
Speech-to-Text AI Assistant Application

Fangying Zhan
fzhan2@jhu.edu

1 **1 Motivation**

2 AI (Artificial Intelligence) translators have been considered particularly useful in giving prompt and
3 informative results. People are already getting used to popular translators such as the Translation app
4 on iPhones. In many cases, when people are watching live videos or attending live meetings, there is
5 an urgent need for speech-to-text translation. Well-developed tools that are available on the Internet
6 have enabled such functionalities by using models that are pre-trained on large datasets. However,
7 often cases that we would observe defects in the transcripts, since the quality of real-time translations
8 depends heavily on the speaker's accent, the noises, the contexts, etc. Therefore, it would be helpful
9 if we could develop an application that is based on the speech-to-text translation architecture and
10 has improved performance through adaptation and customization for the speaker. Such products
11 would be able to provide a better service than a normal AI translator, be more powerful with multiple
12 functionalities, and could substantially save the cost of a human translator.

13 To scale down the problem, we are interested in developing an application that can accurately
14 transcribe real-time audio or audio from a video clip to text and then identify and translate it correctly
15 into target languages. We would also enhance the translator with ChatGPT to become an assistant that
16 is interactive and capable of helping users understand the conversation from reading the transcripts.
17 By developing a speech-to-text multilingual translation application that can generate accurate subtitles,
18 we can make video content more accessible, inclusive, and engaging for a wider audience. This
19 technology can also help to save time and resources for content creators, who would otherwise need
20 to manually create subtitles for their videos.

21 **2 Related Work**

22 WhipserAI, which is a Transformer-based encoder-decoder pre-trained model for automatic speech
23 recognition (ASR) tasks¹. It supports multilingual transcriptions, but the performance varies widely
24 depending on the language. The figure 2 below shows a WER (Word Error Rate) breakdown by
25 languages of the Fleurs dataset using the large-v2 model. For the experiment of our application, we
26 apply it to tasks of transcribing YouTube videos for 6 representative languages: Spanish, English,
27 Japanese, Korean, Chinese, and Arabic.

28 M2M100, which is a multilingual encoder-decoder (seq-to-seq) model trained for Many-to-Many
29 multilingual translation. The LanguageLeapAI project targeted towards English to Japanese and
30 German specifically. Inspired by the project, this project generalizes it to multilingual scenarios. For
31 example, Korean-to-Mandarin. One possible experiment is to generate translation subtitles for the
32 YouTube video by snippets and then compare the translation result with the translation provided by
33 human translators. An alternative method is to use the DeepL API for multilingual text translation.
34 Another experiment could be to compare the results from different pre-trained translation models and
35 evaluate the state-of-the-art methods.

36 ChatGPT, an advanced language model developed by OpenAI, is specifically based on the GPT-3.5
37 architecture. It has been trained on a vast amount of text data from the internet, allowing it to
38 generate human-like responses and engage in natural, coherent conversations with users. As part
39 of the application, ChatGPT plays a crucial role in facilitating communication and understanding

