Распознавание и суммаризация документов в мобильном приложении HSE App X

Индивидуальный программный проект

Выполнил студент группы #БПМИ224 Вдонин Алексей Васильевич

Принял руководитель проекта Юнаш Игорь Игоревич, сотрудник НИУ ВШЭ

Предметная область + Актуальность

- Корпоративная СЭД НИУ ВШЭ генерирует множество документов — договоры, приказы, отчёты.
- Ручное чтение может заметно замедлять согласования: поиск сведений в длинных скан-копиях затруднён.
- Рост объёма документов и переход на онлайн-архив требуют автоматизации извлечения смысла.

Цель и задачи

1. **Цель**
   Создать self-hosted сервис, который принимает документ и за адекватное время возвращает суммаризацию.
2. **Собрать и обработать обучающий датасет**
3. **Обучить сильную модель суммаризации**
4. **Разработать веб-интерфейс**
5. **Дистиллировать модель для CPU-инференса**
6. **Провести экспериментальную оценку**

Анализ существующих решений / подходов

SaaS

Microsoft Syntex, AWS Textract + Bedrock — высокое качество, но нет офлайн-режима, данные уходят в облако.

Open-source

«Tesseract + BART/T5» — конфиденциально, но требует тонкой настройки и русскоязычного корпуса.

Наш выбор

Self-Hosted Models — контроль данных, оптимизация под CPU.

Функциональные требования

Вход

.pdf, .doc, .docx, .txt либо вставленный текст.

Выход

Суммаризация предоставленного документа.

Ограничения

Быстрый инференс на CPU-only сервере

Выбор методов и план эксперимента

OCR

pdf2image + pytesseract (rus+eng).

Teacher

ru-mbart-large-summ, дообучение на собранном датасете.

Метрики

BERTScore F1 (основная), CE-Loss.

Эксперимент

Перебор конфигураций mBART, mT5; логирование W&B.

Сбор датасета

Сбор данных

2 059 публичных документов НИУ ВШЭ парсером requests + BeautifulSoup.

Обработка

Подготовка — 8 356 чанков ⩽ 1 000 MBART-токенов.

Генерация таргетов

Таргеты сгенерированы «o3-mini». Итог: 8 356 пар *{text, summary}*.

Результат

Средняя длина summary ≈ 100 токенов (12 % входа).

Обучение сильной модели

1x

NVIDIA A100 80 GB

Платформа обучения

27

GPU • ч

Время обучения

0.76

BERTScore F1

Лучший чек-пойнт **mbart2_3_1** на валидации

Графики CE-Loss и F1 показывают стабильную сходимость.

mBART превосходит mT5 по качеству и читабельности.

