Speech Recognition and Summarization

Team Members:
Glenys Lion
Fuqian Zou
Iris Lee
Kavya Bhat
Liana Bergman-Turnbull

Northwestern | McCORMICK SCHOOL OF ENGINEERING

Table of Contents

Problem Statement

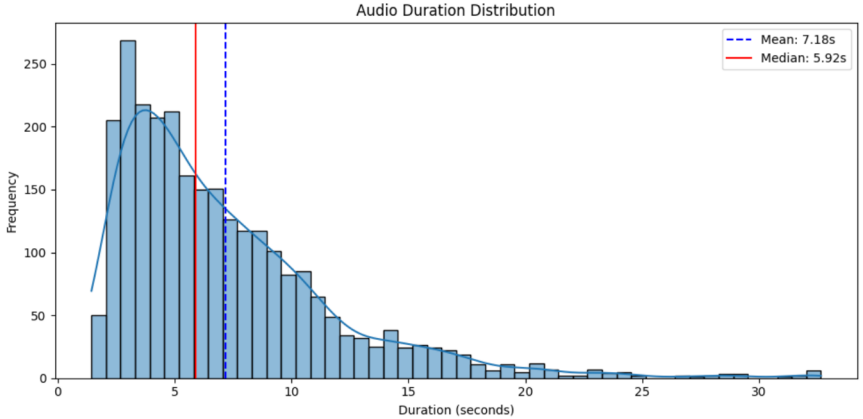
Extracting insights from spoken language can be challenging due to the unstructured nature of audio data. This project aims to develop an interactive interface that processes audio files from the LibriSpeech dataset by performing automatic speech recognition (ASR), text cleaning, named entity recognition (NER), and multi-level summarization. The goal is to transform raw spoken content into structured and meaningful textual information that users can explore through a user-friendly tool.

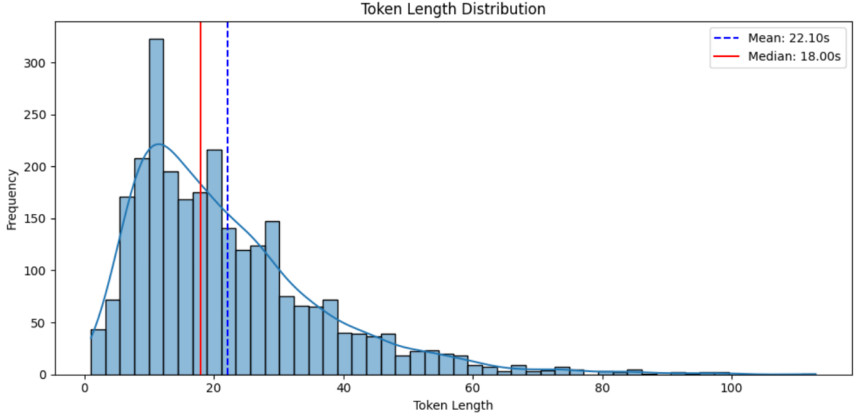
Data Overview

- The data examined comprised the dev-clean subset of the LibriSpeech dataset
- It has 2703 clean English audio files
- After preprocessing, a DataFrame was created comprising three columns:
  - filepath (location of the audio file)
  - duration (in seconds)
  - text (the corresponding transcript)

