Streaming ASR Provider Selection Worksheet

Whisper · Deepgram Nova-3 · AssemblyAI Universal-Streaming — pick by four questions

1 · Four questions, in order

	Question	If yes
Q1	Must the audio stay on your own servers? (medical, legal, data-residency)	YES → self-hosted Whisper
Q2	Voice agent, or just a transcript? (does it need to know when to speak?)	AGENT → Deepgram / AssemblyAI
Q3	Do words need to appear in under ~300 ms? (instant captions / responsive UX)	YES → hosted API
Q4	Very high, very steady volume, plus an ML team to run a GPU fleet?	YES → self-host can win on cost

2 · The three options (vendor-reported, 2026)

Option	Latency	Price	Langs	Best when
Whisper (self-host)	~3.3 s avg	GPU only	99	On-prem; very high scale
Deepgram Nova-3	< 300 ms	~$0.46/hr	45+	Lowest latency; broad features
AssemblyAI Universal-Streaming	~300 ms (immutable)	$0.15/hr	99+	Voice agents; immutable transcripts

3 · Pre-flight checklist before you ship

- ☐ Decided exactly which results are final (immutable) vs partial (revisable) — and act only on finals for irreversible actions.
- ☐ Measured latency and word error rate on your own audio, not vendor demo clips.
- ☐ Confirmed language coverage matches your real user base (not just English).
- ☐ Checked data-residency / compliance: can the audio legally leave your servers?
- ☐ If a voice agent: end-of-turn detection chosen (native in Flux / Universal-Streaming, or built yourself).
- ☐ Modelled cost at real, peaky volume — not steady-state — including idle GPU capacity if self-hosting.
- ☐ Extracted clean audio from the WebRTC / call pipeline (correct sample rate, mono, no clipping).
- ☐ Load-tested concurrent streams at expected peak before launch.