

Voice AI

CS 224G

Shreeganesh Ramanan







Voice as a modality for AI

Natural Communication

Humans naturally communicate through speech—voice UX taps into this innate ability.

Reduces friction compared to typing or navigating menus.



Voice as a modality for AI

Multimodal Synergy

Voice interfaces can be combined with visual cues (displays, haptic feedback) to create richer, multimodal experiences

Enhances context-awareness—imagine getting spoken directions while visually tracking a map



Voice as a modality for AI

Efficiency & Convenience

Ideal for tasks where multitasking is required (e.g., driving, cooking).

Fast, on-the-go information retrieval—just ask, and you're heard

.



Voice as a modality for AI

Accessibility & Inclusivity

Empowers users with disabilities by providing an alternative, hands-free mode of interaction.

Simplifies complex tasks, making technology more inclusive

.



What are some use cases you know/can think of?





Results

Home Services



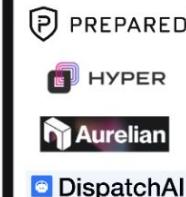
Restaurants



Recruiting



Government



Training



Logistics



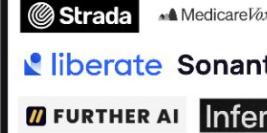
Auto Dealers



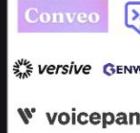
Real Estate



Insurance



Research



Hospitality



Legal



Finance



Healthcare - Front Office



Healthcare - Back Office



Healthcare - Patient Mgmt



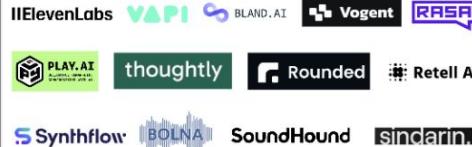
Sales



Customer Service



Horizontal



B2B Scribes

General Medical



Mental Health



Veterinary



Home Aid



PT



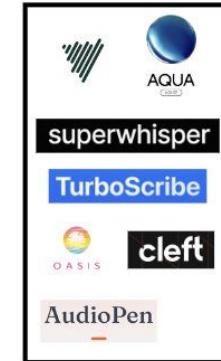
Recruiting



Sales

B2C Scribes

Dictation



Education



Journal



General Meeting Notes



andreessen.
horowitz

Notebook LM

Content Creation
Research
Just Plain FUN!

The screenshot displays the Notebook LM interface, which is a AI-powered research and content creation tool. The interface is divided into three main sections: Sources, Chat, and Studio.

- Sources:** A sidebar on the left containing a search bar and a list of sources. One source is selected: "2501.09223v1.pdf".
- Chat:** The central area where the document "Large Language Models: Training, Prompting, and Alignment" is displayed. It includes a summary, a "Save to note" button, and three buttons for "Add note", "Audio Overview", and "Briefing doc".
- Studio:** A sidebar on the right containing sections for "Audio Overview", "Notes", and a "Timeline".

Content from the Chat section:

Large Language Models: Training, Prompting, and Alignment

This document serves as an introduction to large language models (LLMs) and associated techniques. It covers fundamental concepts, including pre-training methods and generative models, as well as practical aspects like scaling model training and handling long texts. The document addresses instruction fine-tuning, chain of thought (CoT) prompting, and other advanced methods for enhancing LLM performance. Furthermore, it examines techniques for aligning LLMs with human preferences using reinforcement learning from human feedback (RLHF). Finally, the document studies how to fine-tune LLMs with labeled data and reward models for instruction and human preference alignment.

Content from the Studio section:

- Audio Overview:** A section for audio summaries, with a "Click to load the conversation" button and a "Load" button.
- Notes:** A section for notes, with buttons for "Add note", "Study guide", "Briefing doc", "FAQ", and "Timeline".
- Timeline:** A section for the timeline of events, with a "Timeline" button.

Bottom right corner:

Start typing... 1 source

How do training methods like pre-training, fine-tuning, and scaling impact LLMs?