Exploratory Data Analysis

Audio & Token Characteristics



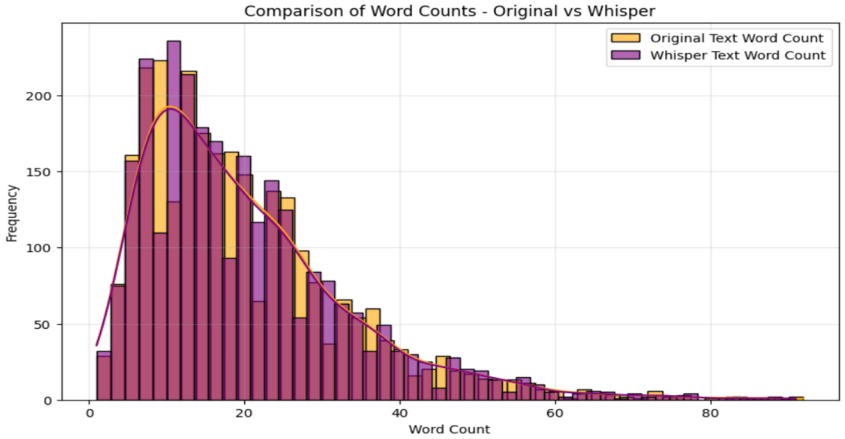


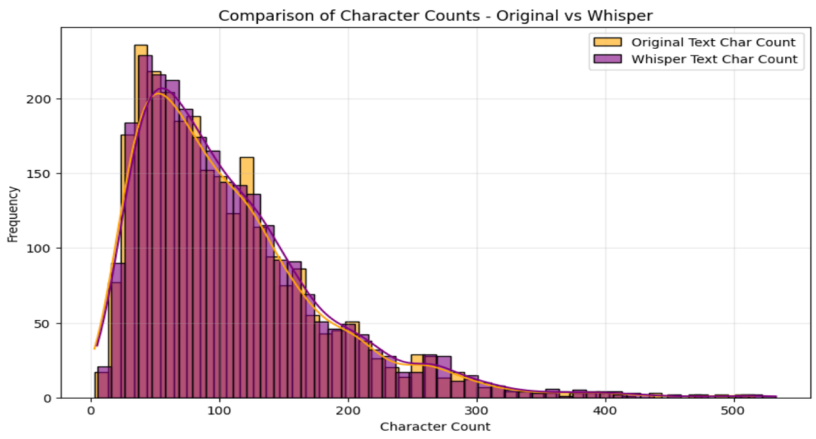
- The audio duration distribution is **right-skewed**, with most clips under **10 seconds**, indicating short and focused utterances
- The highest concentration of clips falls between **2 to 7 seconds**, aligning well with typical sentence-level speech suitable for transcription and summarization tasks

- The token length distribution follows a **right-skewed** pattern, with the majority of transcriptions containing between **10 to 30 tokens**
- The close alignment of the **mean (22.10)** and **median (18.00)** token lengths highlights consistent transcription granularity

Exploratory Data Analysis

Whisper Transcriptions vs Original Text

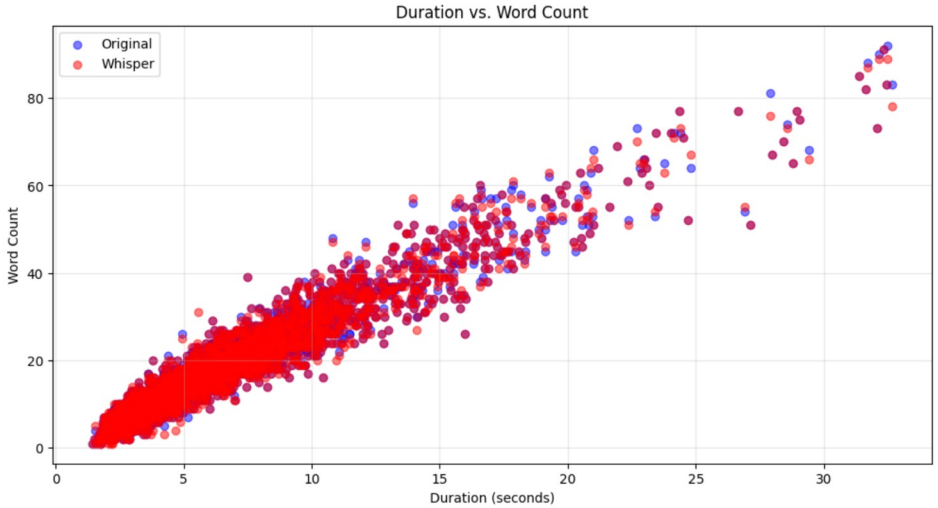




- Whisper's pre-trained model produces transcripts that **closely mirror the original text in both word and character count distributions.**
- Most audio clips are short (**10–30 words, 50–150 characters**), suggesting that the dataset contains clear, sentence-level speech ideal for summarization and NER tasks.

