**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

* Поддержка русского, казахского и английского языков.
* Работа с шумными/искривлёнными сканами.
* Структурированный JSON-вывод.
* Валидация данных (даты, IBAN, суммы).

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
* Демонстрация на сканах.
* Метрики качества.
* Потенциал внедрения (KYC, кредиты, архивы).

7. Метрики и критерии оценивания

Техническая часть (70 баллов):

1. Точность OCR (25)

* CER (Character Error Rate).
* WER (Word Error Rate).
* Normalized Levenshtein Distance.

1. Извлечение данных (25)

* Field-level Accuracy (правильность каждого поля).
* F1-score для каждого поля.
* Exact Match per Document (% документов, извлечённых полностью верно).

1. Работа с шумом (10)

* CER/WER на noisy subset.

1. Структурированность (10)

* JSON Validity (% корректных JSON).
* Schema Consistency (% JSON с обязательными ключами).

Презентация (30 баллов):

* Понятность и структура подачи (10).
* Демо на реальных сканах (10).
* Обоснование технологий (10).

C:\Users\Yerasyl.Akhmer\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\65C64651.tmp

**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).