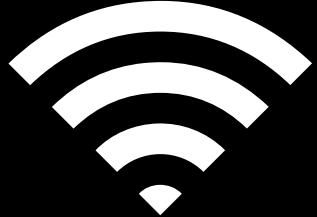
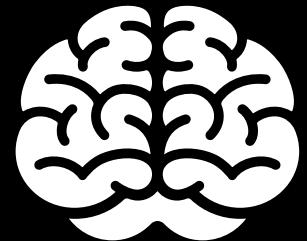


цифровой
прорыв ↑

сезон: ии



Команда Psyper_v3



Разработка QnA чат-бота
на основе базы знаний



Министерство
экономического развития
 Российской Федерации



На
хакатон!



Состав команды Psyper_v3



Семён Семёнов

✉ ruster478@gmail.com ↗ @lux_astrum



MACHINE LEARNING

NLP

COMPUTER VISION

ЕЩЁ

Lead, ml engineer



Антон Барановский

✉ 89021397725@mail.ru ↗ @anton2010000



СИСТЕМНАЯ АНАЛИТИКА

ТЕОРИЯ МАШИННОГО ОБУЧЕНИЯ

Data analyst, ml engineer



Никита Шабан

✉ nikich11let@gmail.com ↗ @karthusclear



ТЕОРИЯ МАШИННОГО ОБУЧЕНИЯ

PYTHON

PYTORCH

ЕЩЁ

ML engineer



Никита Кувшинов

✉ KuvshinovND@yandex.ru ↗ @NDK619



PYTHON

DJANGO

АВТОМАТИЗАЦИЯ ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССОВ

Backend



Денис Пушко

✉ pushkodenis@gmail.com



Frontend



Министерство
экономического развития
Российской Федерации



цифровой
прорыв

сезон: ИИ

Проблематика

Нехватка
квалифицированных
специалистов по
управлению персоналом

Большое количество
документов для
ознакомления

Большое количество
предприятий

Обучение специалистов
требует много времени

Частое изменение и
актуализация нормативных
документов

Постоянный найм новых
сотрудников



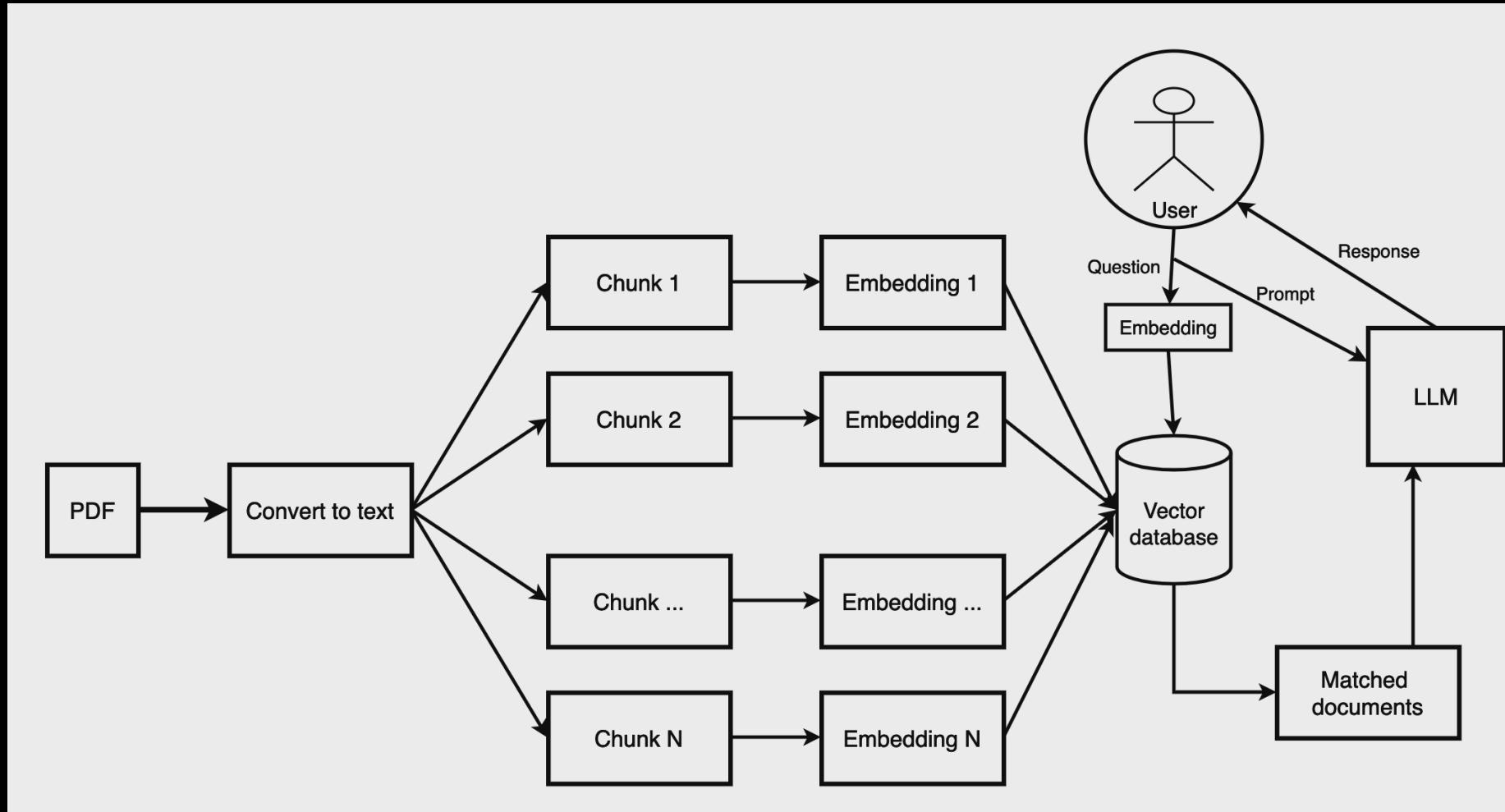
Министерство
экономического развития
 Российской Федерации