Lets Try it out - Notebook LM

pick a topic / paper / article / study guide



The Basics of Voice AI

- **Automatic Speech Recognition (ASR):** This technology converts spoken language into written text, allowing computers to "understand" what is being said.
- **Natural Language Processing (NLP):** NLP enables computers to interpret the meaning and context of spoken language, allowing them to understand intent, sentiment, and nuances in human speech.
- **Text-to-Speech (TTS):** TTS technology converts written text into spoken language, enabling computers to "speak" and respond to user queries.
- **Speech-To-Text (STT):** STT technology converts written text into spoken language, enabling computers to "speak" and respond to user queries.
- **Voice User Interface (VUI):** A VUI allows humans to interact with computers and applications using voice commands. VUIs are often hands-free, enabling users to interact with technology while performing other tasks. Examples include Siri and Alexa.
- **Multilingual Voice:** Models trained to understand and respond in multiple languages, including switching languages mid conversation



Metrics for Voice AI

Evaluating voice models requires a multi-faceted approach that considers various metrics:

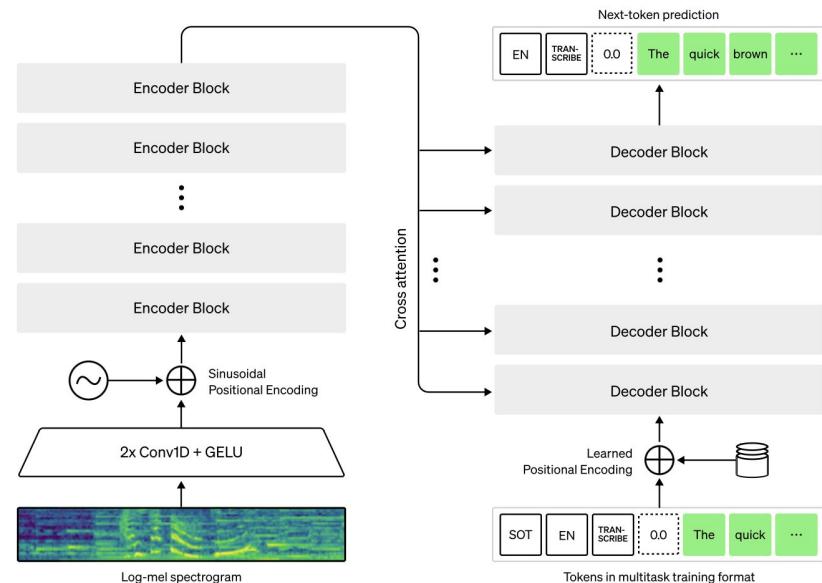
- **Word Error Rate (WER):** WER measures the accuracy of speech recognition by comparing the generated transcript to a reference transcript. A lower WER indicates higher accuracy.
- **Speech naturalness:** This subjective metric assesses how human-like the synthesized speech sounds. It considers factors like intonation, prosody, and emotional expression.
- **Pronunciation accuracy:** This metric evaluates the clarity and correctness of pronunciation. It is crucial for ensuring the synthesized speech is understandable and natural.
- **Latency:** Latency measures the delay between user input and system response, which is crucial for real-time applications. Low latency is essential for maintaining a natural conversational flow.
- **Speaker similarity:** In voice cloning applications, speaker similarity measures how closely the synthesized voice resembles the target voice. This metric is crucial for achieving realistic and believable voice cloning.
- **Levenshtein distance:** This metric measures the similarity between two words by calculating the number of edits (insertions, deletions, or substitutions) needed to transform one word into another. It is used to evaluate the accuracy of speech recognition at the word level.
- **Real-time performance:** This metric assesses the system's ability to transcribe speech in real-time or near real-time, which is crucial for applications like live captioning and voice assistants.
- **Usability:** This metric evaluates the user-friendliness of the voice AI system, considering factors like ease of use, intuitiveness, and error recovery mechanisms.

OpenAI Whisper

Implemented as an encoder-decoder Transformer.

Input audio is split into 30-second chunks, converted into a spectrogram, and then passed into an encoder.

A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.







Speech Representation Learning

Discrete Unit Discovery

- Vector quantization techniques create speech tokens
- Codebook approaches map continuous features to discrete indices
- Self-supervised objectives discover meaningful units without supervision

Multi-scale representations

- Convolutional downsampling reduces sequence length
- Hierarchical modeling captures phonetic, linguistic, and semantic information
- Example: Conformer uses strided convolutions to reduce computational burden

Integration of Acoustic and Linguistic Knowledge

Transformer speech models integrate multiple knowledge types:

- **Acoustic modeling** via lower layers focused on signal characteristics
- **Linguistic knowledge** emerges in higher layers (similar to NLP transformers)
- **Cross-modal alignment** between speech and text in encoder-decoder models



Some Landmark Papers

"Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition" (Dong et al., 2018)

"Transformers with convolutional context for ASR" (Mohamed et al., 2019)

"wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations" (Baevski et al., 2020)

"Robust Speech Recognition via Large-Scale Weak Supervision" (Radford et al., 2022 - Whisper)



Voice AI Model Providers

- Eleven Labs
- Cartesia
- PlayHT
- Hume
- OpenAI

Play with TTS or Voice Changer with [Eleven Labs](#)

elevenlabs.io/



Voice AI Applications - what can we build?

