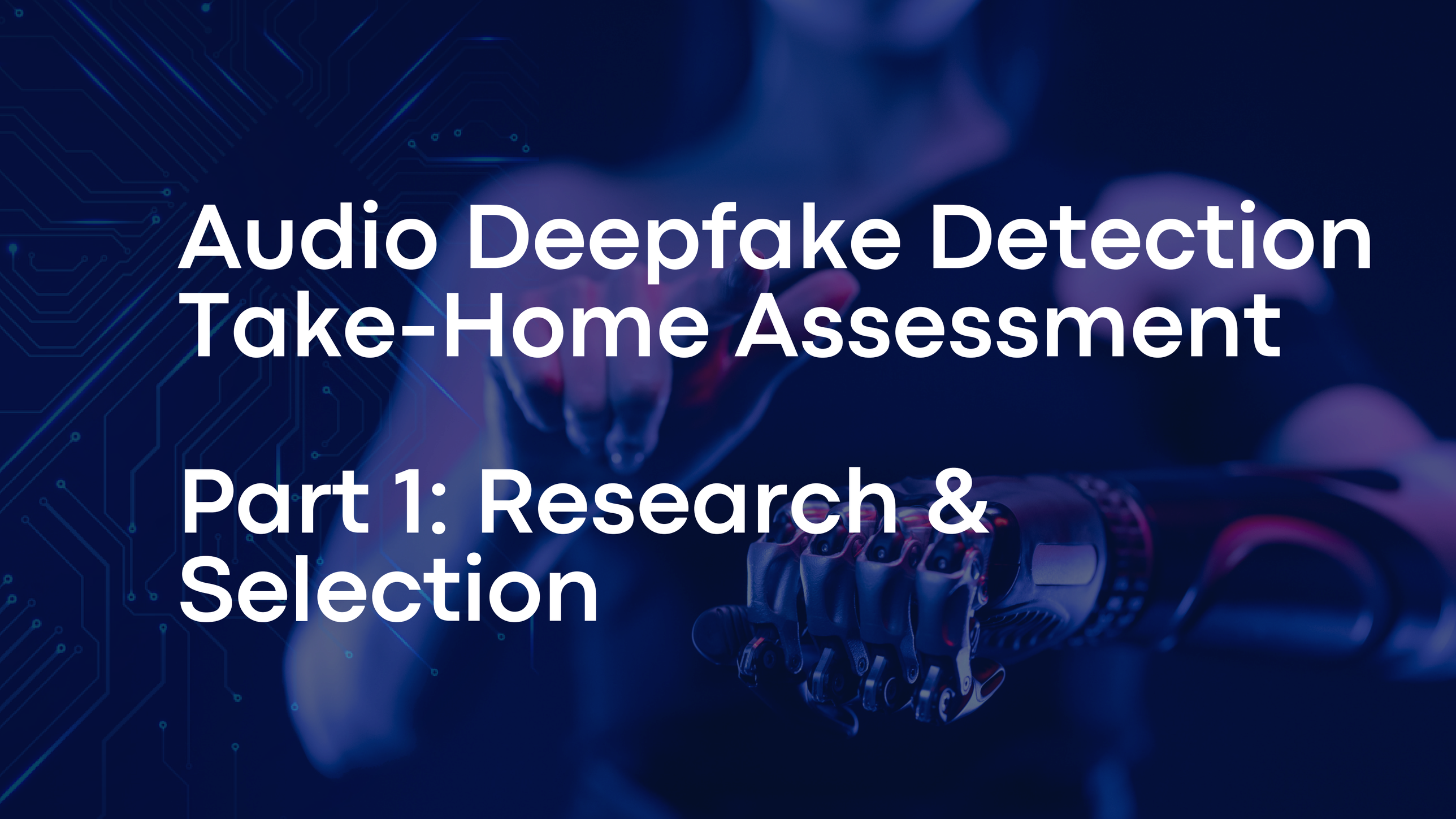
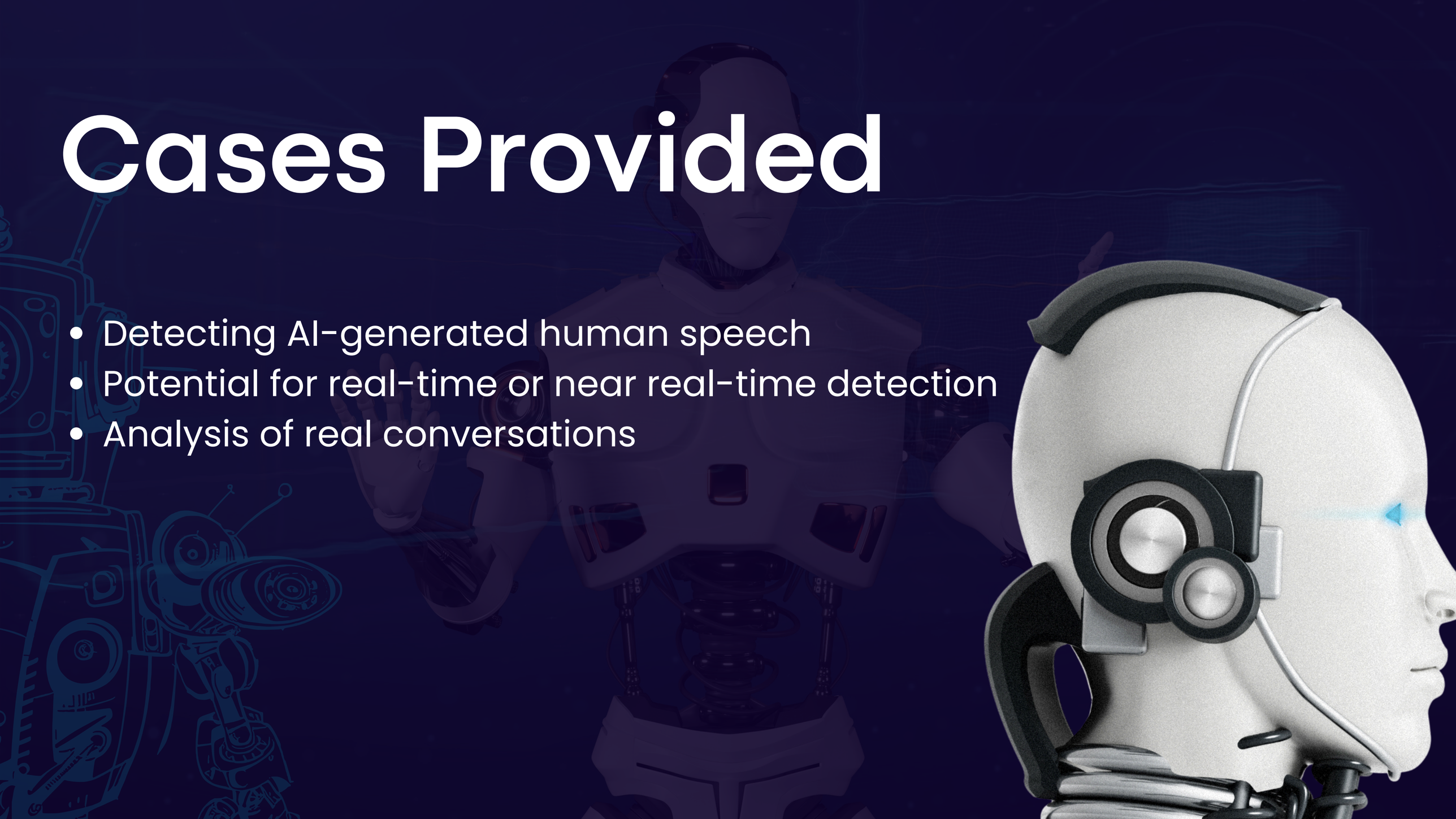
Audio Deepfake Detection Take-Home Assessment

