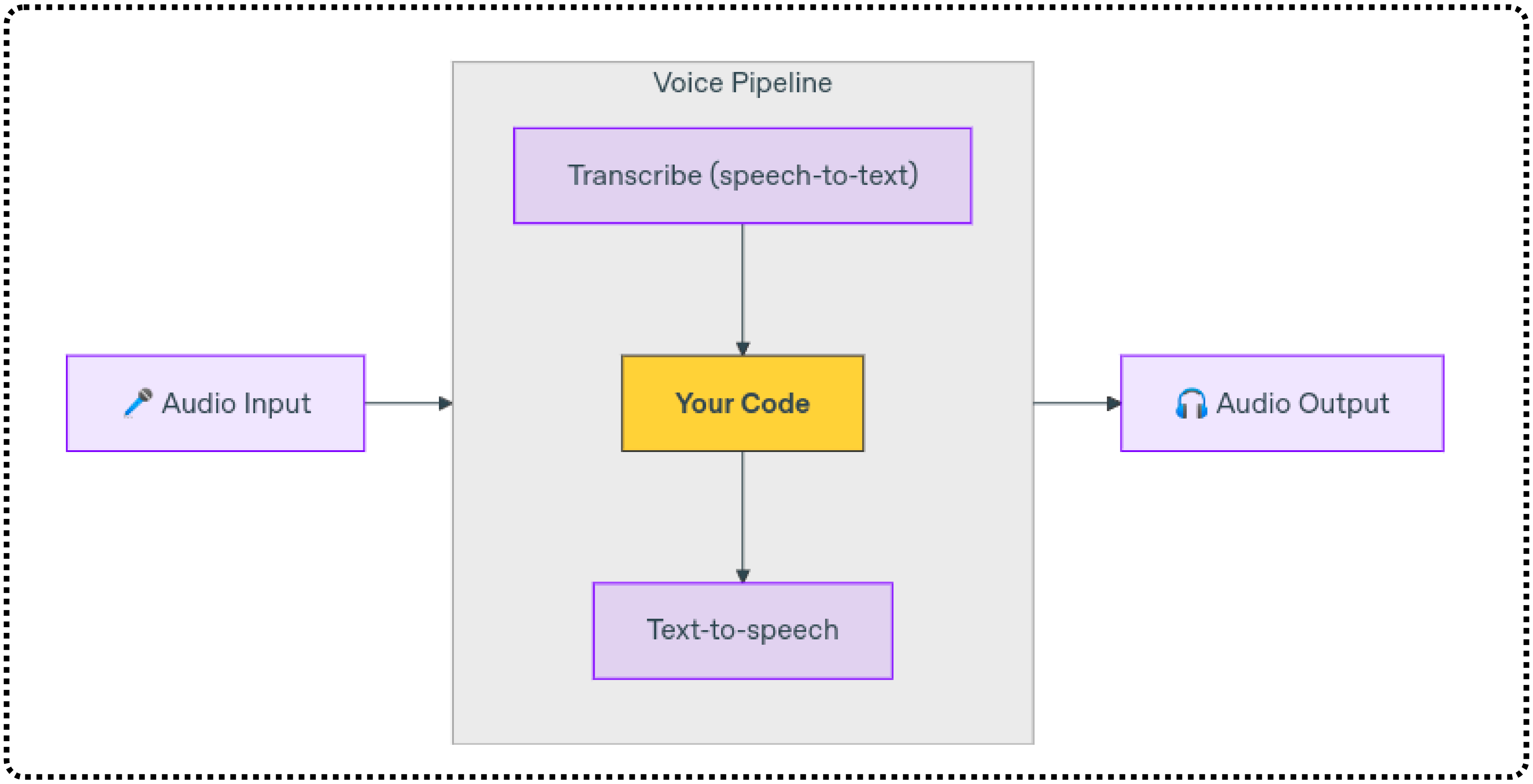