График BERTScore F1 от Step

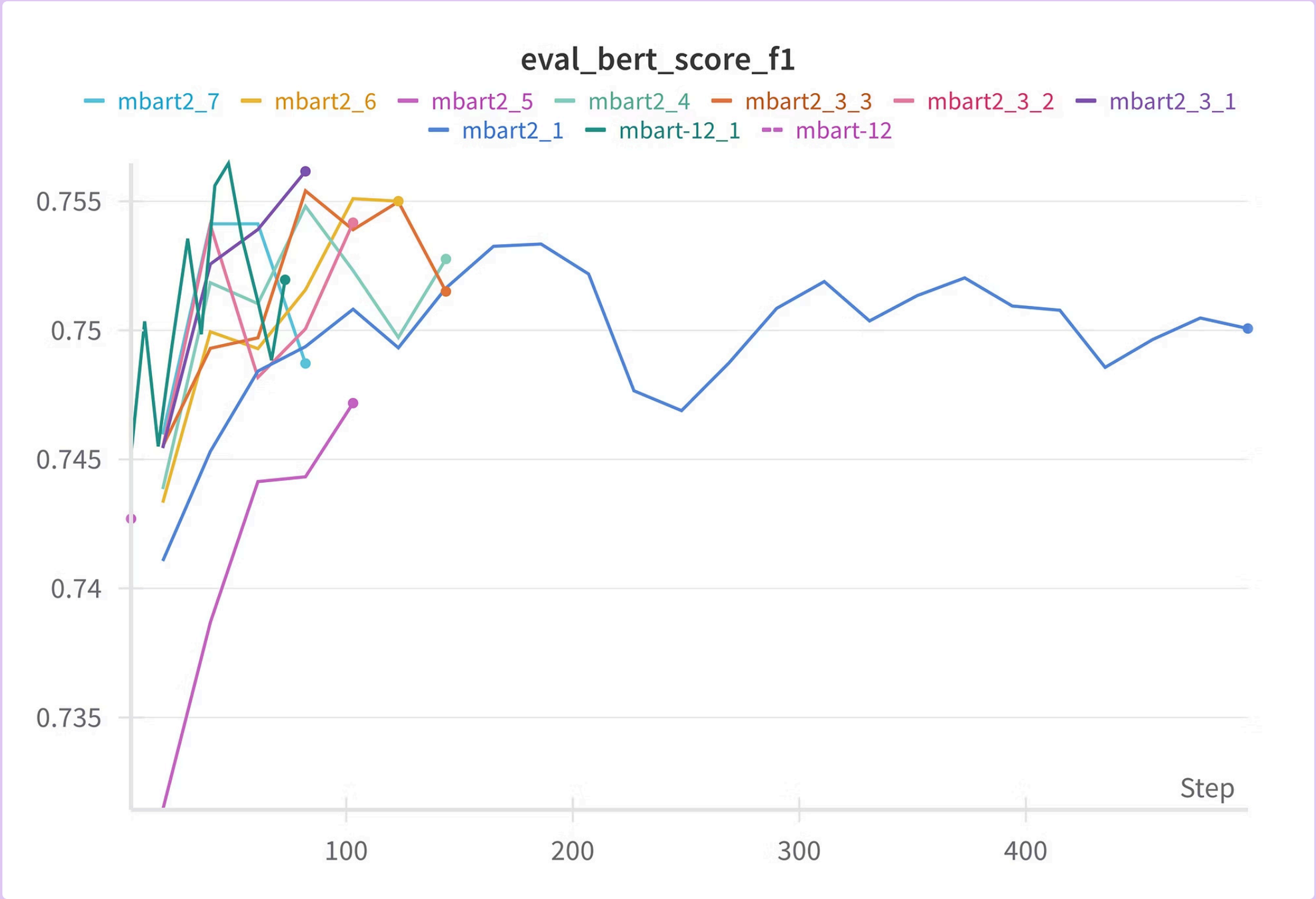