Part 1: Research & Selection



Cases Provided

- Detecting AI-generated human speech
- Potential for real-time or near real-time detection
- Analysis of real conversations



Case 1: Detecting AI-generated human speech

Best Model: Pengi (Microsoft, 2023.05)

Key Technical Innovation:

- Pengi is an Audio Language Model (ALM) that frames all audio-related tasks as text-generation tasks, leveraging transfer learning for a unified approach to audio processing.
- It enables open-ended and close-ended tasks without requiring fine-tuning or task-specific modifications.

Reported Performance Metrics:

- Microsoft has demonstrated its ability to handle multi-modal inputs, but specific AI-generated speech detection metrics are limited in publicly available reports.
- It effectively processes and generalizes across various audio tasks due to its language model-based approach.

Why This Approach is Promising:

- Because AI-generated speech follows distinct token patterns, Pengi's text-generation framing makes it a strong candidate for identifying subtle anomalies in synthetic speech.
- Its transfer learning ability allows it to recognize variations in speech across different domains and speakers.

Potential Limitations or Challenges:

- Lack of direct AI-generated speech detection benchmarks may require additional custom fine-tuning.
- Since it's primarily designed for text generation from audio, it may not explicitly focus on speech authenticity.

Case 2: Real-Time or Near Real-Time Speech Detection

Best Model: USM

Key Technical Innovation

- Built for large-scale Automatic Speech Recognition (ASR) across 100+ languages.
- Uses a universal speech model trained on massive multilingual datasets for robustness.
- Optimized for low-latency transcription, making it ideal for real-time applications.

Reported Performance Metrics:

- Achieves state-of-the-art WER (Word Error Rate) on multilingual ASR benchmarks.
- Scales efficiently across varied acoustic conditions, making it resilient in real-world use.
- Can process spoken language in real-time, making it suitable for live detection tasks

Why This Approach is Promising:

- Since real-time detection needs fast and accurate transcription, USM's ASR capabilities make it highly suitable.
- Its ability to handle noisy conditions and diverse accents ensures high reliability.

Potential Limitations or Challenges:

- Not explicitly designed for AI-generated speech detection, so it might require an additional classification layer for deepfake detection.
- Computationally expensive for real-time applications if deployed on edge devices.

Case 3: Analyzing Real Conversations

Best Model: SpeechGPT

Key Technical Innovation

- Functions as a spoken dialogue language model, making it capable of both speech understanding and generation.
- Can perceive and generate multi-modal content, meaning it understands not just words, but tone, emotion, and context.
- Optimized for human-like conversational flow, making it well-suited for analyzing interactions.

Reported Performance Metrics:

- Outperforms traditional models in dialogue coherence and instruction-following ability.
- Benchmark comparisons against OpenAI’s Whisper and Microsoft’s VALL-E show strong conversational AI capabilities

Why This Approach is Promising:

- **Unlike basic ASR models, SpeechGPT can interpret the intent and nuances of dialogue, making it valuable for context-aware speech analysis.**
- **It can generate responses, meaning it could also simulate conversations for training AI models on real-world interactions.**

Potential Limitations or Challenges:

- Not optimized for large-scale transcription, so it may struggle with long-form content analysis.
- May require fine-tuning for domain-specific conversation analysis (e.g., medical, legal, or customer service dialogues).

Thank You!

-SAINATH

sainathbamandi018@gmail.com