Exploratory Data Analysis

Whisper Transcriptions vs Original Text

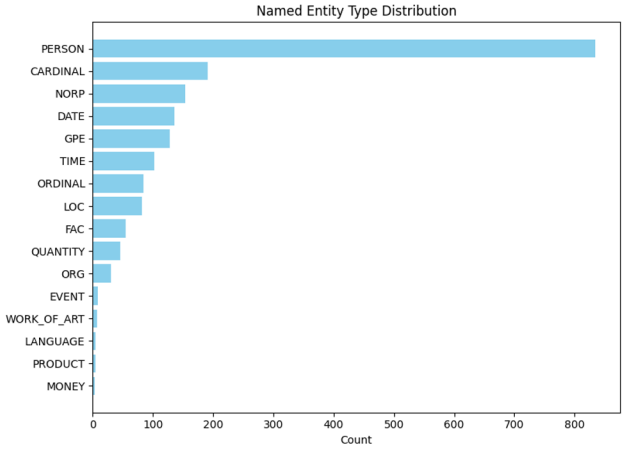


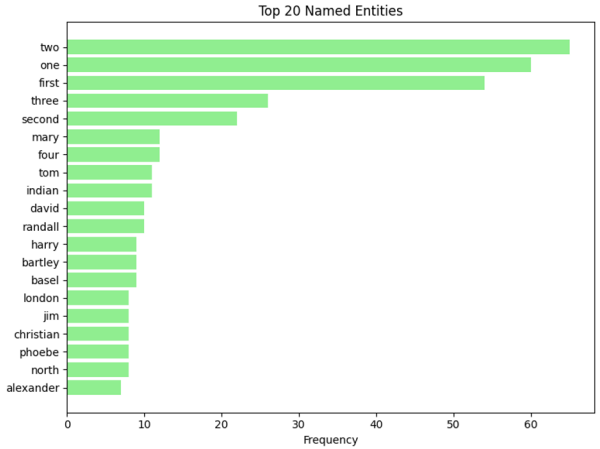
- To validate consistency between the original and Whisper-transcribed texts, we analyzed the **relationship** between **audio duration and word count.** Both showed a **strong positive correlation (r = 0.96),** highlighting Whisper's accuracy in preserving content length.

NER Model

Name Entity Recognition (NER) Model - EN_CORE_WEB_TRF

- 97 combined audio texts
- average of 19.27 entities per audio text
- 98.97% entity coverage (only 1 audio text has no entities detected)





Summarization

Initial Summarization Using T5-Large

What We Did:

- Used the pre-trained **T5-Large** model to summarize our dataset
- Used 3 styles: **Tiny, Short, and Long** by changing the length size

What We Observed:

- Summaries were repetitive. Just **shorter cuts** of the original text
- There **was not much difference between** Tiny, Short, and Long versions
- Important points were **unclear** in the summaries

Output Sample:

Original Text:
mr quilter is the apostle of the middle classes and we are glad to welcome his gospel nor is mr quilters manner less interesting than his matter he tells us that at this festive season of the year with christmas and roast beef looming before us similarly drawn from eating and its results occur most readily to the mind he has graved doubts whether sir frederick laytons work is really greek after all and can discover in it but little of rocky ithaca lynelle's pictures are a sort of upgrades and atom paintings and masons exquisite titles are as national as a jingo poem mr birkett fosters landscapes smile at one much in the same way that mr carker used to flash his teeth and mr john collier gives his sitter a cheerful slap on the back before he says like a shampoo or a turkish bath next man it is obviously unnecessary for us to point out how luminous these criticisms are how …

Tiny Summary:
mr quilter is the apostle of the middle classes and we are glad to welcome his

Short Summary:
mr quilter is the apostle of the middle classes and we are glad to welcome his gospel . he graved doubts whether sir frederick laytons work is really greek after all and can discover in it but little of rocky ithaca lynelle's pictures are a sort of upgrades and atom

Long Summary:
mr quilter is the apostle of the middle classes and we are glad to welcome his gospel . he graved doubts whether sir frederick laytons work is really greek after all and can discover in it but little of rocky ithaca lynelle's pictures are a sort of upgrades and atom paintings . but he has failed even to make himself the tupper of painting by harry quilter mason .