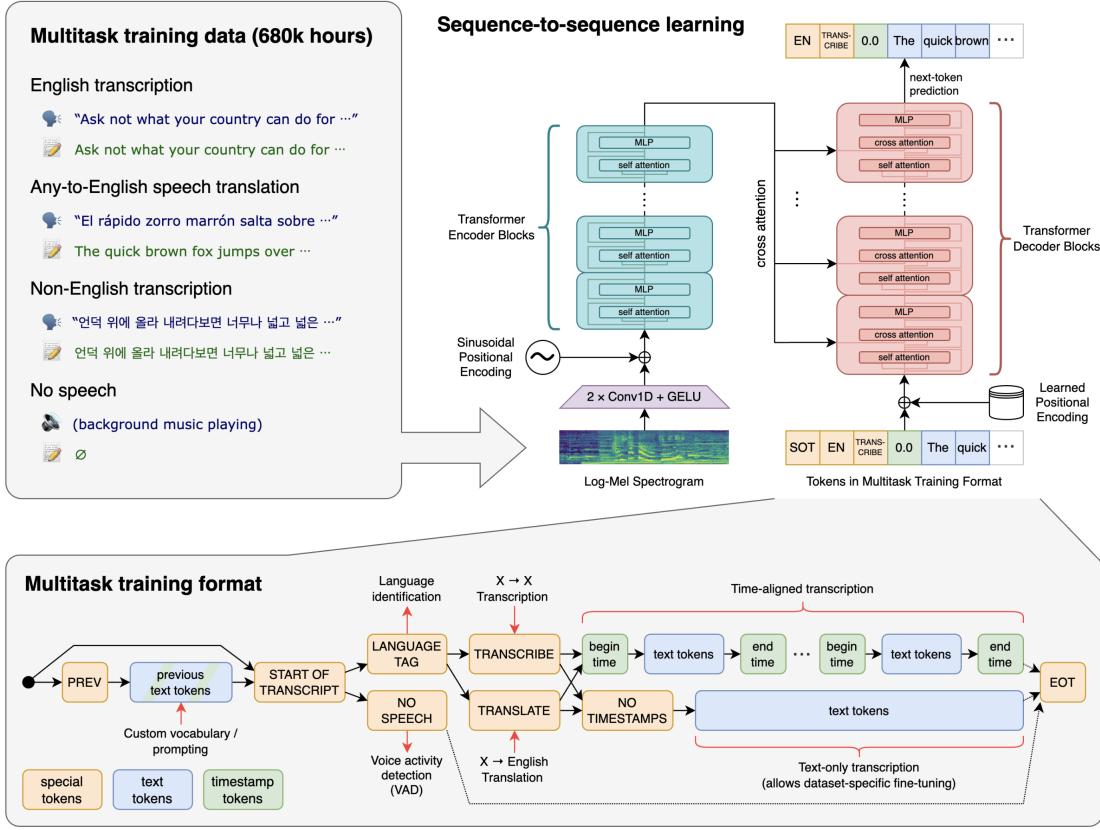


Figure 1: WhisperAI Pipeline

40 between users and the system. It leverages its language processing capabilities to interpret user
 41 queries, provide relevant information, and even provide improved versions of the translation.

42 3 Hypothesis

43 By leveraging the advancements in natural language processing and deep learning, the application is
 44 assumed to be able to accurately transcribe spoken language and generate text-based subtitles in real
 45 time. The application is designed for general purposes so that it recognizes different accents, dialects,
 46 and languages, thereby making video content accessible to a wider audience. An expected outcome
 47 would be that the final transcript from speech-to-text translation is readable, fluent, and accurate in
 48 many personalized details. It is also designed to support various formats of video content, including
 49 live streams, pre-recorded videos, and archived content. The customized Audio Translator ChatGPT
 50 would be a helpful assistant that answers questions about the transcribed and translated contents.

51 4 Application Design & Experiment

52 4.1 ASR Model

53 For the automatic speech recognition (ASR) task, we use the Whisper model by OpenAI. Whisper is
 54 a Transformer based encoder-decoder model, also known as the sequence-to-sequence model. It was
 55 pre-trained on 680k hours of labeled speech data annotated using large-scale weak supervision. The
 56 multilingual model enables both transcription and translation tasks. Though it would be interesting
 57 to incorporate other machine translation (MT) models like M2M100 into our application pipeline,
 58 the process of matching the inputs and outputs requires a lot of work with the consideration of
 59 tokenization. The most decent approach is to make use of the Whisper pipeline that includes both
 60 ASR and MT.

