Техническое задание

R&D: локальная (on-premise) платформа распознавания речи и суммаризации встреч для казахского, русского и английского языков

Цель

Подобрать, проверить и при необходимости дообучить открытые модели так, чтобы полный цикл обработки аудио (распознавание речи, разделение по спикерам, итоговое резюме) работал полностью локально на одном вычислительном узле класса NVIDIA DGX Spark, без передачи данных во внешние облачные сервисы. Качество должно быть измерено отдельно по каждому языку, а производительность узла — рассчитана.

Требования к лицензиям

Допускаются только модели и данные с лицензиями, разрешающими коммерческое использование, модификацию и дообучение (MIT, Apache 2.0, CC-BY и аналогичные).

1. Бенчмарк распознавания речи (ASR)

- Сравнить открытые ASR-модели на публичных наборах (FLEURS, Common Voice и др.) для казахского, русского и английского.
- Эталон для сравнения — коммерческий облачный сервис транскрибации (ElevenLabs).
- Метрики: WER, CER, скорость обработки (RTF), потребление памяти.
- Сравнить качество на 8 кГц (телефония) и на 16 кГц и выше.

2. Смешанная речь и шала-казахский

- Оценить качество на речи с переключением казахского и русского внутри фразы.
- Проверить поведение системы при наличии в записи других языков.

3. Дообучение

- Провести аудит имеющихся данных и наработок.
- Если на казахском или смешанной речи качество заметно ниже эталона — оценить необходимый объём данных и сроки дообучения (fine-tune).

4. Диаризация

- Оценить качество разделения речи по спикерам (DER) на многоязычных записях.

5. LLM для суммаризации

- Подобрать оптимальную по соотношению качества и ресурсов модель с подходящей лицензией.
- Оценить качество резюме встреч на казахском и русском.

6. Всё на одном узле

- Обеспечить одновременную работу ASR, диаризации и LLM на одном устройстве DGX Spark.

7. Нагрузочное тестирование

- Определить число одновременных пользователей.
- Определить объём обрабатываемого аудио в часах в сутки и в месяц.
- Рассчитать, на сколько клиентов достаточно одного узла.

Ожидаемые результаты

- Сравнительный отчёт по моделям ASR, диаризации и LLM с метриками по каждому языку.
- Рекомендованный стек открытых моделей с коммерчески допустимыми лицензиями.
- Оценка потребности в дообучении для казахского и смешанной речи.
- Рабочий прототип полного цикла обработки на одном узле DGX Spark.
- Расчёт пропускной способности узла (пользователи, часы аудио, клиенты).