Название программы: Цифровое моделирование и суперкомпьютерные технологии

Название команды: MAI’s Hive

Team lead (ФИО, tg): Мишин Сергей Алексеевич, @achiona

Ментор (ФИО, tg): Лобанов Захар Олегович, @zak_l

Паспорт проекта

Система автоматического анализа электронного документа (PDF)

1. Общая информация

- **Краткое описание проекта**:
  Проект упрощает обработку PDF-документов для бухгалтеров и секретарей. Система позволяет определить тип документа на основе содержимого и проверить наличие определенных критериев: подпись, печать. Для удобства использования система представлена в виде веб-приложения. Бэкенд-часть приложения использует большую языковую модель для анализа содержимого.

- **Команда**:

  1. **Мишин Сергей Алексеевич**: Team lead
  2. **Барменков Артемий Сергеевич**: UX/UI-дизайнер
  3. **Бондаренко Виктория Арсеновна**: Технический писатель
  4. **Борзова Дарья Максимовна**: Frontend-разработчик
  5. **Иванченко Макар Дмитриевич**: Backend-разработчик
  6. **Казец Полина Олеговна**: Аналитик
  7. **Савинов Никита Олегович**: Frontend-разработчик
  8. **Сергеев Владимир Андреевич**: Data-инженер
  9. **Федоров Ярослав Артемович**: QA-инженер
  10. **Шведов Александр Иванович**: Backend-разработчик

2. Цель проекта

- **Цель проекта**:
  Разработка интерактивной системы анализа электронных документов

- **Ожидаемые результаты**:
  Интерактивная система анализа электронных документов, обеспечивающая автоматизацию процессов и повышение эффективности работы персонала

3. Задачи и процесс работы

- **К работе (только цифрами)**:
  - В процессе: 0.
  - Завершено: 5.
  - Заблокировано: 0.

4. Прогресс и результаты

- **Текущий статус проекта**:
  Проект находится на ранней стадии разработки. Нами была разработана предварительная архитектура на основе микросервисов и выбран стек разработки.

- **Достигнутые результаты по задачам**:
  Наша команда работала над анализом существующих решений и созданием различных диаграмм, отражающих концепцию проекта. Для организации эффективной работы оформлен репозиторий и трекер задач на GitHub. К защите проекта была подготовлена презентация и оформлен паспорт проекта.

- **Риски и препятствия**:
  Поскольку проект использует большую языковую модель для анализа содержимого документов, для развёртывания потребуются значительные вычислительные мощности. К тому же, система должна быть разработана с учётом больших объёмов данных и должна легко масштабироваться. Не стоит забывать про неточность системы распознавания текста — возможно, придётся искать другие решения.

5. Ресурсы и материалы проекта

- **Используемые инструменты и технологии**:
  - Фронтенд-фреймворк: React JS
  - Бэкенд-фреймворк: FastAPI
  - Связь между микросервисами: gRPC
  - Распознавание текста: Tesseract OCR
  - Анализ текста документа: Ollama, Llama 3.3
  - Анализ подписи, печати: OpenCV

- **Ссылки на внешние ресурсы**:
  - Репозиторий: https://github.com/oryce/mai-ck

  - Трекер задач: https://github.com/users/oryce/projects/1

- **Данные**

  Датасет пока не утверждён.

6. Комментарии и мысли команды

- **Комментарии**:
  Нам будет крайне интересно поработать с большой языковой моделью типа Llama, свободно доступной и отлично функционирующей на суперкомпьютерах.