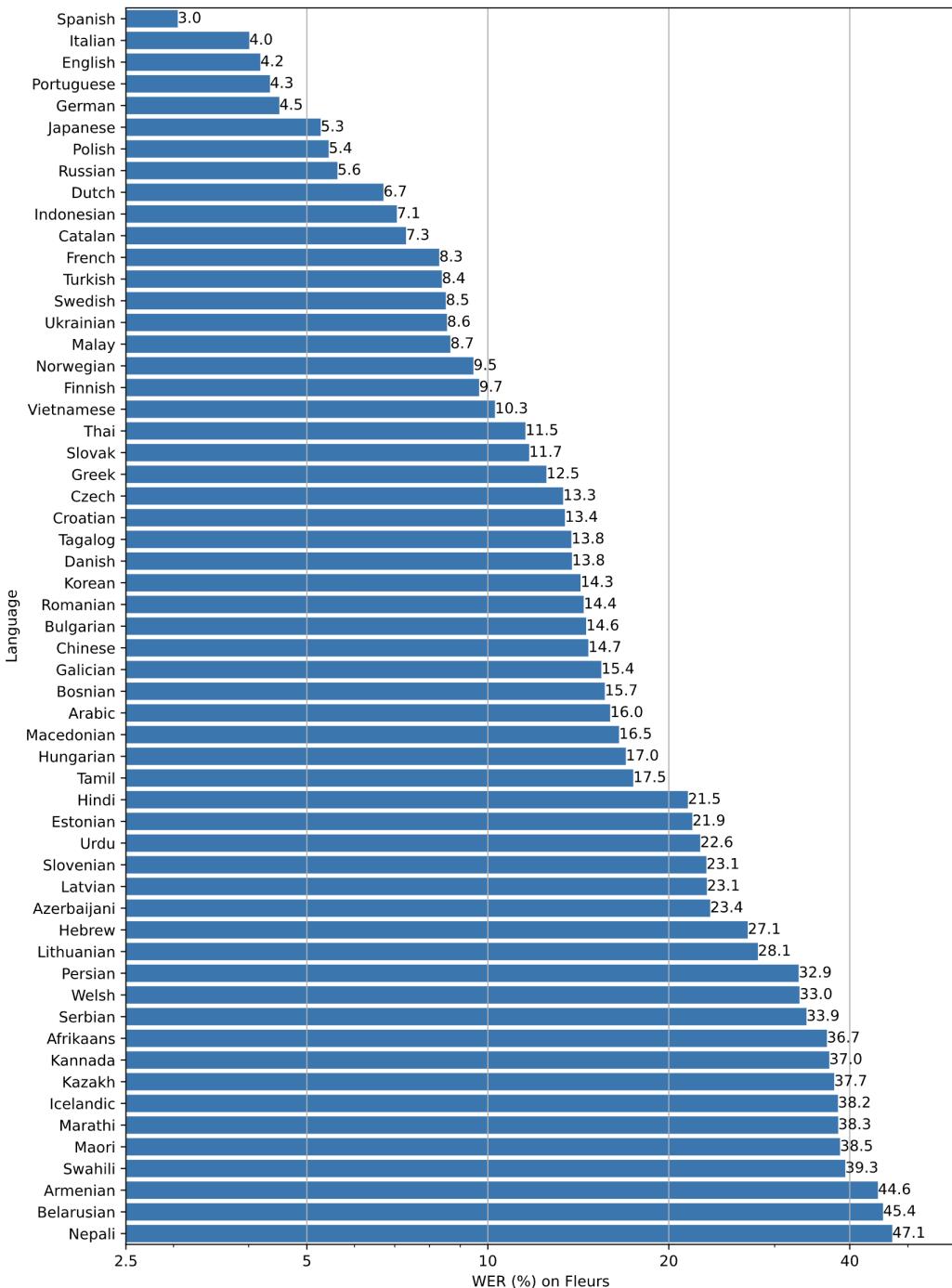


Figure 2: WhisperAI Performance: WER (Word Error Rate) breakdown

61 **4.2 LLM**

62 ChatGPT (GPT-3.5) is a variant of the large language model (LLM) GPT (Generative Pre-trained
63 Transformer) developed by OpenAI. The customized ChatGPT, integrated into the speech-to-text
64 translation AI assistant application, is designed to be easily accessible through API calls, which require
65 a user to obtain an OpenAI API key for personal usage. This API key serves as an authentication
66 mechanism to ensure authorized access to the model and maintain the security and privacy of the
67 system.

68 The benchmark is that after fine-tuning the model by training on a speaker's audio and transcription
69 data, evaluate the model on test audio cases by comparing the translation results with human
70 translation.

71 **4.3 Datasets**

72 To ensure that the speech translation is customized for speakers in real cases, datasets would be
73 the audio and transcriptions manually collected from videos on online platforms like YouTube. An
74 alternative would be to use publicly available datasets that could simulate real scenarios. For testing
75 our application, using YouTube videos with the length of 3 to 30 minutes would be sufficient.

76 **4.4 User Interface**

77 The user interface for the application is developed via Gradio, which can be embedded in Python
78 notebooks and presented as a webpage. It provides good interactions with machine learning models
79 specifically and is supported by Hugging Face Spaces with server support and easy sharing access.

80 **5 Result**

81 **5.1 Qualitative Outcome**

82 To gain a qualitative understanding of the application and the differences between the models,
83 we experiment with a randomly selected 6-minute and 20-second YouTube video titled "Movie
84 Dream Greetings on Stage" and present the results in Table 1, which showcases excerpts from the
85 transcription and translation outputs derived from the video, and Table 2, which showcases results
86 from querying ChatGPT. This selection serves as an evaluation sample obtained from the internet
87 specifically for our analysis.

88 An important question regarding the Whisper model, specifically the trade-off between application
89 performance and model runtime. The Whisper model offers various sizes, including tiny, base,
90 small, medium, large, and large-v2. It is observed that larger models tend to yield more accurate
91 transcriptions and translations, capturing finer details and reducing redundant words. However, it is
92 essential to consider that larger models also come with a runtime cost, typically taking 2 to 4 times
93 longer to execute.

