1. Model Evaluation Summary

This report presents a comparative analysis of multiple Text-to-Speech (TTS) models — Cartesia, ElevenLabs, Deepgram, and Sarvam — evaluated for latency, stability, and voice naturalness. The goal is to identify which model best balances performance with integration simplicity for your Django + FAISS + Celery stack.

2. Model Performance Metrics

Model	Avg Latency (ms)	Naturalness	Consistency	Best Use Case
Cartesia	430	Excellent	Stable	High-quality narration & streaming.
ElevenLabs	520	Very High	Good	Voice cloning & emotional tone.
Deepgram	610	Moderate	Stable	Real-time low-resource cases.
Sarvam	720	Good	Variable	Local/Offline regional inference.

3. VAPI Platform Evaluation

Vapi acts as an abstraction and orchestration layer for voice agents, managing pipeline latency, model routing, and session control. Its integration simplifies complex workflows but introduces a measurable latency overhead (~150–300ms per request).

While Vapi enhances modularity, scalability, and allows flexible back-end swapping (e.g., between TTS and STT vendors), it should not be used for ultra-low-latency real-time TTS or transcription tasks where direct SDK integration offers faster responses. It excels in orchestrating multi-step, conversational flows or when switching between models dynamically is required.

4. Recommendations

• **Use Direct SDK/API Integration** → For latency-critical streaming (TTS < 400ms). • **Use VAPI** → When modular routing, scaling, or concurrent multi-agent handling is required. • **Preferred TTS** → Cartesia for quality + stability balance. • **Avoid Overstacking APIs** → Each intermediary adds ~150–250ms delay cumulatively. • **Keep ASR/TTS model pairing consistent** → Use same vendor or latency-optimized mix (e.g., Cartesia + Deepgram).

5. Final Verdict

Vapi provides long-term scalability but adds extra latency. For real-time applications (live dubbing, streaming TTS), use direct SDK calls. For dynamic routing, multi-agent voice orchestration, or testing hybrid workflows, Vapi integration is justified. Model-wise, Cartesia and ElevenLabs lead for production-level clarity and expressiveness.