gguf.py \  
  --model_path $model_path \  
  --output_path path/to/output.gguf
```

запуск бекенда

```
docker run -p 8080:8080 \  
  -v $model:/models \  
  ghcr.io/ggerganov/llama.cpp:server \  
  -m models/model.gguf \  
  --host 0.0.0.0 \  
  --port 8080
```


сравнительный анализ

Технология	CPU only	GPU regular	GPU rich
llama-cpp	8-14 tps	45-51 tps	49-54 tps
tgi			
vllm			

apple silicon

Чип	CPU only
M1 ultra	37 tps
M2 ultra	41 tps



TEXT GENERATION INFERENCE

обзор

проект от hugging face

- ☑️ оптимизирован для работы на gpu
- ☑️ поддерживает распределенный инференс
- ☑️ поддерживает динамический батчинг

применение

запуск бекенда

```
docker run \  
  --gpus all \  
  -p 8080:80 \  
  -v $volume:/data \  
  ghcr.io/huggingface/text-generation-inference:2.3.0 \  
  --model-id $model
```

сравнительный анализ

Технология	CPU Only	GPU regular	GPU rich
llama-cpp	8-14 tps	45-51 tps	49-54 tps
tgi	N/A	525-545 tps	1700-1850 tps
vllm			

VLLM

обзор

проект от berkeley us

- ✓ использует pagedattention
- ✓ оптимизирован для мощных gpu
- ✓ поддерживает распределенный инференс
- ✓ поддерживает динамический батчинг

применение

запуск бекенда

```
pip install vllm
```

```
vllm serve $model
```

guided_regex

цель: избавиться от лишних символов

Привет! Я бот помогающий разговаривать с вашими данными. Выберите коллекцию и задавайте вопросы!

