

LLM / Prompt Engineering Study

# **Building an LLM app for transcribing and summarising user interviews in the life sciences domain**

Minjoon

## Prototype 만들기까지의 과정

AssemblyAI

pyannote

OpenAI



한정적인 무료체험

성능이 떨어지는 transcription

발화자 구분하기 (speaker diarization 불가능)

# Google Med-PaLM 의 발견



**minjoon** 17 Mar at 15:52

Stumbled across this LLM model from Google for the medical domain.  
It would be cool if this is more suited for our work. Anybody know if it is possible to get access to this via ebi?

<https://sites.research.google/med-palm/>

**sites.research.google**

**Med-PaLM: A Medical Large Language Model - Google Research**

Discover Med-PaLM, a large language model designed for medical purposes. See how we developed our AI system to accurately answer medical questions.

4 replies



**Santiago Insua** 17 Mar at 16:38

I will ask our Google Customer engineer, seems access is under request



**Alex** 18 Mar at 07:09

That model was published mid 2023 so is pretty old technology. There are much more effective models available.



**Santiago Insua** Tuesday at 15:15

Just to confirm, is preferably to use Gemini latest versions from the Google's point of view, as Alex said this has been deprecated in favour of Gemini.



1

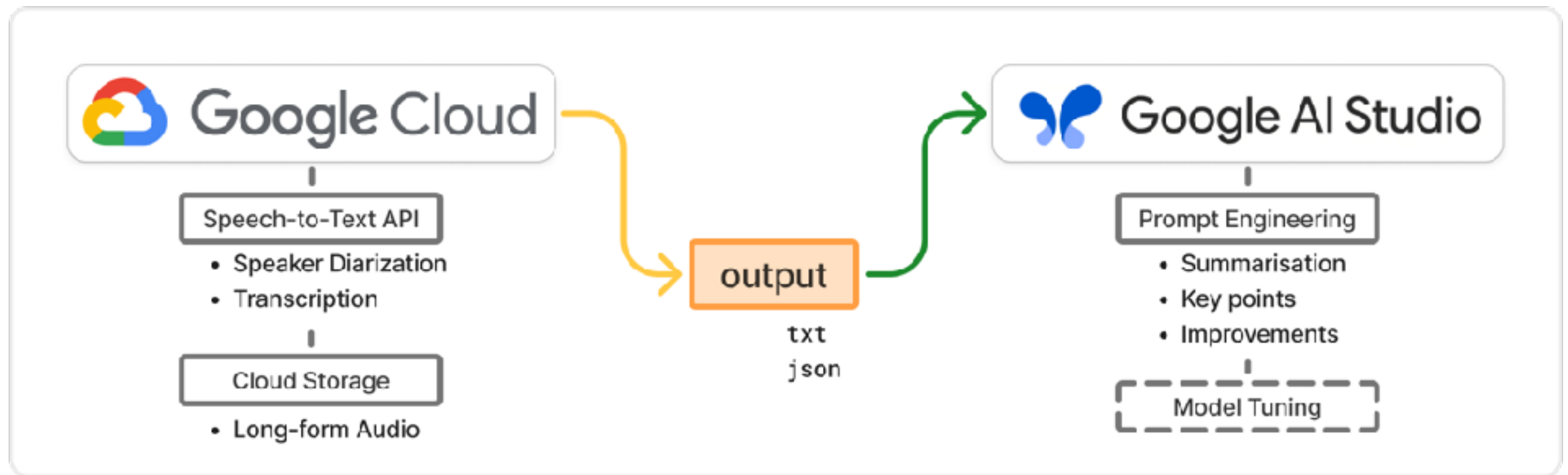


**minjoon** Tuesday at 15:17

Many thanks for checking!

기승전 Gemini

# Interview Transcriber + Summariser



```
# Load environment variables from .env file
load_dotenv()

def transcribe_with_diarization_from_gcs(gcs_uri, language_code):
    """Transcribes an audio file from Google Cloud Storage with speaker diarization"""
    client = speech.SpeechClient()

    audio = speech.RecognitionAudio(uri=gcs_uri)
    config = speech.RecognitionConfig(
        encoding=speech.RecognitionConfig.AudioEncodingMP3,
        sample_rate_hertz=16000, # Adjust based on audio file
        language_code=language_code,
        diarization_config=speech.SpeakerDiarizationConfig(
            enable_speaker_diarization=True,
            min_speaker_count=2,
            max_speaker_count=2,
        )
    )
```

transcription\_output\_from\_gcs.txt

Speaker 0: okay well first off thank you for your time and partly in the i officially start can you kind of explain what you do in terms of your rese your background and yeah yeah so I do both the special mix the big thing s techniques and try and map if you hold sell ones the biggest example of th part of my work I've been working on the ia version of lock it with a diff spectrometry acquisition method right so that's one thing I've been workin project though does have a biological goal as well I am interested in gene ridiculously the safety and toxicity associated with gene therapy so I'm l toxicity of aays aays yeah and how they affect the cell when they are bein transfusing themselves trying to map if with techniques in their lily love view of what's happening to all of the proteins in terms of their funds ch really interesting and I I should know I I am not I'm not a biologic perso understand evervthina but I did I did some buver for medics in underarad a

# Prompt

You are an expert UX researcher specializing in the biomedical domain. Please summarize the following interview transcript, focusing on the user's experiences, pain points, needs, and suggestions related to [mention the specific topic or product being discussed in the interview].

