Техническое задание (краткий бриф)

Повышение точности ИИ-модели распознавания лабораторных анализов в медицинском приложении

Заказчик: ТОО "iBEC Systems" · **Ответственное лицо:** Уважанов Саят

---

1. Контекст

У Заказчика есть действующее мобильное и веб-приложение: личный кабинет, где пользователь хранит результаты своих анализов. Пользователь загружает бланк лаборатории (PDF, фото или скан). Система извлекает из него показатели, сопоставляет их со справочником и показывает динамику. Приложение работает в продакшене, справочник содержит более 500 показателей с нормами.

Стек: серверная часть на PHP (Laravel), PostgreSQL, обработка через асинхронную очередь, распознавание сейчас построено на LLM. Перед сохранением пользователь проверяет распознанные значения на отдельном экране.

2. Проблема

Бланки лабораторий Казахстана сильно различаются: структура, названия показателей, единицы измерения, способ записи норм. Из-за этого модель ошибается в нескольких типичных случаях:

- путает показатели с похожими названиями, например процентное и абсолютное значение (Базофилы % и Базофилы абс.);
- неверно определяет или теряет единицу измерения;
- на фото и сканах ошибается чаще, чем на текстовых PDF: перекос, тени, несколько страниц;
- плохо разбирает бланки без норм и с качественными результатами: мазки, ПЦР, «не обнаружено», «8–12 в п/зр», «+», «++»;
- принимает за анализ посторонние изображения, например скриншоты.

В итоге пользователь вручную правит результаты, а часть показателей теряется.

3. Цель

Сократить долю строк, которые пользователь правит вручную, и сделать качество распознавания измеримым и воспроизводимым.

4. Объём работ

4.1. Аудит и тестовый набор

- Проанализировать текущий конвейер распознавания и типовые ошибки.
- Собрать размеченный эталонный набор из обезличенных бланков. В нём должны быть разные лаборатории и форматы: PDF, фото, сканы, мазки, ПЦР.
- Сделать скрипт оценки, который считает метрики из раздела 6 на этом наборе.

4.2. Улучшение модели

- Повысить точность извлечения полей: название, значение, единица, норма, отметка лаборатории о выходе за норму, биоматериал.
- Улучшить обработку изображений: предобработка, OCR, многостраничные документы.
- Поддержать качественные и полуколичественные результаты и бланки без норм.
- Добавить оценку уверенности по каждой строке, чтобы сомнительные строки уходили пользователю на проверку.
- Добавить классификатор «анализ / не анализ» для входящих файлов.

4.3. Сопоставление со справочником

- Сопоставлять строки бланка с показателями справочника с учётом единиц измерения, биоматериала и пола.
- Различать показатели с похожими названиями по единице и контексту.
- Неопознанные строки не терять: передавать их в очередь на модерацию.

4.4. Интеграция

- Встроить решение в действующую серверную часть через существующую очередь обработки. Формат ответа должен быть совместим с текущим экраном проверки.
- Сохранить обратную совместимость: старые загрузки и API не меняются.

- Покрыть новую логику автотестами.

5. Входные и выходные данные

Вход: файл бланка (PDF, JPG, PNG, HEIC; до 20 файлов в пакете), пол и возраст пользователя, справочник показателей.

Выход (JSON): признак «это анализ», дата анализа, лаборатория и список строк. Каждая строка содержит название с бланка, идентификатор показателя из справочника, значение, единицу, норму, отметку лаборатории и уверенность.

6. Критерии приёмки

Метрики считаются на эталонном наборе, собранном в п. 4.1.

Метрика	Целевое значение
Точность извлечения значения и единицы	≥ 97% строк (текстовые PDF), ≥ 92% (фото и сканы)
Верное сопоставление со справочником	≥ 95% строк
Пропущенные строки бланка	≤ 2%
Ложное «это анализ» на постороннем изображении	≤ 1%
Время обработки одного бланка	≤ 60 с
Стоимость обработки одного бланка	не выше текущей

Целевые значения уточняются после аудита (п. 4.1) по согласованию сторон.

7. Ограничения и требования

- **Данные.** Исполнитель работает только с обезличенными данными. Персональные и медицинские данные пользователей Исполнителю не передаются и не покидают контур Заказчика.
- **Размещение.** Решения, которые работают локально или в изолированном окружении, в приоритете. Если используется внешний API, это нужно согласовать заранее и обосновать.
- **Медицинские выводы.** Модель только переносит данные с бланка. Она не ставит диагнозы и сама не вычисляет «норма / не норма».
- **Код и документация.** Код передаётся в репозиторий Заказчика с описанием архитектуры, инструкцией по запуску и инструкцией по повторной оценке на тестовом наборе.

8. Этапы и результаты

Этап	Результат	Срок
1. Аудит	Отчёт о типовых ошибках, эталонный набор, скрипт оценки, базовые метрики	2 недели
2. Разработка	Улучшенная модель и сопоставление, метрики на эталонном наборе	4–6 недель
3. Интеграция	Решение встроено в продукт, автотесты, документация	2 недели
4. Опытная эксплуатация	Проверка на реальном потоке, исправление замечаний	2 недели

9. Что предоставляет Заказчик

- Доступ к справочнику показателей и правилам сопоставления.
- Обезличенную выборку бланков разных лабораторий.
- Тестовое окружение и описание текущего API.
- Контактное лицо для оперативных вопросов.

10. Требования к Исполнителю

- Опыт в OCR и Document AI либо в извлечении данных из документов с помощью LLM.
- Будет плюсом опыт работы с медицинскими документами и обезличенными данными.
- В заявке указать предлагаемый подход, сроки, стоимость по этапам и примеры похожих проектов.