6:57

Привет! Я бот помогающий разговаривать с вашими данными. Выберите коллекцию и задавайте вопросы!

6:57

как ты?
6:57

как ты?
6:57

(太棒了
6:57

Оценить ▾

Отлично!
6:57

Оценить ▾

результат: время генерации **x2**

сравнительный анализ

Технология	CPU	GPU regular	GRU rich
llama-cpp	8-14 tps	45-51 tps	49-54 tps
tgi	N/A	525-545 tps	1700-1850 tps
vllm	N/A	670-768 tps	1800-2000 tps

3

ОБЩИЕ РЕКОМЕНДАЦИИ

рекомендации

прототипирование и не важна скорость работы - llama.cpp

нагрузка < 40 rps и нужны качественные и быстрые ответы - tgi

высокая нагрузка, быстрые ответы и есть мешок денег - vllm

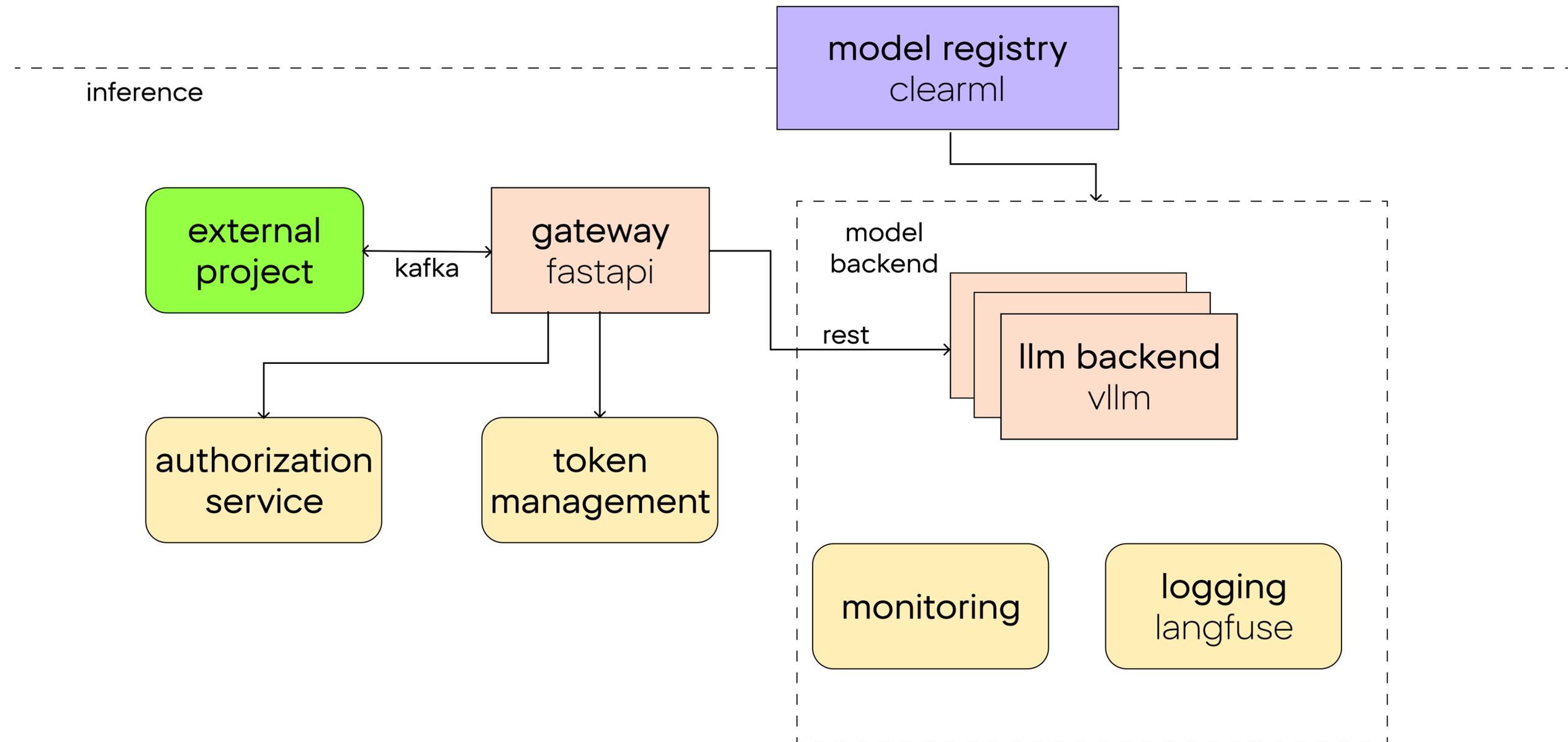
ЧТО СТОИТ УЧИТЫВАТЬ

- ☑ генерация зависит от промптов
- ☑ необходимо подбирать параметры генерации
- ☑ галлюцинации - всегда будут

4

КАК БЕКЕНД УСТРОЕН В X5

Архитектура



langfuse

- ✓ распределенный трейсинг
- ✓ версионирование промптов
- ✓ а/б тесты для промптов
- ✓ хранение датасетов

langfuse

Langfuse v2.12.1

Dashboard

Tracing

Traces

Sessions

Generations

Scores

Models

Users

Prompts

Datasets

Settings

Support

Feedback

Projects

+ New

M Michil

infocenter > Traces > 86fa164b-bb2b-4cca-8117-d18108ccd319

Trace Detail

☆ Private ^ v 🗑️

Session: b2014e2f-d0e3-8597-77dc-ddb3df37e881

Tags

SPAN **paraphrase_clarify_query**

Add score + Add to dataset

26.09.2024, 00:20:32

Latency: 3.37s

Input