цифровой
прорыв

сезон: ИИ

Принцип работы чат-бота



Министерство
экономического развития
 Российской Федерации



цифровой
прорыв

сезон: ИИ

Обоснование выбора модели

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General						
MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code						
HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math						
GSMBK (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning						
ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use						
BFCL	76.1	-	60.4	84.8	-	85.9
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context						
ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual						
Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4



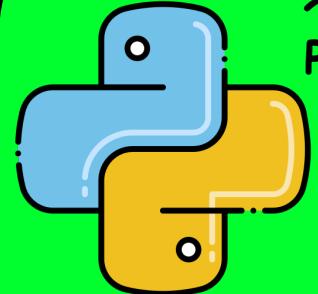
Министерство
экономического развития
 Российской Федерации



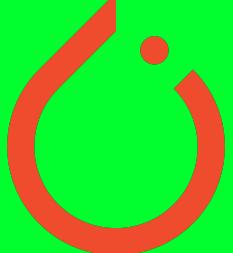
цифровой
прорыв

сезон: ИИ

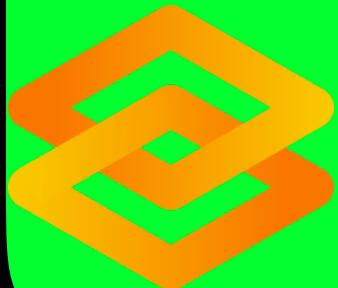
Технологии



Язык программирования
Python



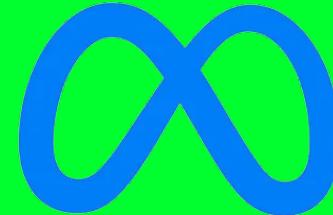
Фреймворк глубокого
обучения PyTorch



Фреймворк для создания
веб-приложений машинного
обучения Gradio



Векторная база данных
Faiss



Llm Llama 3.1 8b_q4



Министерство
экономического развития
Российской Федерации



цифровой
прорыв

сезон: ИИ

Ключевые особенности нашего решения

- Удобный и понятный интерфейс чат-бота → Скорость обработки вопросов не заставляет пользователя долго ждать
- Быстрое развертывания системы и удобство обновления базы знаний → Оптимизированное по потреблению вычислительных ресурсов решение
- Масштабируемость системы → Задел на дальнейшее развитие системы