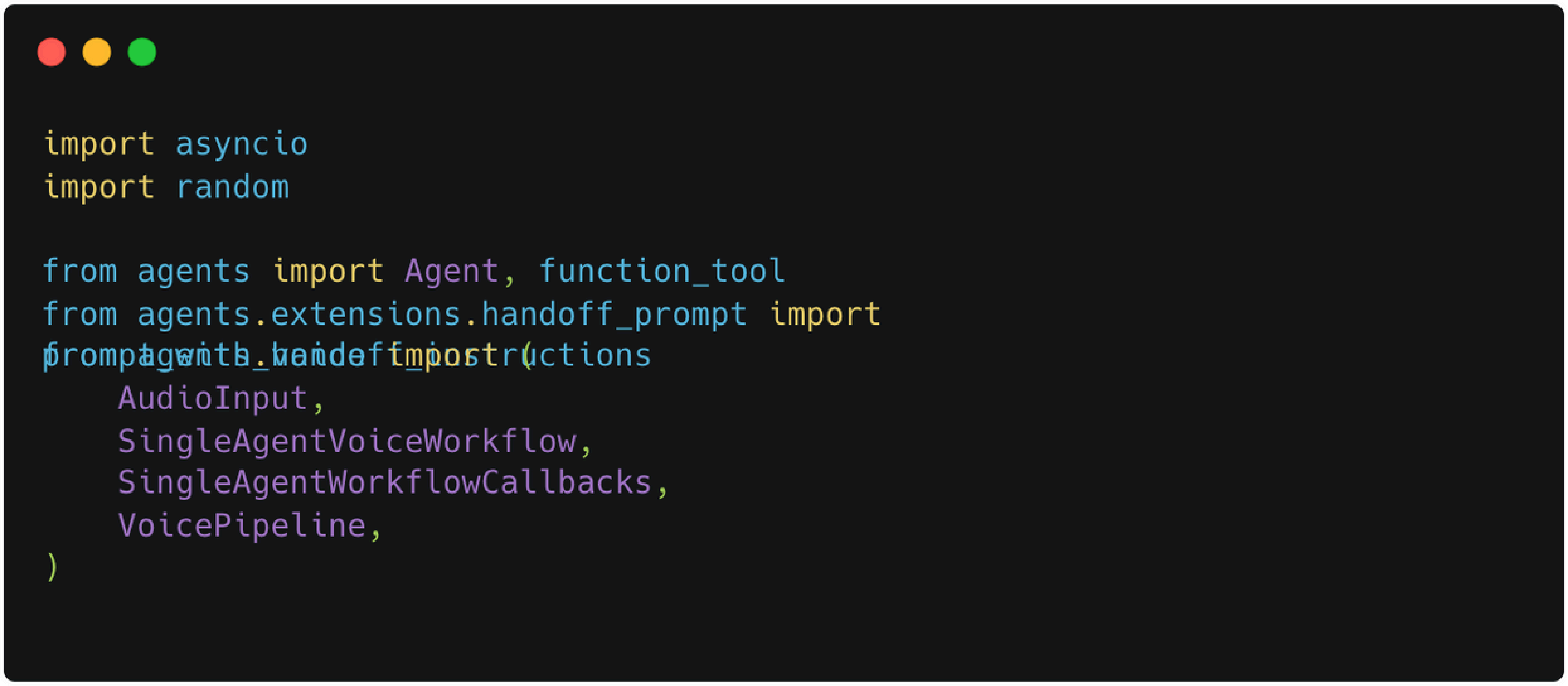
Analytics Vidhya