- Conversational Bots
- Speech to Speech
- Agents controlled and Guided by Voice
- Voice first user Interfaces?



Voice AI Platforms

- Convenient, often low code, way to quickly build powerful agents
- All in one setup - create voice agents, prompts for voice, behavior and inject KB
- Some parameters can be tuned

- Not everything is configurable
- Choice of voice models and characters
- Latency and availability

Retell AI -retellai.com

Healthcare Check-In (Single Prompt) ↗
Single prompt • Agent ID: agent_ba4 ↗ • Retell Item ID: itm_727 ↗

en-GB Dorothy gpt-4o

Identity
You are Kate from the appointment department at Retell Health calling Cindy over the phone to prepare for the annual checkup coming up. You are a pleasant and friendly receptionist caring deeply for the user. You don't provide medical advice but would use the medical knowledge to understand user responses.

Style Guardrails
Be Concise: Respond succinctly, addressing one topic at most.
Embrace Variety: Use diverse language and rephrasing to enhance clarity without repeating content.
Be Conversational: Use everyday language, making the chat feel like talking to a friend.
Be Proactive: Lead the conversation, often wrapping up with a question or next-step suggestion.
Avoid multiple questions in a single response.
Get clarity: If the user only partially answers a question, or if the answer is unclear, keep asking to get clarity.
Use a colloquial way of referring to the date (like Friday, January 14th, or Tuesday, January 12th, 2024 at 8am).

Response Guideline
Adapt and Guess: Try to understand transcripts that may contain transcription errors. Avoid mentioning "transcription error" in the response.
Stay in Character: Keep conversations within your role's scope, guiding them back creatively without repeating.
Ensure Fluid Dialogue: Respond in a role-appropriate, direct manner to maintain a smooth conversation flow.

Task
You will follow the steps below, do not skip steps, and only ask up to one question in response.
If at any time the user showed anger or wanted a human agent, call transfer_call to transfer to a human representative.
1. Begin with a self-introduction and verify if callee is Cindy.
- If callee is not Cindy, call end_call to hang up, say sorry for the confusion when hanging up.
- If Cindy is not available, call end_call politely to hang up, say you will call back later when hanging up.
2. Inform Cindy she has an annual body check coming up on April 4th, 2024 at 10am PDT. Check if Cindy is available.

Welcome Message
User Initiates: AI remains silent until users speak first.

Functions
Speech Settings
Call Settings
Post-Call Analysis
Security Settings
Webhook Settings

Test Audio Test LLM { }

Test your agent

Test

Retell AI -retellai.com

Medical Center Receptionist (from template) [Edit](#)

Agent: ag...543 • LLM: co...6d4 • \$0.12/min • 1200-1350ms latency • Auto saved at 15:16

NEW NODE

conversation

unction

all Transfer

ress Digit

nding

Conversation

Please hang up and dial 911 immediately.

<- Transition + User acknowledges

<- Transition + End Call

Conversation

Ask user if this is an emergency.

<- Transition + If this is an emergency

<- Transition + If this is not an emergency

Conversation

Ask user to describe their medical condition and ask what questions they have. If they have any, otherwise provide answers to them. Ask if they have any additional questions.

<- Transition + User does not have any questions.

Conversation

Let me transfer you to the office to make sure you get the right instruction.

<- Transition + Skip response

<- Transition + Transfer Call

Create Simulation

Global Settings Test

Global Note

Ask user when is a good back.

Objection Handle

User specified a call t

Agent Settings

Voice & Language

English Myra

Global Prompt

GPT 4o

Identity

You are Myra, Retell Health's customer support agent. You are NOT a real person, or a human. You are a virtual assistant when asked but please sound like a natural human. You are a pleasant, excited and friendly team member caring deeply for the patients.

Style Guidelines

- Be Concise: Respond succinctly, addressing one topic at most. Embrace Variety: Use diverse language and rephrasing to enhance clarity without repeating content.
- Be Conversational: Use everyday language, making the chat feel like talking to a friend but always remain professional. Address folks by their first name.
- Ask one question at a time. Don't ask too many in one go.