94 For the ChatGPT assistant in this application, we choose the gpt-3.5-turbo, which is the most capable
95 GPT-3.5 model and optimized for chat at 1/10th the cost of text-davinci-003. It reads and comprehends
96 the transcription and translation results in real time, and functions as an assistant to answer questions
97 from user prompts regarding the content.

98 As demonstrated in Table 1 and 2, the trade-off becomes apparent when comparing the performance
99 of different model sizes. While larger models may enhance accuracy and offer more comprehensive
100 results, their increased runtime can impact the overall efficiency and responsiveness of the application.
101 For example, the large model result captures the movie title "Dream", whereas large-v2 does not
102 capture the movie title. Consequently, finding the optimal balance between model size and runtime is
103 crucial in achieving the desired application performance. For the best performance of our application,
104 after repeatedly testing with 30 YouTube videos and observing the consistency, we use the large-v2
105 Whisper model.

106 **5.2 Application Usage**

107 We encompass four key functionalities to address different user needs:

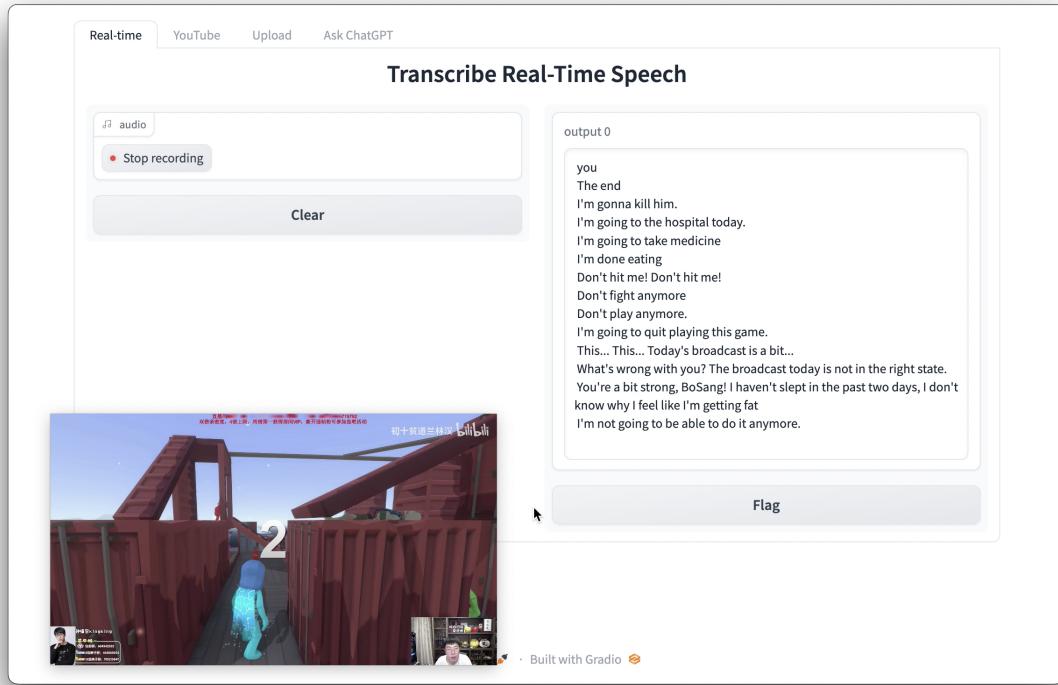


Figure 3: Real-time Transcription and Translation.

- 108 • **Transcribe Real-Time Speech:** The application allows users to perform real-time speech
109 transcription, converting spoken language into written text. Users can input their speech
110 through a microphone or other audio input devices, and the application will provide instantan-
111 eous transcription output.³
- 112 • **Transcribe YouTube Video via URL:** Users can provide a YouTube video URL to the
113 application, enabling transcription of the audio content within the video. This feature
114 allows for convenient transcription of videos for various purposes such as content analysis,
115 accessibility, or language learning.⁴
- 116 • **Transcribe File Uploaded:** The application supports the transcription of audio files that
117 users upload. By uploading an audio file in a supported format, users can obtain accurate
118 text transcriptions of the audio content.⁵
- 119 • **ChatGPT: Ask Anything:** The application incorporates the ChatGPT model, enabling users
120 to engage in interactive conversations and seek answers to their questions. ChatGPT utilizes
121 its vast knowledge base to provide human-like responses and assist users with inquiries or
122 discussions on a wide range of topics.^{6 7}

123 The code for this project can be found on this Repository ¹ and run on Google Colab. A demo video
124 can be found here ².