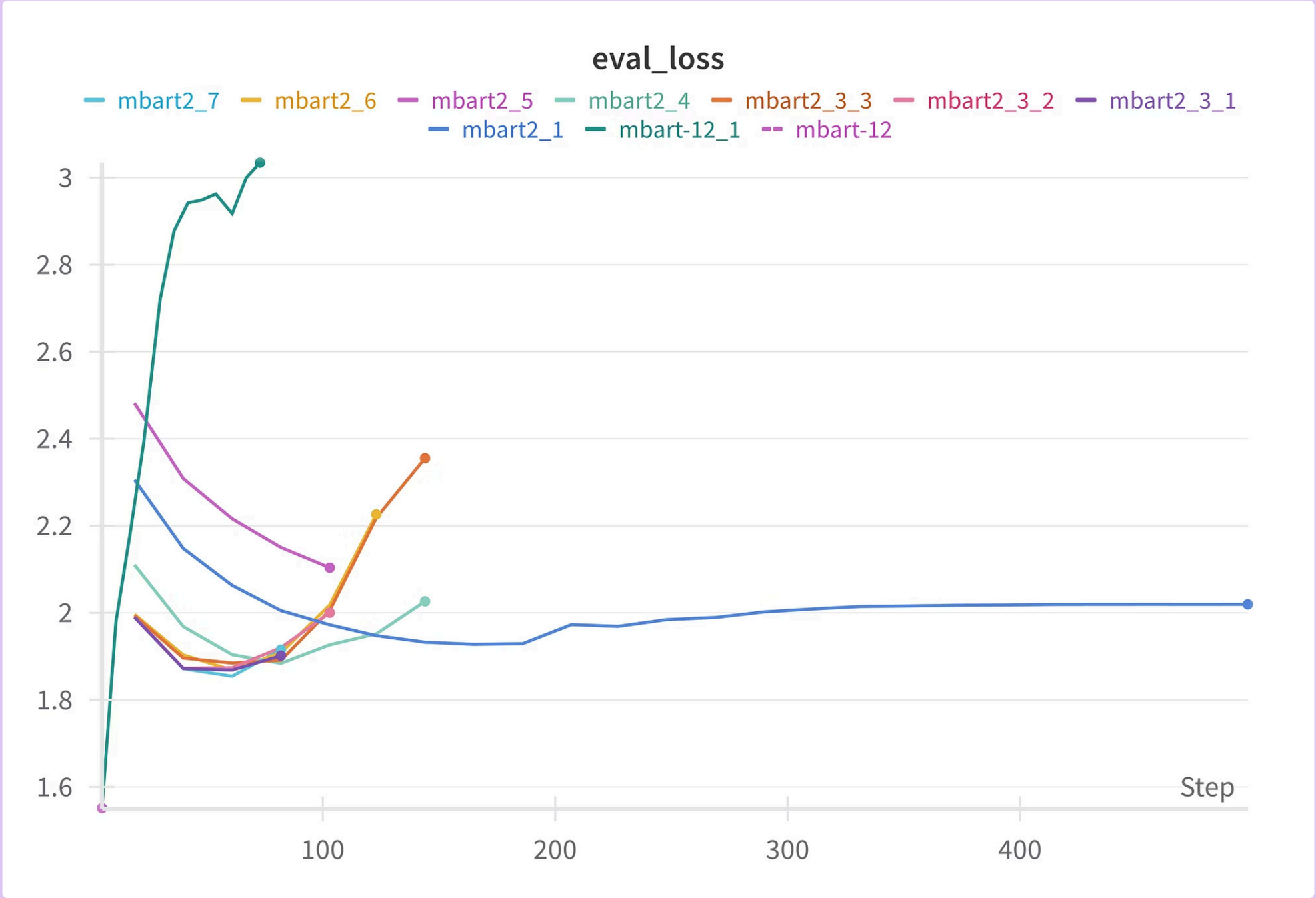
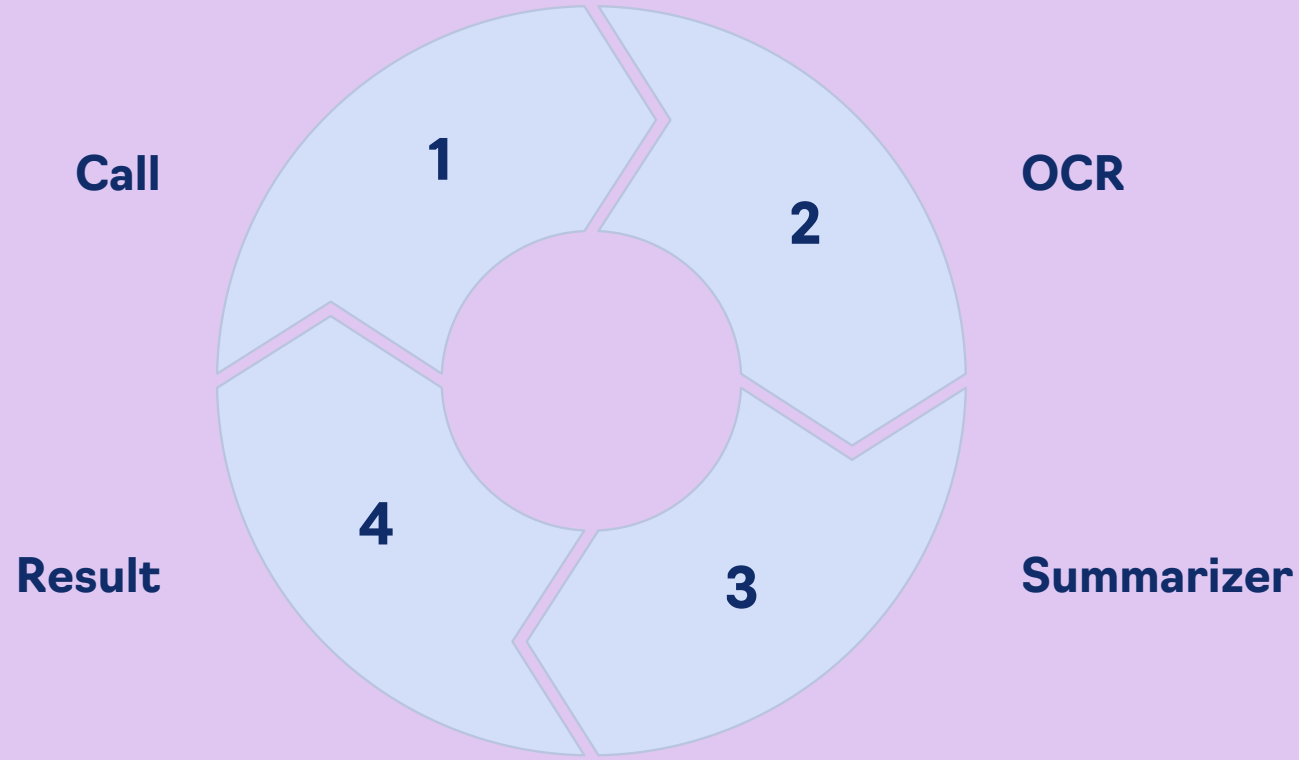