T5-Fine Tuning

Why Fine-Tune T5?

Baseline Model T5-Large:

- Did not generate useful summaries
- Output were too general, sometimes missing the key information
- **Main issue:** Lack of meaningful content

Our Approach:

- Use a **larger model (LLaMA)** to produce high quality summaries
- Do **prompt engineering** on LLaMA to get good quality of summary
- These summaries became the targets to **fine-tune T5-small (faster to train) using LoRA**

Generating High-Quality Targets:

- Applied **prompt engineering** on LLaMA to generate better summaries
- Created **three summary styles**: Tiny, Small, and Large using different prompt styles
- These summaries are the **training targets** to fine-tune T5-small using LoRA

Why LoRA?

- LoRA (Low-Rank Adaptation) is a lightweight fine-tuning method
- Adds **small trainable layers** instead of updating the whole model - making it f**aster and more efficient**

T5-Fine Tuning

LoRA Fine-Tuning Process

- **Fine-tuned 3 separate T5-small models** using LoRA: One for Tiny, one for Small, one for Large summary target
- Each model was trained using LLaMA-generated summaries

What was the result?

- The fine-tuned model produced more **accurate** and **detailed summaries**, based on **ROUGE** scores
- On average, LoRA fine-tuned T5-small **improved performance** by over 50% compared to the base model
- LoRA is **effective** even with a small model like T5-small when trained with high quality targets

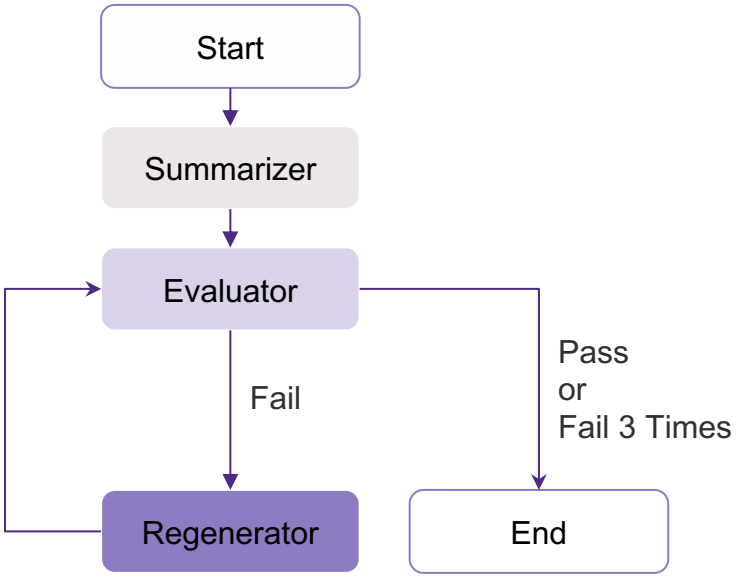
```
Avg ROUGE Score:
Base Model: {'rouge1': 0.22242367494055887, 'rouge2': 0.07391086622071422, 'rougeL': 0.1522966652116537, 'rougeLsum': 0.17819529220407432}
LoRA Model: {'rouge1': 0.48945059088977905, 'rouge2': 0.19646030757948024, 'rougeL': 0.3003381957180717, 'rougeLsum': 0.36425469905416363}
```

Final Application:

- Used the **fine-tuned LoRA weights** to generate 3 summary types (Tiny, Small, Large) for the full dataset
- Each summary types was generated using a separate **T5-small model**, each **fine-tuned** with its **own LoRA weights**
- This approach is **fast and efficient**. No need to retrain or update the full model
- The same weights can be **reused** anytime to generate summaries, including in our **Streamlit interface** for live, interactive summarization

