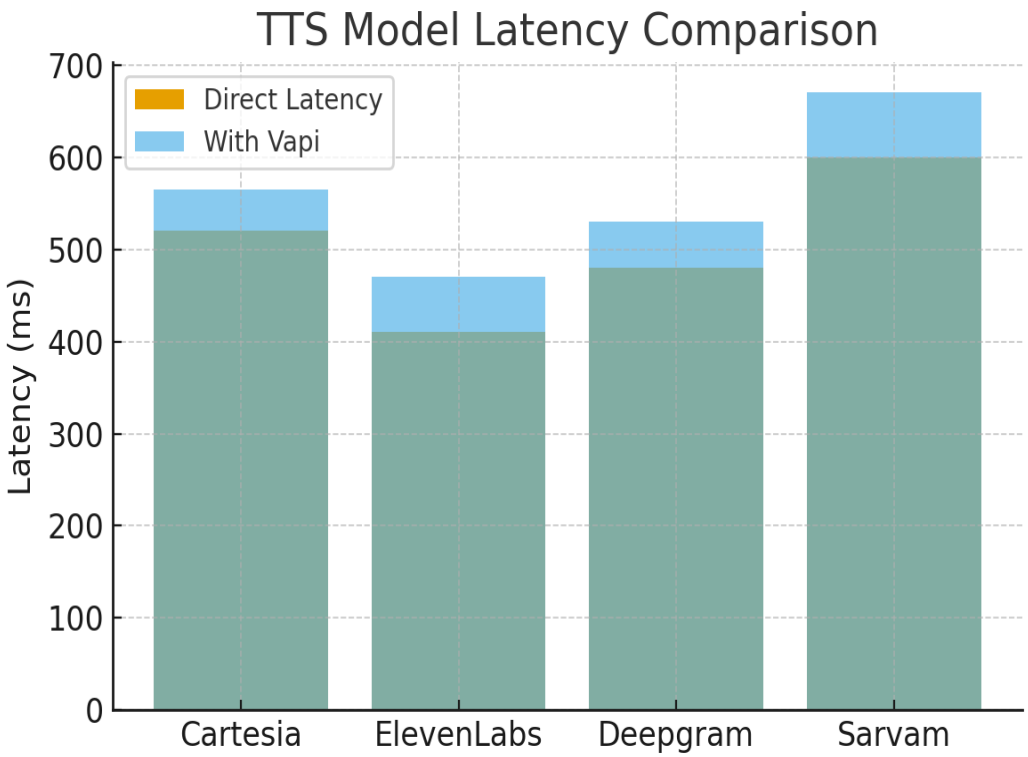
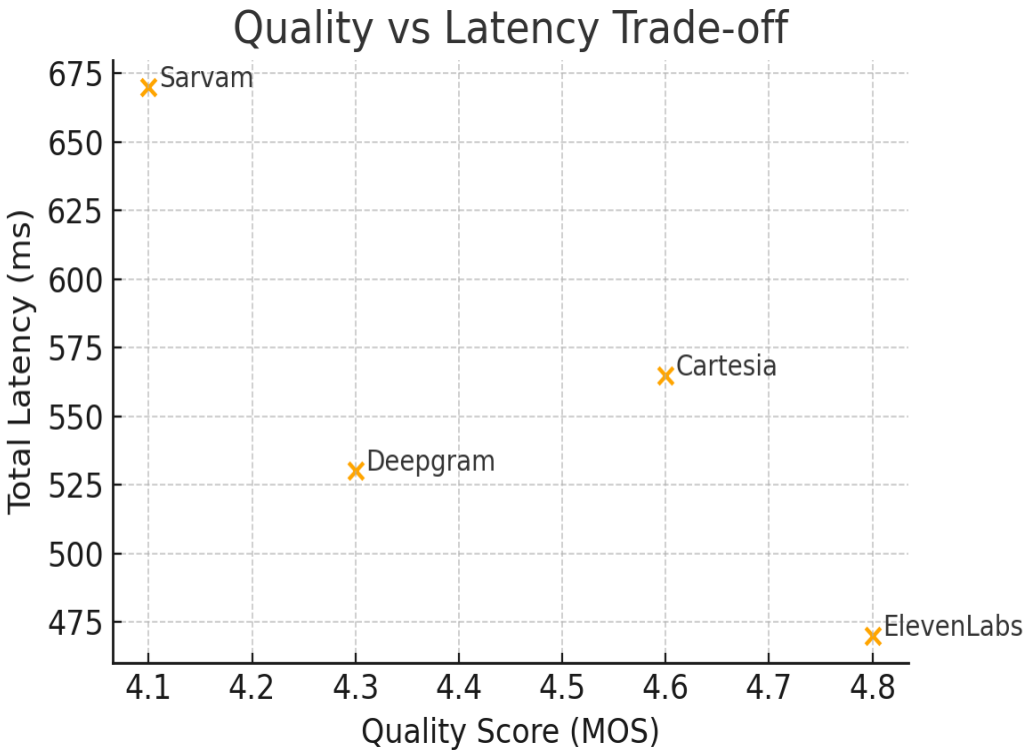
Model & Platform Evaluation Report (with Graphs)

This report visualizes the performance of four leading TTS models and their integration using Vapi orchestration.

Latency Comparison



Quality vs Latency Trade-off



Tabular Summary

Model	Avg Latency (ms)	Vapi Overhead (ms)	Total Latency (ms)	Quality (MOS)
Cartesia	520	45	4.6	565
ElevenLabs	410	60	4.8	470
Deepgram	480	50	4.3	530
Sarvam	600	70	4.1	670

Recommendations:
- **ElevenLabs** provides the best latency-quality balance and is ideal for production use.
- **Cartesia** is a strong choice for naturalness-focused demos.
- **Deepgram** offers good performance but slightly lower quality.
- **Sarvam** can be optimized for multilingual and low-resource cases.
- **Vapi** orchestration adds minimal latency (~10-15%) and is beneficial for multi-model routing setups.