График CE Loss от Step



Архитектура приложения и веб-интерфейс



Веб-UI: Streamlit, работает на CPU-сервере; функции: загрузка файла, прогресс-бар OCR, сравнение teacher/student

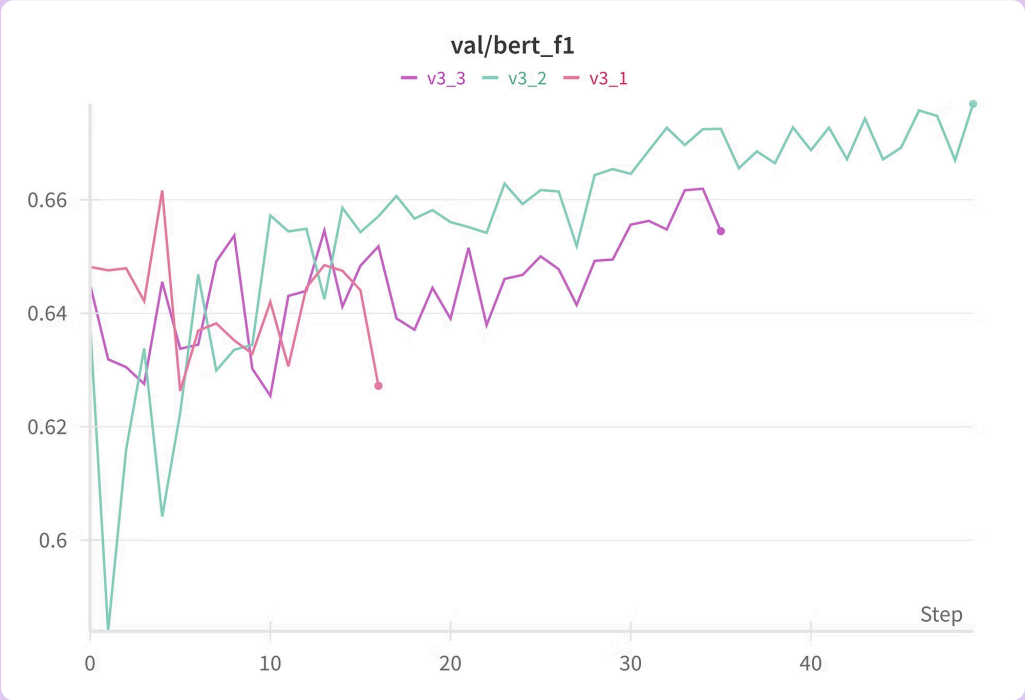
Легкая интеграция в существующую систему СЭД