LLM-Based Agentic Summarization System

- **Model:** Llama 3 8B Instruct
- **In-Context Learning (ICL) Prompt**



Summarizer

Prompt:
Instruction - Summarize the following text into …
ICL - Original text examples & summarized text examples
Input - Original text
Response: Summarized text

Evaluator

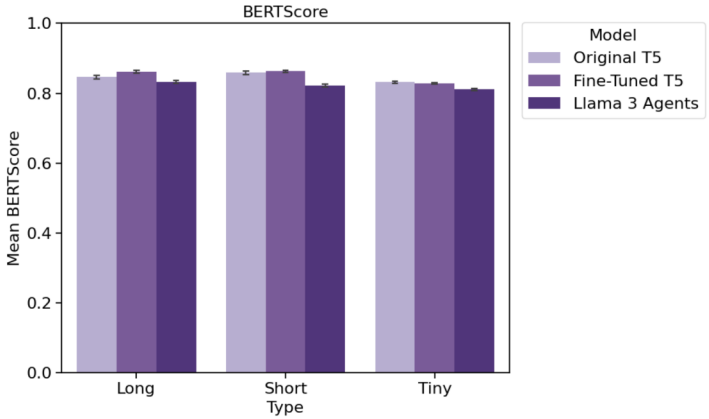
Prompt:
Instruction - Evaluate the summary based on …
ICL - Original text examples & summarized text examples & evaluation result examples & feedback examples
Input - Original text & summarized text
Response: Evaluation result & feedback

Regenerator

Prompt:
Instruction - Address the feedback and summarize the text …
Input - Original text & last summary & feedback on the last summary
Response: New summarized text

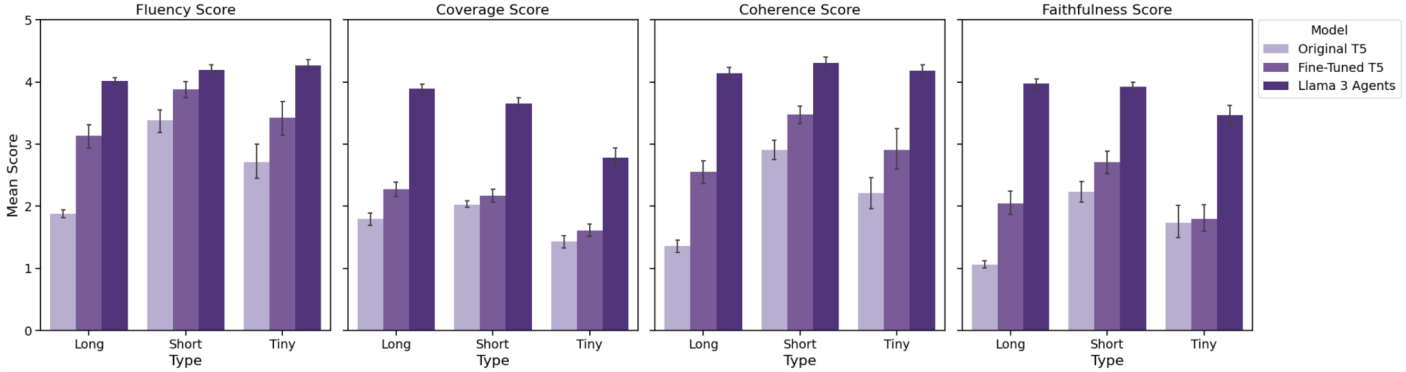
Model Evaluation and Comparison

- **BERTScore:** Evaluates the semantic similarity between two texts by comparing the contextual embeddings of their tokens using BERT
- **Results:**
  - The fine-tuned T5 outperforms the original T5 and Llama 3 agents for both long and short summaries
  - The original T5 is the best for tiny summaries
- **Pros:** Unlike ROUGE or BLEU, BERTScore uses contextual embeddings, capturing meaning beyond exact word overlap
- **Cons:**
  - Summaries generated by the original T5 model often include repeated or copied sentences from the original text, which can inflate BERTScore values and reduce its reliability as an indicator of true summarization quality
  - Scores may degrade when comparing texts with very different lengths (e.g., 300-word source vs. 20-word tiny summary)



