Streaming TTS Engine Selection Checklist

2026 latency + price comparison, host-vs-rent, and a pre-flight before you ship a voice feature

1 · The 2026 streaming-TTS options at a glance

Engine	Host / rent	TTFA P50	Price	Langs	Best for
Kokoro-82M	Self-host	—*	~$1 / M chars	8	Volume, privacy, cost floor
ElevenLabs Turbo v2.5	Hosted API	264 ms	$0.10 / 1k	32	Broad langs + steady latency
ElevenLabs Flash v2.5	Hosted API	288 ms	$0.05 / 1k	32	Low-cost low-latency
Cartesia Sonic-3	Hosted API	188 ms	from $0	40+	Fastest median, latency-first
OpenAI gpt-4o-mini-tts	Hosted API	stream	~$0.015/min	multi	Batch narration in OpenAI stack
OpenAI TTS-1-HD	Hosted API	2,295 ms	$30 / M	multi	Batch only — not real-time

*TTFA = median time-to-first-audio (Coval, May 2026). *Kokoro latency depends on your hardware (~210x real-time on one RTX 4090). Aim for < 300 ms TTFA; measure TTFA, not time-to-first-byte.*

2 · Host it, or rent it

Criterion	Host it (Kokoro)	Rent it (ElevenLabs / Cartesia / OpenAI)
Cost model	GPU + electricity (near-zero per use)	Per character of text
Setup	You run and scale the servers	Start in an afternoon
Privacy	Audio stays in your network	Audio leaves to the vendor
Languages	8	32 - 40+
Best at volume	Very high volume	Low to mid volume

3 · Pre-flight before you ship a streaming voice feature

- ☐ Decided real-time vs batch — if nobody waits live, pick on quality/price; if conversational, latency rules everything.
- ☐ Set a latency budget: target < 300 ms TTFA (under 200 ms is excellent), and remember TTS sits after STT + LLM in the chain.
- ☐ Measured time-to-first-audio, not time-to-first-byte — strip the file header before timing the first sound.
- ☐ Weighed latency consistency (spread / IQR), not just the median — a wide spread breaks a fraction of calls at scale.
- ☐ Costed your monthly volume in characters (~900 per minute of speech) against each engine's per-character rate.
- ☐ Modelled the self-host break-even: at high volume Kokoro's near-zero marginal cost beats any per-character bill.
- ☐ Confirmed your language coverage need (Kokoro 8 vs ElevenLabs 32 vs Cartesia 40+).
- ☐ Checked the privacy constraint: if audio cannot leave your network, self-host an open model.
- ☐ Chose the right transport: SSE when the full text is known up front, a reused WebSocket when text streams from an LLM.
- ☐ Tuned the chunk schedule against real LLM output and kept one WebSocket open per session to save 50-100 ms/turn.
- ☐ Picked a server region near your users (or self-hosted close to them) to cut network travel time.
- ☐ Verified the worst case, not just the demo: tested end-to-end TTFA from your users' real locations.