Build Multilingual Voice Agent Using OpenAI Agent SDK



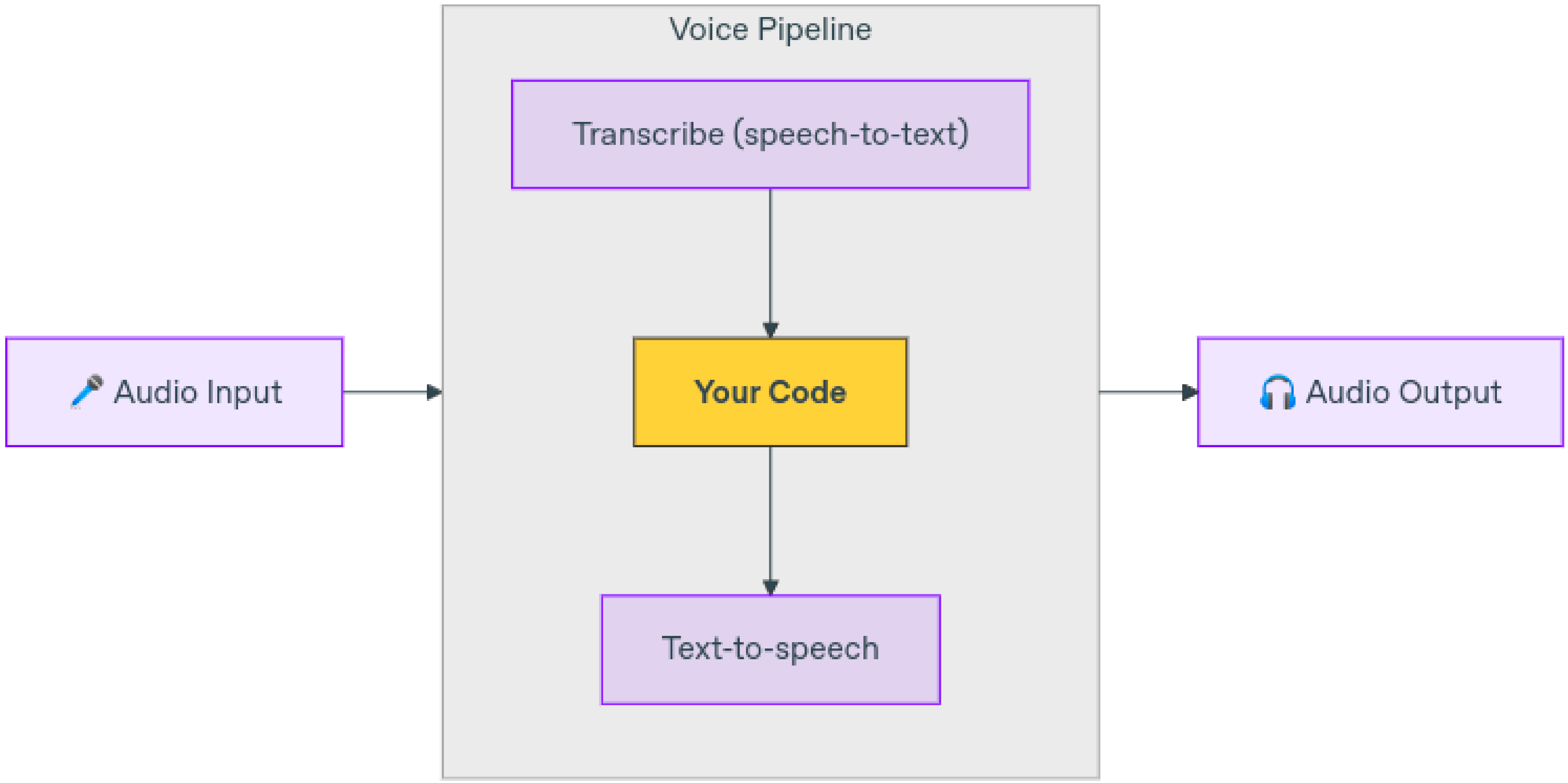


```
import asyncio
import random

from agents import Agent, function_tool
from agents.extensions.handoff_prompt import
prompt_with_handoff_instructions
from agents.voice import (
    AudioInput,
    SingleAgentVoiceWorkflow,
    SingleAgentWorkflowCallbacks,
    VoicePipeline,
)
```

What is a Voice Agent?

A Voice Agent is a system that listens to your voice, understands what you're saying, thinks about a response, and then replies out loud. The magic is powered by a combination of speech-to-text, language models, and text-to-speech technologies.