Model Evaluation and Comparison

- **LLM-as-a-Judge:** Uses an LLM to evaluate the quality of summaries on a 1–5 scale (1 = worst, 5 = best) across four key dimensions:
  - **Fluency score:** Assesses grammar and readability
  - **Coverage score:** Evaluates inclusion of key information
  - **Coherence score:** Measures logical flow and structural clarity
  - **Faithfulness score:** Checks factual consistency with the original text
- **Results:** Llama 3 agents outperform both fine-tuned T5 and original T5 across all summary lengths
- **Pros:** Offers flexible, rubric-based scoring with human-like judgment across multiple aspects of quality
- **Cons:** May introduce variability or bias, and lacks full transparency in decision-making

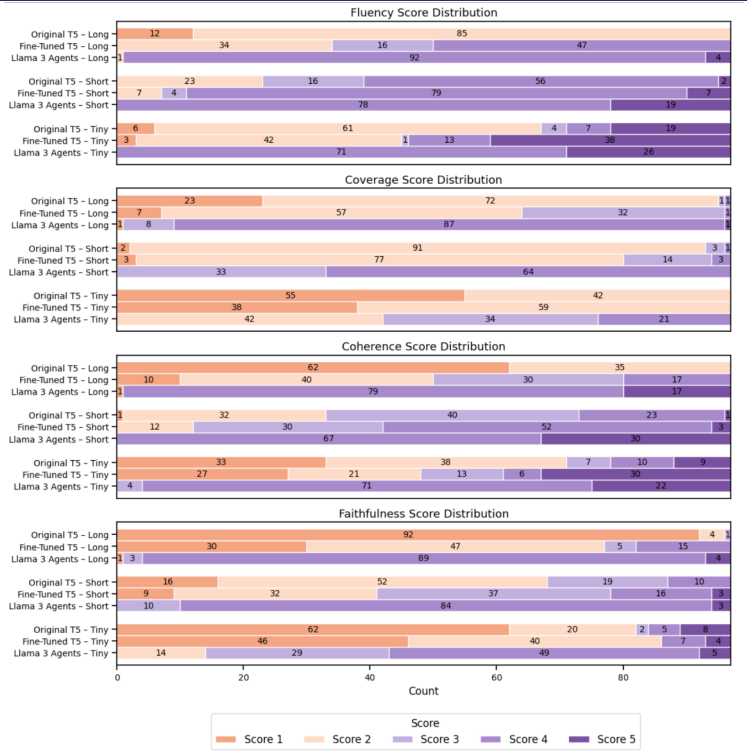


Model Evaluation and Comparison

- **Llama 3 Agents Significantly Outperform Other Models:**
  - Across all four evaluation metrics (fluency, coverage, coherence, faithfulness), Llama 3 Agents consistently achieve the highest mean and median scores
  - Statistical tests (Mann–Whitney U) show extremely low p-values, indicating the improvements of Llama 3 Agents over both Original T5 and Fine-Tuned T5 are highly significant

	Metric	Model A	Model B	B > A p-value	A mean	B mean	A median	B median	B > A significant
0	Fluency	Original T5	Fine-Tuned T5	1.057953e-17	2.656357	3.481100	2.0	4.0	True
1	Fluency	Original T5	Llama 3 Agents	1.068750e-55	2.656357	4.161512	2.0	4.0	True
2	Fluency	Fine-Tuned T5	Llama 3 Agents	2.764822e-15	3.481100	4.161512	4.0	4.0	True
3	Coverage	Original T5	Fine-Tuned T5	1.659052e-08	1.752577	2.020619	2.0	2.0	True
4	Coverage	Original T5	Llama 3 Agents	1.026022e-83	1.752577	3.446735	2.0	4.0	True
5	Coverage	Fine-Tuned T5	Llama 3 Agents	1.304788e-68	2.020619	3.446735	2.0	4.0	True
6	Coherence	Original T5	Fine-Tuned T5	3.017125e-16	2.161512	2.979381	2.0	3.0	True
7	Coherence	Original T5	Llama 3 Agents	1.055738e-76	2.161512	4.213058	2.0	4.0	True
8	Coherence	Fine-Tuned T5	Llama 3 Agents	1.326584e-40	2.979381	4.213058	3.0	4.0	True
9	Faithfulness	Original T5	Fine-Tuned T5	5.278079e-12	1.676976	2.185567	1.0	2.0	True
10	Faithfulness	Original T5	Llama 3 Agents	1.830298e-78	1.676976	3.790378	1.0	4.0	True
11	Faithfulness	Fine-Tuned T5	Llama 3 Agents	7.629166e-61	2.185567	3.790378	2.0	4.0	True

