Памятка по выбору провайдера streaming ASR

Whisper · Deepgram Nova-3 · AssemblyAI Universal-Streaming — выбор по четырём вопросам

1 · Четыре вопроса, по порядку

	Вопрос	Если да
B1	Должно ли аудио оставаться на ваших серверах? (медицина, право, место хранения)	ДА → Whisper на своём хостинге
B2	Голосовой агент или просто транскрипт? (нужно ли знать, когда говорить?)	АГЕНТ → Deepgram / AssemblyAI
B3	Нужны ли слова быстрее ~300 мс? (мгновенные субтитры / отзывчивый UX)	ДА → облачный API
B4	Очень большой и очень ровный объём, плюс ML-команда для парка GPU?	ДА → свой хостинг может выиграть

2 · Три варианта (по данным вендоров, 2026)

Вариант	Задержка	Цена	Языки	Когда выбрать
Whisper (свой хостинг)	~3,3 с	только GPU	99	On-prem; очень большой масштаб
Deepgram Nova-3	< 300 мс	~$0,46/час	45+	Мин. задержка; много функций
AssemblyAI Universal-Streaming	~300 мс (неизм.)	$0,15/час	99+	Голосовые агенты; неизм. транскрипты

3 · Чек-лист перед запуском

- ☐ Решено, какие результаты финальные (неизменяемые), а какие частичные (пересматриваемые) — необратимые действия только по финалам.
- ☐ Задержка и WER измерены на вашем аудио, а не на демо-роликах вендора.
- ☐ Подтверждено, что набор языков покрывает вашу реальную аудиторию (не только английский).
- ☐ Проверён комплаенс / место хранения: может ли аудио легально покинуть ваши серверы?
- ☐ Если голосовой агент: выбрано определение конца реплики (нативно в Flux / Universal-Streaming или своё).
- ☐ Стоимость посчитана на реальном пиковом объёме, а не на среднем, с учётом простоя GPU при своём хостинге.
- ☐ Чистое аудио извлечено из WebRTC / пайплайна звонка (верная частота дискретизации, моно, без клиппинга).
- ☐ Параллельные потоки нагрузочно протестированы на ожидаемом пике до запуска.