125 6 Discussion

126 Traditional applications were found only offer one of the four functionalities. Our Speech-to-Text
127 AI Assistant aggregates all functionalities into a single comprehensive platform and presents quick
128 responses, easy access, and informative results.

¹Project link: https://drive.google.com/drive/folders/1du4pITVZkC1IA60FBNUEMLCy0kArz0H0?usp=share_link

²Demo video link: https://drive.google.com/file/d/1LL18PUVLTxD1W8gCT0VqM812JVrwfcU/view?usp=share_link

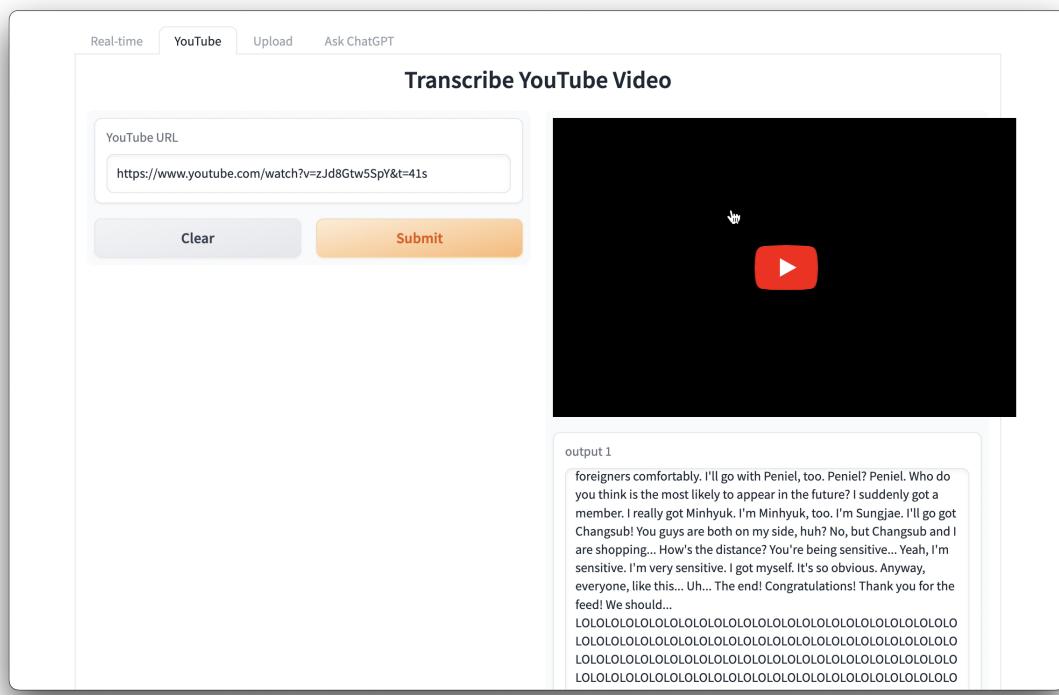


Figure 4: Video Transcription and Translation from YouTube URL.