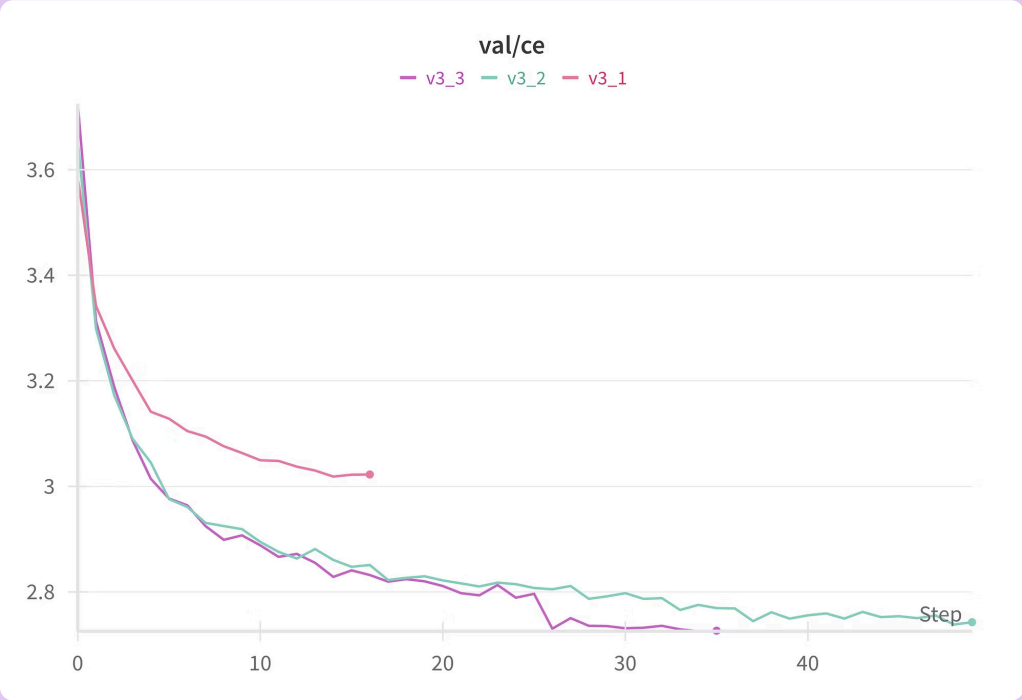
Дистилляция

Платформа	1x NVIDIA A100 80 GB, 50 GPU • ч
Метод	soft-label KD (T = 2.0) по распределению teacher
Student 1	LSTM + MHA — F1 0.69, 4.66 с, ускорение ×2.4
Student 2	чистый LSTM — F1 0.68, 2.23 с, ускорение ×5.0

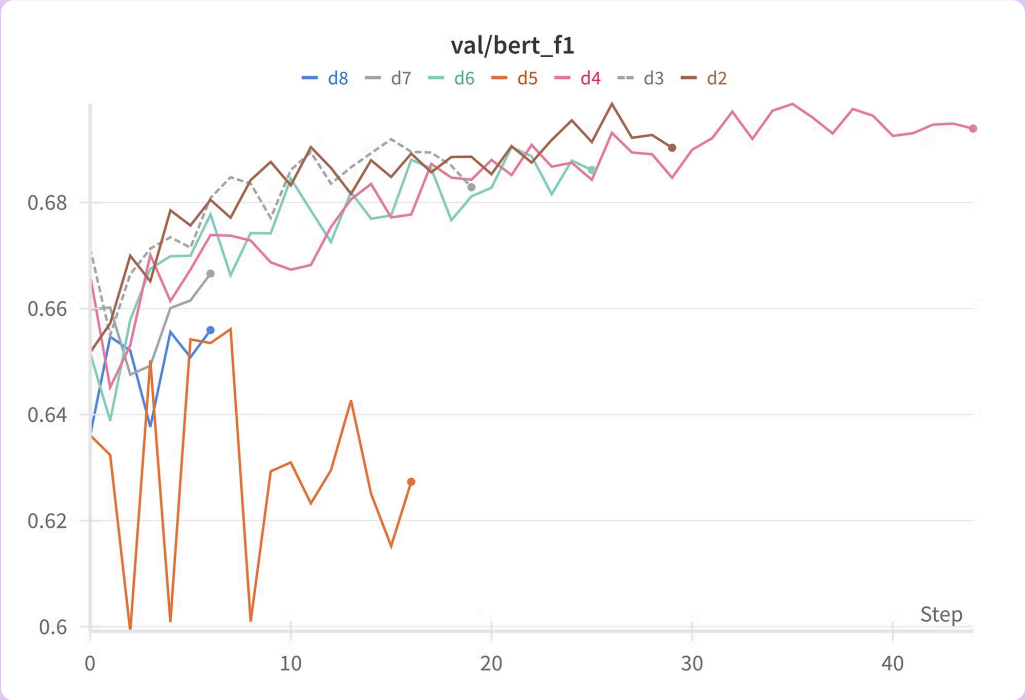
Итог: пакет из трёх моделей (teacher + 2 студента) покрывает сценарии «качество <—> скорость».