Pairwise Mann–Whitney U Test Results for LLM-as-a-Judge Scores



Webapp Demo

Audio Input: self-recorded audio of Google's VEO 3 from a news article

Actual Audio Text: On Tuesday announced Veo 3, an AI video generator that can also create and incorporate audio. The artificial intelligence tool competes with OpenAI's Sora video generator, but its ability to also incorporate audio into the video that it creates is a key distinction. The company said Veo 3 can incorporate audio that includes dialogue between characters as well as animal sounds."Veo 3 excels from text and image prompting to real-world physics and accurate lip syncing," Eli Collins, Google DeepMind product vice president, said in a blog Tuesday. The video-audio AI tool is available Tuesday to U.S. subscribers of Google's new $249.99 per month Ultra subscription plan, which is geared toward hardcore AI enthusiasts. Veo 3 will also be available for users of Google's Vertex AI enterprise platform.

Transcript: Google on Tuesday announced **VO3,** an AI video generator that can also create an incorporate audio. The artificial intelligence tool competes with open AI's **SOARA** video generator, but it's ability to also incorporate audio into the video that creates a key distinction. The company said **VO3** can incorporate audio that includes dialogue between characters as well as animal sounds. **VO3** excels from text and image prompting to real world physics and accurate lip syncing. Eli Collins, Google DeepMind product vice president said in the blog Tuesday. The video audio AI tool is available Tuesday to US subscribers of Google's new 249 and 99 cent dollars per month ultra subscription plan. Which geared towards hardcore AI enthusiasts. **Vi3** will also be available for users of Google's vertex AI enterprise platform.

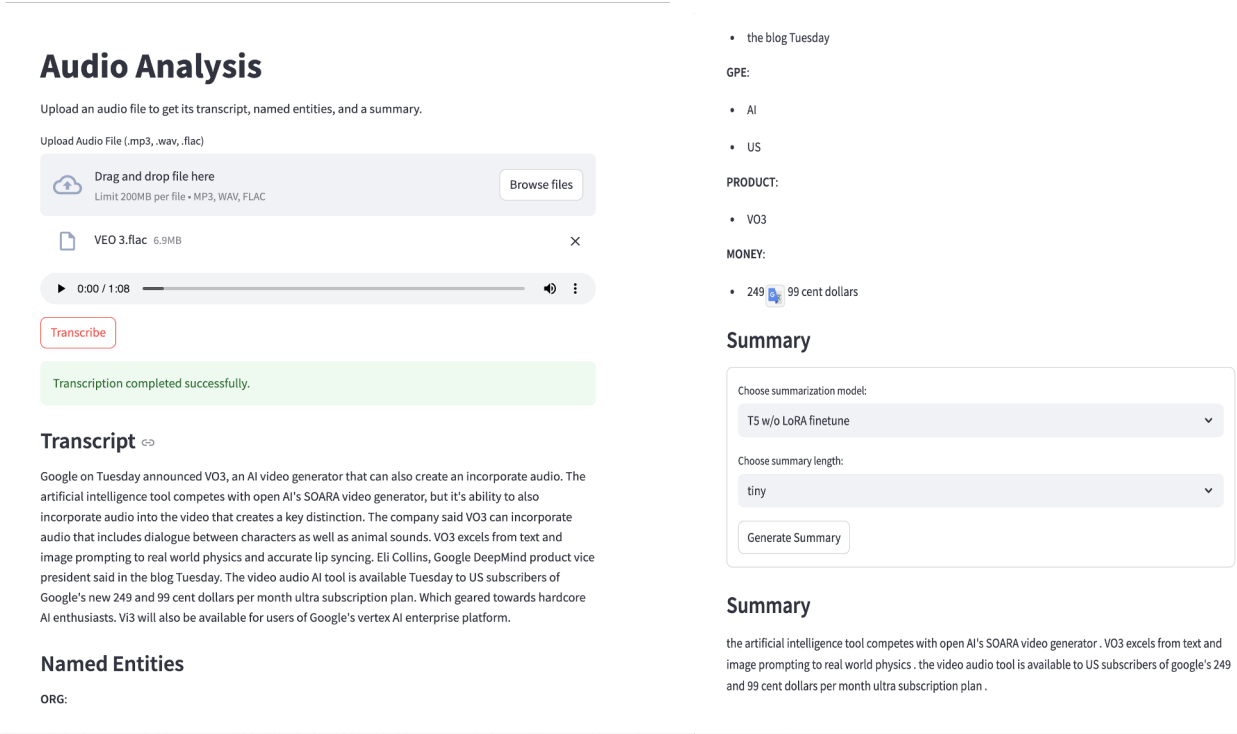
Summary (Short):