```
{\n  query: "Как войти на портал TTS с личного телефона?"\n  context: "Всего найдено документов: 5 Документ 1 **Портал TTS** | | | -- | -- | | **Портал «TTS» - это система обратной связи. Портал предназначен для регистрации обращений Гостей.** | Портал размещен по адресу: <https://crm-cc.x5.ru/> | | | -- | -- | -- | | * Вход в систему TTS только через браузеры **Mozilla firefox и Google Chrome;** | | | * **Не использовать** браузер Internet Explorer для работы с системой. | **Для входа в систему** **TTS** **необходимо использовать учетную запись и пар...**"\n}
```

Output

```
{\n  paraphrased_query: "Для входа на портал TTS с личного телефона необходимо авторизоваться в VPN или Airwatch, в зависимости от устройства. Затем открыть браузер Google Chrome или Mozilla Firefox и перейти по прямой ссылке http://fr-dtw.xcloud-prod.x5.ru. Можете ли вы подтвердить, что хотите получить инструкцию по входу на портал TTS с личного телефона?"\n}
```

TRACE jaicp

4.49s

SPAN process_query

4.44s

SPAN paraphrase_history

0.57s

SPAN retrieve_context

0.47s

SPAN process_context

0.00s

SPAN process_context

0.00s

SPAN _is_query_relevant

0.02s

SPAN **paraphrase_clarify_query**

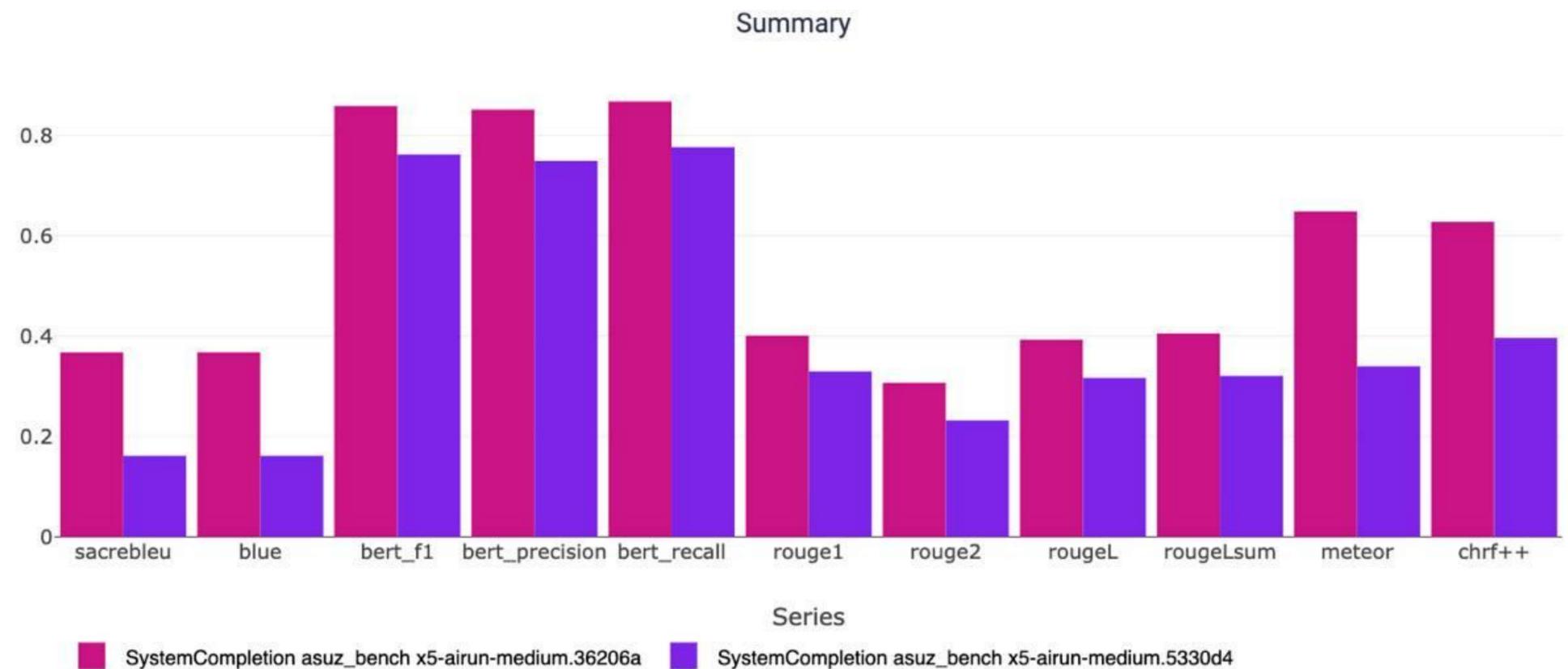
3.37s

SPAN send_answer



как мы выбираем модель

- ✓ следим за обновлениями
- ✓ прогоняем на бенчмарках
- ✓ катим в дев
- ✓ катим в прод



КАКИЕ МОДЕЛИ МЫ ТЕСТИЛИ

- ✔️ mistral-7b-saiga, mixtral-8x7b
- ✔️ llama-2-13b-code, llama-2-70b-instruct
- ✔️ suzume-llama-3, llama-3-saiga, llama-3-instruct
- ✔️ c4ai-command-r-35b, c4ai-command-r-35b-0824
- ✔️ huyalm-100b

почему несколько?

- ☑ для решения разных задач на специализированных моделях
- ☑ возможность комбинировать ответы от разных моделей
- ☑ маленькие модели - быстрая скорость ответа

ИТОГИ

ИТОГИ

- 📁 вкладывайтесь в инфраструктуру для минимизации утечки данных
- 📁 выбирайте бекенд для инференса в зависимости от задачи
- 📁 IIm можно использовать без gpi
- 📁 используйте специализированные сервисы для мониторинга IIm

ВАШИ ВОПРОСЫ

МИЧИЛ ЕГОРОВ

РУКОВОДИТЕЛЬ РАЗРАБОТКИ, X5 TECH

ДЛЯ СВЯЗИ

MICHIL.EGOROV@X5.RU

@LOALKOTA