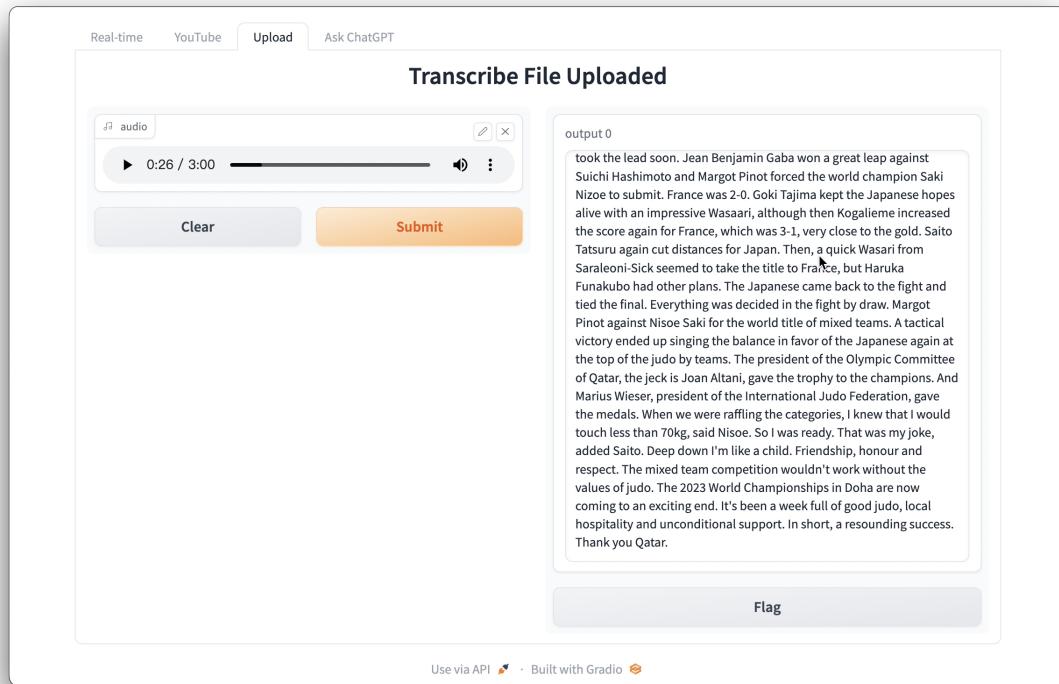


Figure 5: Video Transcription and Translation from User Upload.

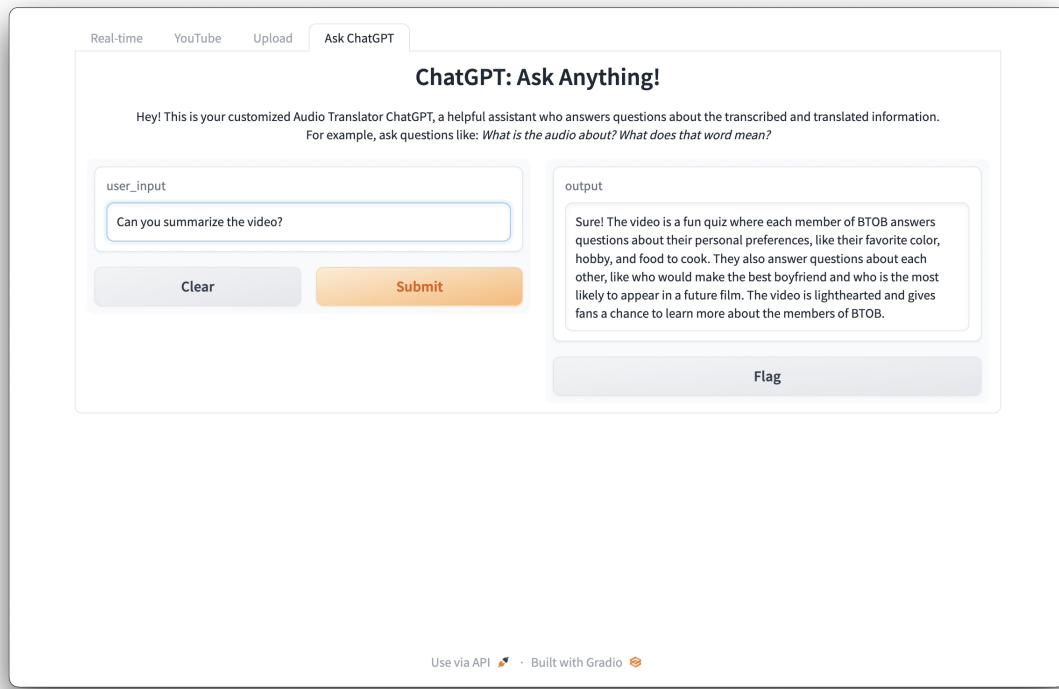


Figure 6: Customized ChatGPT Summarizing the Video.