Original T5: the artificial intelligence tool competes with open AI's SOARA video generator . VO3 excels from text and image prompting to real world physics and accurate lip syncing . the video audio tool is available to US subscribers of google's 249 and 99 cent dollars per month ultra subscription plan .

Fine-Tuned T5: The artificial intelligence tool competes with open AI's SOARA video generator, but it's ability to incorporate audio into the video that creates a key distinction. The company said it can incorporate audio that includes dialogue between characters and animal sounds. It excels from text and image prompting to real world physics and accurate lip syncing.

Llama 3 Agents: Google announced VO3, an AI video generator that can also incorporate audio, distinguishing it from open AI's SOARA. VO3 can include dialogue between characters and animal sounds, exceling in text and image prompting, real-world physics, and accurate lip syncing. The tool is available to US subscribers of Google's 249−99/month Ultra plan, geared towards hardcore AI enthusiasts, and will also be available on Google's Vertex AI enterprise platform.

Webapp UI


Audio Analysis
Upload an audio file to get its transcript, named entities, and a summary.
Upload Audio File (.mp3, .wav, .flac)
Drag and drop file here
Limit 200MB per file • MP3, WAV, FLAC
Browse files
VEO 3.flac 6.9MB
0:00 / 1:08
Transcribe
Transcription completed successfully.
Transcript
Google on Tuesday announced VO3, an AI video generator that can also create an incorporate audio. The artificial intelligence tool competes with open AI's SOARA video generator, but it's ability to also incorporate audio into the video that creates a key distinction. The company said VO3 can incorporate audio that includes dialogue between characters as well as animal sounds. VO3 excels from text and image prompting to real world physics and accurate lip syncing. Eli Collins, Google DeepMind product vice president said in the blog Tuesday. The video audio AI tool is available Tuesday to US subscribers of Google's new 249 and 99 cent dollars per month ultra subscription plan. Which geared towards hardcore AI enthusiasts. Vi3 will also be available for users of Google's vertex AI enterprise platform.
Named Entities
ORG:
the blog Tuesday
GPE:
AI
US
PRODUCT:
VO3
MONEY:
249 99 cent dollars
Summary
Choose summarization model:
T5 w/o LoRA finetune
Choose summary length:
tiny
Generate Summary
Summary
the artificial intelligence tool competes with open AI's SOARA video generator . VO3 excels from text and image prompting to real world physics . the video audio tool is available to US subscribers of google's 249 and 99 cent dollars per month ultra subscription plan .

Northwestern | McCORMICK SCHOOL OF ENGINEERING

Thank You

Northwestern | McCORMICK SCHOOL OF ENGINEERING