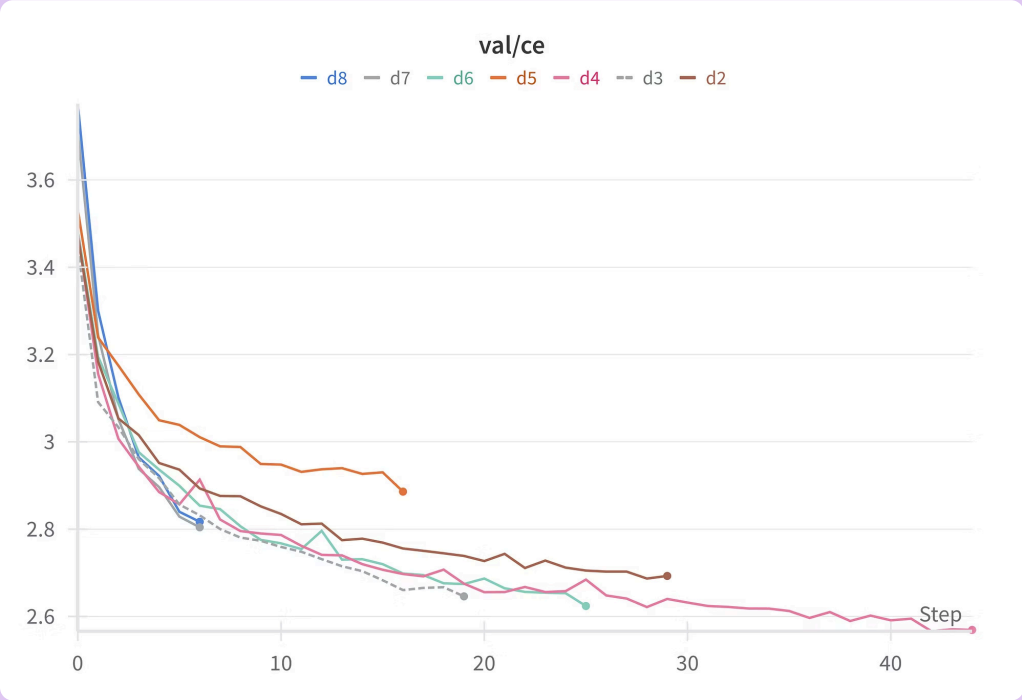
Графики обучения LSTM



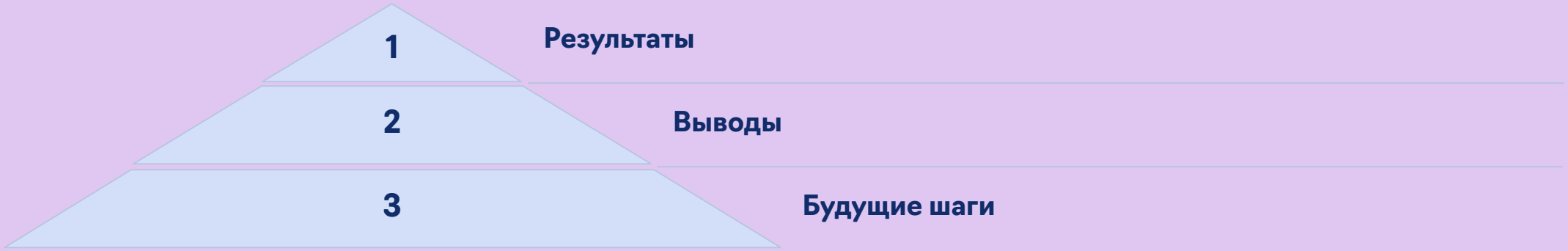


Графики обучения LSTM + MHA





Результаты, выводы, дальнейшая работа, источники



Результаты:

- Полный цикл «OCR + Summarize» реализован как self-hosted сервис.
- Teacher: F1 0.76; студенты ускоряют инференс до 5× на CPU.
- Streamlit-UI, легкая интеграция.

Выводы: сервис готов к интеграции, обеспечивает конфиденциальность и потенциальную экономию времени сотрудников.

Будущие шаги:

1. Расширить датасет (многоязычные документы, новые типы).
2. Гибридная экстрактивно-абстрактивная модель.
3. Авто-выбор модели по размеру входа и SLA.

Спасибо за внимание!

Буду рад ответить на Ваши вопросы