Министерство
экономического развития
Российской Федерации



цифровой
прорыв ↑

сезон: ИИ

User interface

The screenshot shows a user interface for a QnA chat-bot. At the top right are 'Чат' and 'Документы' buttons. The main title is 'QnA чат-бот на основе базы знаний КУ РЖД'. On the left is a large red and blue circular logo. A white callout box contains text about the challenges faced by new employees at RZhD regarding document overload and lack of time to understand benefits. The bottom features the RZhD Corporate University logo, a 'цифровой прорыв' (Digital Breakthrough) banner, and the RZhD logo.

КОРПОРАТИВНЫЙ УНИВЕРСИТЕТ РЖД

цифровой прорыв

сезон: или

РЖД

КОРПОРАТИВНЫЙ УНИВЕРСИТЕТ РЖД



Министерство
экономического развития
 Российской Федерации



цифровой
прорыв

сезон: или

User interface

The screenshot shows a web-based chatbot interface. At the top left is the logo of the 'КОРПОРАТИВНЫЙ УНИВЕРСИТЕТ РЖД' (RZhD Corporate University). At the top right is a link labeled 'Документы' (Documents). The main title 'Чат-бот корпоративного университета РЖД' (Chatbot of the RZhD Corporate University) is centered above a text block stating: 'Чат-бот может ответить на ваши вопросы по нормативным документов из внутренней базы знаний корпоративного университета РЖД. База документов, по которой осуществляется поиск, была получена на основе следующего ресурса: [РЖД_Документы | Компания](#)'. Below this is a large white input area labeled 'Chatbot'. At the bottom of this area are two buttons: 'Удалить предыдущий вопрос/ответ' (Delete previous question/answer) and 'Очистить весь чат' (Clear entire chat). A text input field contains the placeholder 'Какой у вас вопрос?' (What's your question?). To its right is an orange 'Отправить' (Send) button. Below the input field are three examples: 'Сколько длится ежегодный основной отпуск?' (How long is the annual main vacation?), 'Кто такой представитель работодателя?' (Who is the employer's representative?), and 'На основе какой статьи трудового кодекса Российской Федерации исчисляется размер среднечасового заработка?' (Based on which article of the Labor Code of the Russian Federation is the size of the average hourly wage calculated?). At the very bottom are links: 'Использовать через API' (Use via API) and 'Создано с помощью Gradio' (Created with Gradio).