The OpenAI Agent SDK makes this incredibly accessible through something called a VoicePipeline—a structured 3-step process:

- Speech-to-text (STT)
- Agentic logic
- Text-to-speech (TTS)

Choosing the Right Architecture

Depending on your use case, you’ll want to pick one of two core architectures supported by OpenAI:

1. Speech-to-Speech (Multimodal) Architecture

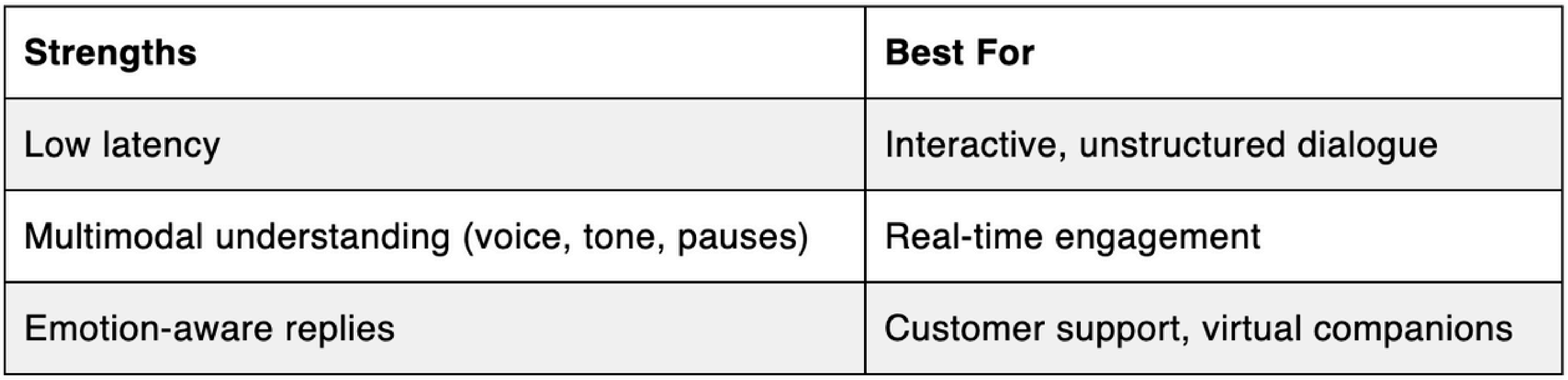
This is the real-time, all-audio approach using models like gpt-4o-realtime-preview. Instead of translating to text behind the scenes, the model processes and generates speech directly.

Why use this?

- Low-latency, real-time interaction
- Emotion and vocal tone understanding
- Smooth, natural conversational flow

Perfect for:

- Language Tutoring
- Live conversational agents
- Interactive storytelling or learning apps



Strengths	Best For
Low latency	Interactive, unstructured dialogue
Multimodal understanding (voice, tone, pauses)	Real-time engagement
Emotion-aware replies	Customer support, virtual companions

2. Chained Architecture

The chained method is more traditional: Speech gets turned into text, the LLM processes that text, and then the reply is turned back into speech. The recommended models here are:

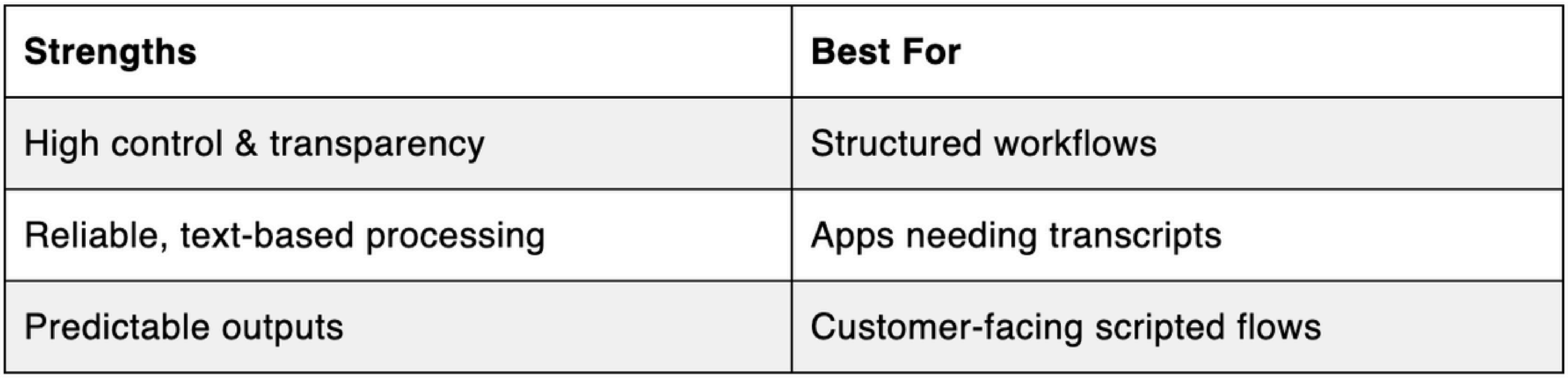
- gpt-4o-transcribe (for STT)
- gpt-4o (for logic)
- gpt-4o-mini-tts (for TTS)

Why use this?

- Need transcripts for audit/logging
- Have structured workflows like customer service or lead qualification
- Want predictable, controllable behaviour

Perfect for:

- Support bots
- Sales agents
- Task-specific assistants



Strengths	Best For
High control & transparency	Structured workflows
Reliable, text-based processing	Apps needing transcripts
Predictable outputs	Customer-facing scripted flows

How Does Voice Agent Work?

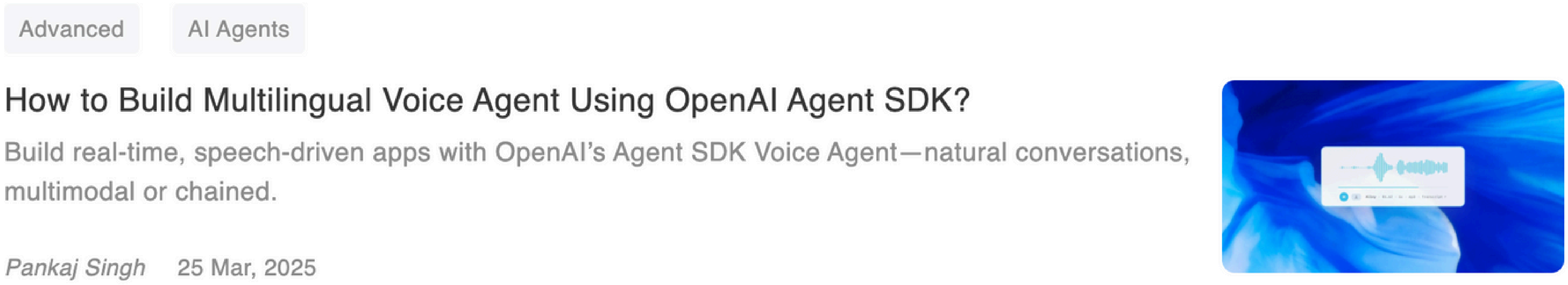
We set up a VoicePipeline with a custom workflow. This workflow runs an Agent, but it can also trigger special responses if you say a secret word.

Here’s what happens when you speak:

1. Audio goes to the VoicePipeline as you talk.
2. When you stop speaking, the pipeline kicks in.
3. The pipeline then:
   - Transcribes your speech to text.
   - Sends the transcription to the workflow, which runs the Agent logic.
   - Streams the Agent’s reply to a text-to-speech (TTS) model.
   - Plays the generated audio back to you.

It’s real-time, interactive, and smart enough to react differently if you slip in a hidden phrase.

For more information, kindly visit this article



Advanced AI Agents

How to Build Multilingual Voice Agent Using OpenAI Agent SDK?

Build real-time, speech-driven apps with OpenAI’s Agent SDK Voice Agent—natural conversations, multimodal or chained.

Pankaj Singh 25 Mar, 2025