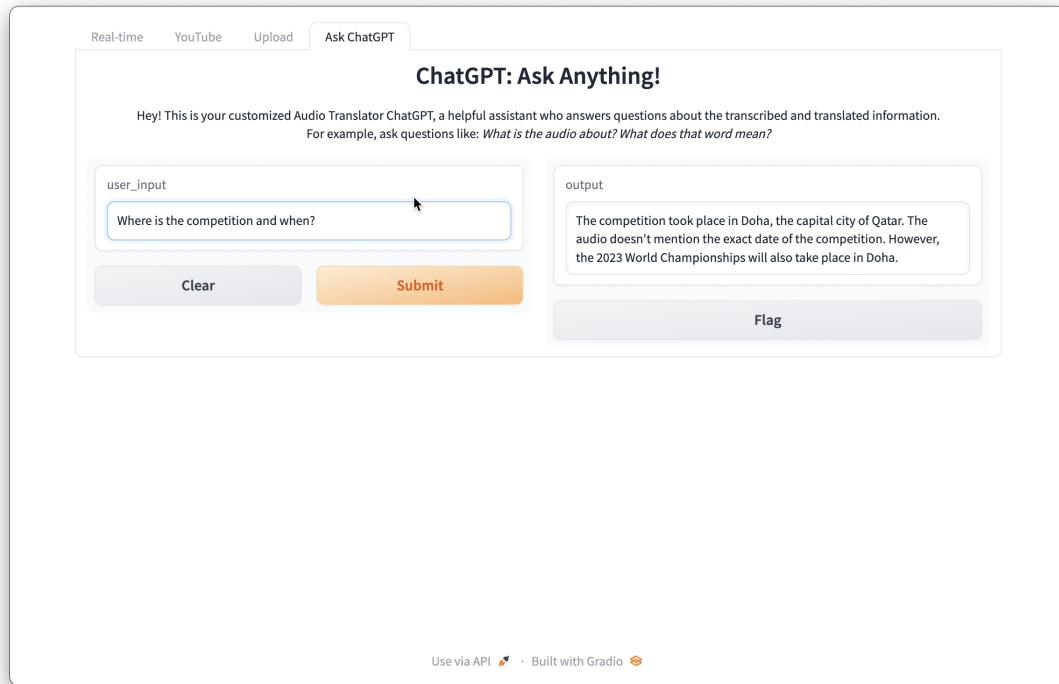


Figure 7: Customized ChatGPT Answering Questions about the Video.

129 Although the Whisper model is capable of performing multilingual machine transcription and
130 translation, changing the target language setting would require substantial effort to load the large-v2
131 model. Consequently, for the purpose of this presentation, we focus solely on the English version of
132 the application, while it is important to note that versions for other languages are also available.
133 For multilingual LLMs, the performance varies across different languages. The imbalance in
134 performance across languages is also found in the Whisper ASR model. Huang et al. (2023)
135 discussed the notion that not all languages receive equal treatment in LLMs. They propose a potential
136 solution in the form of a generic template prompt known as cross-lingual-thought prompting (XLT),
137 which aims to stimulate cross-lingual and logical reasoning skills. Future efforts should be directed
138 toward addressing and improving this issue.

139 References

- 140 [1] Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., & Wei, F. (2023). Not All Languages
141 Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. arXiv
142 preprint arXiv:2305.07004.
- 143 [2] LanguageLeapAI: The project integrates three AI systems to form a pipeline that pro-
144 vides a real-time multilingual translation: WhisperAI, DeepL Translator, and Voicevox.
145 <https://github.com/SociallyIneptWeeb/LanguageLeapAI>.
- 146 [3] Li, X., Wang, C., Tang, Y., Tran, C., Tang, Y., Pino, J.M., Baevski, A., Conneau, A., & Auli, M. (2020).
147 Multilingual Speech Translation from Efficient Finetuning of Pretrained Models. Annual Meeting of the
148 Association for Computational Linguistics.
- 149 [4] O'Regan, J. (2022). Continued finetuning as single speaker adaptation. TMH QPSR, 3. Retrieved from
150 <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-314269>
- 151 [5] Rabinovich, E., Mirkin, S., Nath Patel, R., Specia, L., and Wintner, S., Personalized Machine Translation:
152 Preserving Original Author Traits, arXiv e-prints, 2016. doi:10.48550/arXiv.1610.05461.
- 153 [6] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust
154 Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356. Retrieved
155 from <https://arxiv.org/abs/2212.04356>