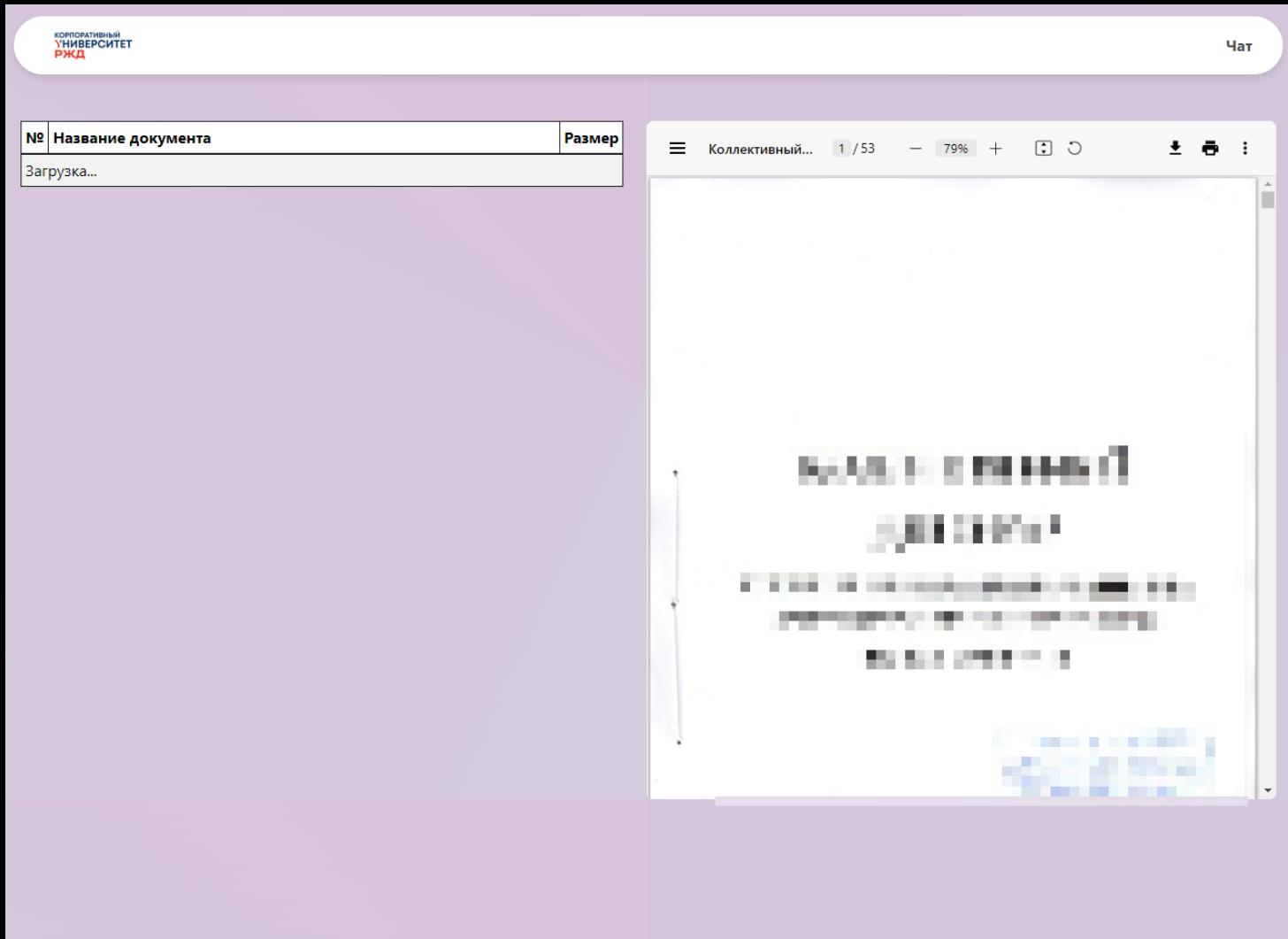


Министерство
экономического развития
Российской Федерации



цифровой
прорыв ↑
сезон: ИИ

User interface



Министерство
экономического развития
Российской Федерации



цифровой
прорыв

сезон: ИИ

Стоимость инференса

Yandex cloud

1xNvidia T4, 4vCPU, 32Гб RAM
49305,6 руб.



Yandex Cloud

Selectel

1xNvidia T4, 4vCPU, 32Гб RAM
38277,70 руб.



Министерство
экономического развития
Российской Федерации



цифровой
прорыв

сезон: ИИ

Планы на развитие

- Добавить возможность загрузки и удаления документов из веб-приложения → Добавить авторизацию пользователя
- Внедрение AirFlow для автоматического сбора и обновления документов → Сделать микросервисную архитектуру через docker контейнеры
- Отказаться от Gradio чата в пользу самописной версии → Интеграция в платформу КУ РЖД



Министерство
экономического развития
Российской Федерации



цифровой
прорыв ↑

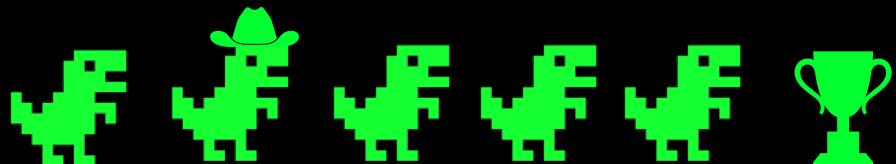
сезон: ИИ

цифровой ↑
прорыв

сезон: ИИ



Команда: Psyper_v3



Министерство
экономического развития
Российской Федерации

