Чек-лист запуска Pyannote в продакшене

Диаризация дикторов — кто и когда говорил — на своей инфраструктуре

1 · Трёхстадийный конвейер

#	Стадия	Модель	Что делает
1	Сегментация	segmentation-3.0	Найти речь + наложение в окне 10 с (powerset)
2	Эмбеддинг	ResNet / WeSpeaker	Превратить каждый соло-голос в числовой отпечаток
3	Кластеризация	агломеративная	Объединить отпечатки; число дикторов — само собой
–	Агрегация	—	Собрать таймлайн кто-когда говорил (RTTM)

2 · Self-hosting vs хостинговый сервис

Критерий	Self-hosted pyannote	Хостинговый сервис
Стоимость ПО	Бесплатно (лицензия MIT)	Плата за час / минуту
Точность (DER)	Высокая; под ваши данные	Часто выше «из коробки»
Скорость	~40x реального времени на GPU	Управляется вендором, быстрее
Данные	Остаются на ваших серверах	Уходят в облако вендора
Запуск	GPU, модели, токен, настройка	API-ключ
Когда лучше	On-prem; стабильный объём	Быстро; высшая точность без эксплуатации

3 · Чек-лист перед запуском

- ☐ Подтвердили, что нужны именно метки дикторов — если нужны только слова, используйте ASR, а не диаризацию.
- ☐ Создали токен Hugging Face и приняли условия обеих моделей: segmentation-3.0 и speaker-diarization-3.1.
- ☐ Извлекли чистое моно-аудио 16 кГц — вход, который ожидает пайплайн.
- ☐ Передали число дикторов (num_speakers) или диапазон (min/max_speakers) везде, где продукт его знает.
- ☐ Проверили резидентность данных: может ли аудио легально покинуть ваши серверы или должно остаться on-prem?
- ☐ Сообщили DER честно: указали зазор и оценивалась ли перекрывающаяся речь; сравнивали одинаково.
- ☐ Задали ожидание, что наложение и короткие реплики — самые трудные случаи и могут быть помечены неверно.
- ☐ Построили путь человеческой коррекции для записей, где неверная метка стоит дорого (медицина, право).
- ☐ Рассмотрели настройку (порог кластеризации δ) или дообучение модели сегментации на доменном аудио.
- ☐ Смоделировали стоимость на реальном объёме (GPU-часы = часы аудио / коэф. реального времени) vs хостинг.