ТЕХНИЧЕСКОЕ ЗАДАНИЕ (ТЗ) 2

🎯 Общие правила для обоих кейсов

- У каждой задачи есть baseline (Tesseract для OCR, rule-based шаблоны для маркетинга). Цель — превзойти baseline.
- Для оценки решений используется открытый тестовый набор + скрытый тестовый набор (чтобы исключить подгонку).
- Итоговый балл = сумма по технической части (70) + презентации (30).

Дается два кейса на выбор. Необходимо выбрать **один**.

📍 КЕЙС 2. OCR 2.0 ДЛЯ БАНКОВСКИХ ДОКУМЕНТОВ

1. Общие сведения

Разработать интеллектуальную OCR-систему нового поколения для обработки банковских документов (чеки, договора, выписки), которая работает лучше Tesseract.

2. Подробное описание задачи

Классические OCR плохо справляются с шумными документами, не понимают структуру и затрудняют извлечение нужных данных.
Необходимо:

- Считывать текст даже с плохих сканов.
- Извлекать ключевые поля.
- Строить JSON-структуру.
- Проверять корректность данных.

3. Ожидаемый результат

- MVP: загрузка документа → извлечение текста и ключевых данных.
- Формат вывода: JSON с полями.
- Демо на реальных сканах.

4. Требования

- Vision Transformers (Donut, LayoutLMv3).
- OCR: PaddleOCR, TrOCR.
- LLM для пост-обработки текста.
- Python, Streamlit/Gradio для демо

5. Стек технологий

- Vision Transformers (Donut, LayoutLMv3).
- OCR: PaddleOCR, TrOCR.
- LLM для пост-обработки текста.
- Python, Streamlit/Gradio для демо.

6. Структура презентации

- Проблема OCR в банке.
- Архитектура решения (Vision Transformer + LLM).
- Демонстрация на сканах.
- Метрики качества.
- Потенциал внедрения (KYC, кредиты, архивы).

7. Метрики и критерии оценивания

Категория	**Критерии**	**Комментарии**	**Баллы**
Техническая часть (70 баллов)	Точность OCR	- CER (Character Error Rate). - WER (Word Error Rate). - Normalized Levenshtein Distance.	25
	Извлечение данных	- Field-level Accuracy (правильность каждого поля). - F1-score для каждого поля. - Exact Match per Document (% документов, извлечённых полностью верно).	25
	Работа с шумом	CER/WER на noisy subset.	10
	Структурированность	- JSON Validity (% корректных JSON). - Schema Consistency (% JSON с обязательными ключами).	10
Презентация (30 баллов)	Понятность и структура подачи		10
	Демо на реальных сканах		10
	Обоснование		10

	технологий		

Дедлайн: до 23:59, 14 сентября (GMT+5)
ссылка https://forms.gle/Swj5t73Sn9ht5TBL8