Identify the key themes and insights that emerged during the conversation. Pay close attention to any mentions of specific biological, medical, or bioinformatics terms and ensure they are accurately represented in the summary.

If multiple speakers were present (indicated by "Speaker 1:", "Speaker 2:", etc.), please attribute the key points to the correct speaker where possible.

The transcript is as follows:

[FULL TRANSCRIPT HERE]

Provide an outline of 3-5 paragraphs with main themes of the interview that captures the essence of the interview and highlights the most important information for UX research purposes.

Would prefer the outline to be structured with subsection titles.

## Prompt

**Role Definition:** *"You are an expert UX researcher..."*

- Helps the LLM adopt the appropriate persona and focus.

**Specific Focus:** *"...focusing on the user's experiences, pain points, needs, and suggestions related to [topic]..."*

- Guides the LLM to extract the most relevant information for UX research.

**“Key Themes and Insights”:** Instructs the LLM to identify overarching patterns and important findings.

**Domain Specificity:** *"Pay close attention to any mentions of specific biological, medical, or bioinformatics terms..."*

- Emphasises the need for accuracy in handling specialised vocabulary.

**Speaker Attribution:** *"If multiple speakers were present..."*

- Tells the LLM to consider speaker roles if the transcript includes diarization information.

**Desired Output Format and Length:** *"Provide a concise summary of approximately [length]..."*

- Gives the LLM guidance on the expected length and format of the summary.



# Output

Okay, here's a summary of the interview transcript, focusing on the user's experiences, pain points, needs, and suggestions related to UniProt and Music Maps.

## ✓ I. User Background and Research Focus

The user is a biomedical researcher employing high-throughput techniques, specifically LOPIT (Localization of Organelle Proteins by Isotope Tagging), with mass spectrometry, to map protein localization within cells. Their research has a significant biological goal: investigating the safety and toxicity of adeno-associated viruses (AAVs) in gene therapy. They aim to understand how AAVs affect cells at the protein level during transduction using techniques like LOPIT to gain an overall view of protein changes. The user's core workflow involves identifying proteins via mass spectrometry and then characterizing them to understand their function and behavior within the cellular context.

## ✓ II. UniProt Usage: Needs and Pain Points

UniProt is a crucial resource in the user's workflow. After identifying proteins through mass spectrometry, they use UniProt to quickly gain a summary of the protein's function and relevant literature. The most frequently used sections are the function summary, the sub-cellular location information (including associated publications), and details on sites and binding sites. The user values the linked publications, which provide easy access to validation data. The user compares their LOPIT localization data with UniProt's localization data to identify novel findings or validate existing knowledge. While the user appreciates the comprehensive information provided by UniProt, they mentioned that scrolling through large amounts of data can be time-consuming. A key suggestion was the ability to toggle sections on and off, allowing for a streamlined view focused on the information most relevant to the current task, such as function and sub-cellular location. The user uses mainly Swiss-Prot entries (the curated ones) but also comes across TrEMBL entries depending on the search.

## ✓ III. Music Maps: Potential Integration with UniProt

The user is familiar with Music Maps but hasn't used it extensively, as they find the necessary information readily available on UniProt, especially due to the PubMed links. The user suggested integrating Music Maps directly into the UniProt interface as a tab or section. The user expressed interest in using Music Maps if it could simplify complex data visualizations. Specifically, they proposed the ability to select two proteins and view only their related pathways or nodes side-by-side. The user also suggested a Music Maps version of the existing sub-cellular location panel in UniProt, with a similar layout for easy understanding and intuitive navigation. For instance, clicking on 'cytoplasm' in the Music Maps version would highlight or zoom in on the corresponding region or pathways within the cellular network. A consistent visual direction (e.g., top-down or left-right) in Music Maps was also suggested for easier comprehension.

## ✓ IV. Key Themes and Insights

- **Focus on Protein Function and Localization:** The user's primary interest lies in understanding protein function and spatial distribution within the cell, driven by their research on AAV toxicity in gene therapy.
- **UniProt as a Central Hub:** UniProt serves as a central resource for quick access to protein information, literature, and validation data.
- **Need for Streamlined Information Access:** The user desires a more efficient way to navigate and filter information within UniProt, particularly when working with multiple proteins.
- **Potential for Music Maps Integration:** Integrating Music Maps into UniProt could enhance data visualization and simplify complex pathways, but it needs to be intuitive and focused on relevant information like sub-cellular localization.
- **Importance of Contextual Information:** The user values the links to publications and the ability to compare their experimental data with curated information in UniProt.