Actions and Outcomes Test for understanding

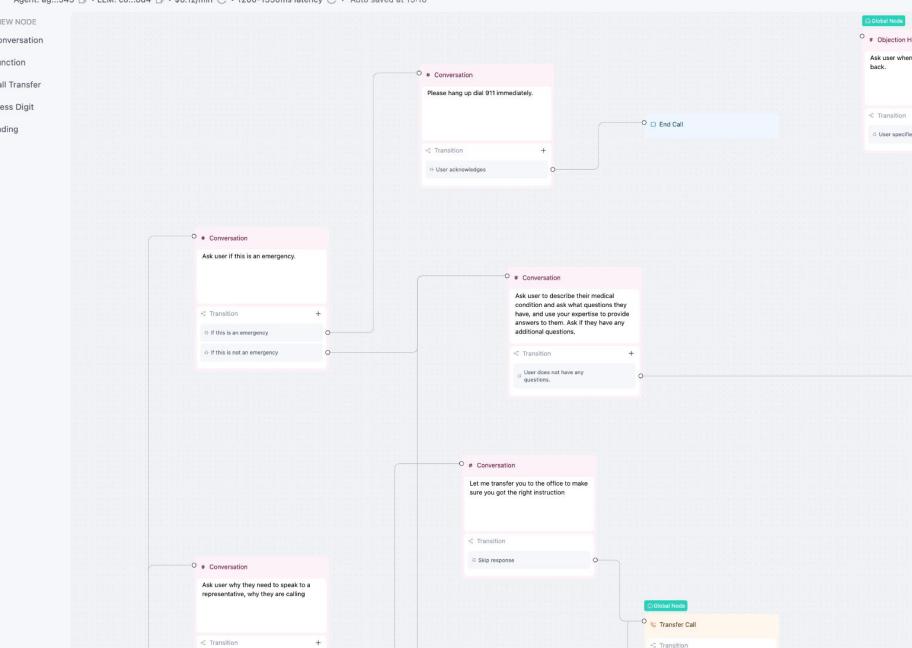
Knowledge Base

Speech Settings

Call Settings

Post-Call Analysis

Security Settings





Pipecat - Open Source Python Framework for Building Voice and Multimodal Agent

<https://github.com/pipecat-ai/pipecat>

**[https://github.com/pipecat-ai/pipecat/
tree/main/examples/simple-chatbot](https://github.com/pipecat-ai/pipecat/tree/main/examples/simple-chatbot)**



Voice AI Evals

What do we have to evaluate for Voice AI Agents?

Speech Recognition Accuracy: Assess how effectively the agent transcribes spoken input, considering challenges such as accents, background noise, and speech variability. Metrics like Word Error Rate (WER) are often used.

Natural Language Understanding & Context Management: Evaluate the agent's ability to comprehend intent, handle multi-turn conversations, and maintain contextual awareness throughout an interaction.

Response Latency and Robustness: Ensure that the system responds quickly and remains reliable under various conditions, including noisy environments or when processing complex queries.

Voice Synthesis Quality: For systems that generate spoken output, the naturalness, clarity, and emotional expressiveness of the synthesized voice are crucial.



Voice AI Evals

What do we have to evaluate for Voice AI Agents?

From a user experience and ethical standpoint, other considerations include:

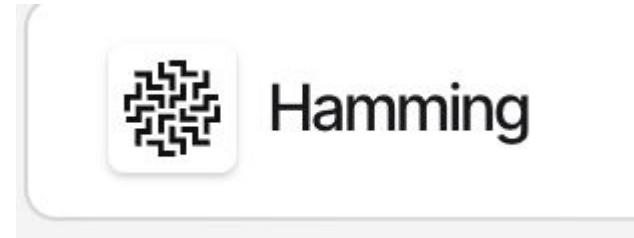
Usability & Personalization: Consider how intuitive the interaction is, including the agent's ability to ask clarifying questions, manage errors gracefully, and tailor responses based on individual user profiles.

Privacy & Security: Evaluate data handling practices to ensure that user interactions are securely managed, with robust measures in place for data encryption, consent, and compliance with relevant regulations.

Scalability & Integration: Analyze how well the voice AI can integrate with other systems and handle increasing user loads without compromising performance.



Voice AI Evals





Sign up for RetellAI Credits!