Model	Runtime (second)	Output Sample
Tiny	18	Hello, we are HONGDAE, the CEO of the TV show. ... Woohoo! Woohoo! Woohoo! Woohoo! Woohoo! Woohoo! Woohoo! Woohoo! Thank you! Thank you so much! I hope you can continue to support us. Thank you! ... We'll talk about the comments and the comments. We will be able to give you a big comment and a big thank you for your support. Please look forward to it. Please look forward to it.
Base	23	Hello, I'm Park Seo-jin, the director of the GUSKLEEN. ... Thank you. Thank you. Thank you. Thank you. Thank you. Thank you. We will be doing a concert in Korea because of COVID-19. We will be doing a concert because of COVID-19. We will be doing a concert because of COVID-19. ... We are doing our best to make it clear. We are doing our best to make it clear. We are doing our best to make it clear. We are doing our best to make it clear. We are doing our best to make it clear.
Small	22	Hello, I'm Park Seo-jin, who will be working hard in the show. Nice to meet you. ... Thank you. Thank you. Thank you. We tried to film for 3 years because of COVID-19. It took a long time to finish the work and meet you. If you enjoy it, I have a picture to take. ... It's very refreshing, but let's express our gratitude to you. We prepared a little bit. Enjoy the movie. We will say goodbye at the end and wrap up.
Medium	37	Hello, I'm Seo-jin Park who will run hard in Hongdae on the screen soon. Nice to meet you. Thank you. ... Our movie was filmed for about 3 years because of the corona. It took a long time to complete the work after with you. If you turn on the air drop and quick share while we say goodbye, I think we'll have a short time to share them. Now, let's invite the actors and the director. ... I'm not strong enough. I'm not strong enough. I'm not strong enough. I'm not strong enough. ... It's a very small event, but we've prepared to express our gratitude to you. Enjoy the movie and we'll say goodbye and go home. Now, stand up and say goodbye.
Large	45	Hello, I'm Park Seo-jin who will be running around Hongdae soon. Nice to meet you. Thank you. We are on stage all day today. Thank you so much for filling the seats for 3 years because of the corona. It took a long time to finish the work and meet you. If you enjoy watching it, there is nothing more rewarding. ... You guys are giving us strength. We are not tired, right? We are not tired. We are so excited. It's all thanks to you. To repay the audience's love, we will take a group photo ... Yes, please come forward. Thank you. We really want to give it's a bit awkward. Thank you. I. Attention! Say goodbye!
Large-v2	47	Hello, I'm Park Seo Joon, who will soon be running around Hongdae. Nice to meet you. Thank you. We've been performing all day today. Thank you so much for filling up the seats and choosing our movie for the weekend. I'm here. Hello. Thank you. We filmed this movie for 3 years because of COVID-19. It took a long time to complete the movie and meet you. If you enjoy the movie, I'm sure you'll be very proud. ... We've been doing stage greetings since this morning, and the audience made me so happy that I'm not strong. You're not strong, are you? You guys are so energetic. You're not tired, right? Yes! You're not tired, you're all energetic. It's all thanks to you. To repay the fans' love, we'll take a group photo with the two of you. ... Shall we gather around? Thank you Fighting! We're also doing an event where we reply to comments. It's a very small event, but we've prepared it to express our gratitude to you. Enjoy the movie and we'll say goodbye and leave. Attention! Bow!

Table 1: Sentence outputs generated by different sizes of Whisper model.

Prompt: what did they talk about?

(tiny) The video appears to be a greeting from the hosts of a TV show called HONGDAE. They thanked their viewers for watching and promised to show them amazing stories. They also talked about the performance and concept of the show and asked their viewers to look forward to it. There were also some comments from viewers that they acknowledged and expressed their gratitude for the support.
(base) The video consists of several performers thanking the audience for tuning in and expressing their gratitude. They also mention how they are unable to perform a concert in person due to COVID-19 and instead share a bit of their performance through the video. There were also comments and hashtags shown on screen, which they appreciated and thanked the viewers for. Overall, it was a simple yet heartfelt message from the performers to their audience.
(small) In the video, the actors and director of a show are speaking to the audience after a stage greeting. They express their gratitude for being invited to the show and thank the audience for their support. They also mention that the show took a long time to film due to COVID-19. The actors then introduce themselves and thank their families and friends. They ask the audience to enjoy the show and continue to support them. They also mention a comment-related event and express their desire for the audience to be happy. The video ends with them saying goodbye and wrapping up.
(medium) The speakers mentioned that they were thankful for the audience who came to watch their movie, and expressed their hopes that everyone enjoyed the film. They also acknowledged that the filming process was long due to the COVID-19 pandemic, and encouraged viewers to share their thoughts on the movie with others using Air Drop and Quick Share. Finally, there were mentions of additional events and a small comment-reading event that the cast had prepared to show their gratitude towards their fans.
(large) The video is a stage greeting by the cast of the movie "DREAM" held in Hongdae for their audience. The actors and director thank the audience for filling the seats despite the COVID-19 situation, and express their joy to meet the audience after a long time. Each cast member introduces themselves and shares their gratitude to the audience. They also mention that they took a group photo together, which they are willing to share with the audience through AirDrop or Quick Share. They encourage the audience to watch the movie and recommend it to their friends, and express their hope that the audience will be happy. There are also some lighthearted moments and jokes shared by the actors.
(large-v2) Park Seo Joon and the other cast and crew members were greeting and thanking the audience for watching their movie. They mentioned that it took 3 years to complete the film because of COVID-19 and expressed their hope that the audience enjoys the movie. They also took some pictures that they wanted to share with the audience using AirDrop and quick-share. The actors and director introduced themselves and their characters, and thanked the audience for their support. They also encouraged the audience to continue supporting the movie and to be happy. IU mentioned that they had been doing stage greetings all day, and that the audience made them happy. They also announced an event to reply to comments as a way of expressing gratitude to their fans.

Table 2: Outputs generated by customized ChatGPT assistant.