An Open-Domain Avatar Chatbot by Exploiting a Large Language Model

Takato Yamazaki, Tomoya Mizumoto, Katsumasa Yoshikawa, Masaya Ohagi, Toshiki Kawamoto, Toshinori Sato

LINE Corporation

{takato.yamazaki, tomoya.mizumoto}@linecorp.com

Abstract

With the ambition to create avatars capable of human-level casual conversation, we developed an open-domain avatar chatbot, situated in a virtual reality environment, that employs a large language model (LLM). Introducing the LLM posed several challenges for multimodal integration, such as developing techniques to align diverse outputs and avatar control, as well as addressing the issue of slow generation speed. To address these challenges, we integrated various external modules into our system. Our system is based on the award-winning model from the Dialogue System Live Competition 5. Through this work, we hope to stimulate discussions within the research community about the potential and challenges of multimodal dialogue systems enhanced with LLMs.

1 Introduction

We present a demonstration of an open-domain avatar dialogue system that we have developed, with the goal of facilitating natural, human-like conversations. With the advent of large language model (LLM) technologies such as LLaMA (Touvron et al., 2023) and ChatGPT (OpenAI, 2022), the fluency of text-based dialogue systems has significantly improved. One of the next directions in this field involves dialogue systems that utilize voice, facial expressions, and gestures through an avatar, contributing to a more engaging and interactive conversation experience (Hyde et al., 2015).

As part of the efforts in dialogue system research, the Dialogue System Live Competition 5 (DSLC5) was held in Japan, a competition of avatar chat dialogue systems (Higashinaka et al., 2022). It was hosted within the academic conference of dialogue systems, where a large number of researchers evaluate the demonstrations performed on the stage to determine their ranking. We developed a dialogue system based on the LLM for this competition (Yamazaki et al., 2022), and encountered a variety of

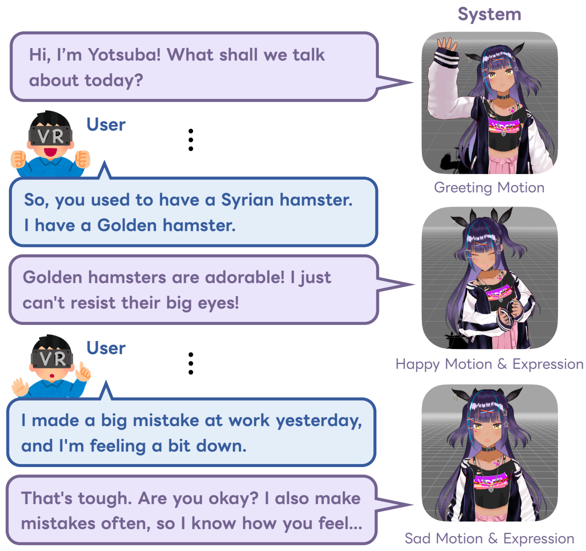

Figure 1: Sample interactions between users and our avatar chatbot, translated from Japanese. The avatar, accessible via VR headset or display, exhibits its emotions through motions and expressive facial cues.

challenges on integrating LLMs into a multimodal dialogue system. For instance, due to the real-time nature of spoken dialogue, it is essential to return some form of response quickly, which is a challenge when using computationally intensive LLMs. Furthermore, when the system involves an avatar, methods of controlling the avatar's facial expressions and motions present another challenge.

We strive to address such missing capabilities of an LLM-based dialogue system by integrating several external modules. Such modules encompass the incorporation of filler phrases and thinking motions during LLM's computational time, task parallelization to speed up responses, and detection of errors in speech recognition. Simultaneously, we paid close attention to the content of dialogue, aspiring to create a system that allows users to engage in deep, prolonged, and safe interactions. As a result, our system achieved the best human evaluation results in the competition. However, among

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 428–432
September 11–15, 2023. ©2023 Association for Computational Linguistics

the metrics, the naturalness of avatar's speaking style received the lowest score, indicating a need for improvement on its motions and expressions.

In this demonstration, we present an avatar dialogue system that improves from DSLC5. The improvement includes addition of emotion recognizer to enhance naturalness of the avatar expressions and motions. Additionally, aiming to provide a more immersive dialogue experience, we offer a system that allows conversation with an avatar through a virtual reality (VR) headset. Although the system is originally designed to respond in Japanese, we provide translated responses for English speakers. Through this demo, we hope to stimulate discussion within the research community about the potential and challenges of integrating LLMs into multimodal dialogue systems.

2 System Overview

We first provide a overview of the features of our proposed system, followed by a more detailed explanation in the subsequent subsections. The system architecture is illustrated in Figure 2.

Initially, the user's vocal input is transcribed into text utilizing a speech-to-text (STT) module. As the STT results frequently lack punctuation, a module called the *Punctuationizer* is utilized to append period marks and question marks. The punctuated text is then fed into the *Dialogue System*.

The Dialogue System leverages an LLM to generate responses. To elicit more engaging responses from the LLM, a process termed *Prompt Creation* is performed beforehand, providing the model with contextually rich information. Following the response generation, an *Editing & Filtering* phase is undertaken. During this phase, any responses that are dull or ethically inappropriate are identified and either edited or discarded as necessary.

Once the response text is determined, the system proceeds to control the avatar and text-to-speech system (TTS). The avatar's motions and expressions are decided based on the outcome of the Emotion Analyzer. To help accurately reading Kanji characters of Japanese, we add *Pronunciation Helper* to convert into the phonetic script of Hiragana.

2.1 Dialogue System: Prompt Creation

In the Prompt Creation phase, the system utilizes different types of few-shot prompts based on the given user inputs. All prompts are created based on a common template, which includes instructions such as the system character's profile, the current date and time, and the manner of speech. Here, we introduce a few of the prompts that we employ in our system.

STT Error Recovery Prompt There is a risk of LLM generating unintended responses when it receives user utterances containing STT errors. To mitigate this, we implemented an STT error detector based on fine-tuned BERT (Devlin et al., 2019) with dialogue breakdown detection data (Higashinaka et al., 2016). In cases where errors are detected, the system discards the user input and employs a prompt to inform the LLM about the inaudibility of the received utterance. Figure 3 shows an example shot of this prompt.

Knowledge-Response Prompt In cases where a user engages in a deep conversation on a specific topic, specialized knowledge or the latest information not included in the LLM's parameters might be required. Moreover, it is empirically known that the more niche a topic is, the more likely the LLM is to generate dull responses or only refer to well-known topics. Thus, satisfying users who wish to delve into more core conversations is challenging. To accommodate this, we introduced a search system for topic-related knowledge from online sources (e.g. Wikipedia) and inserts them into the prompt. This prompt is triggered when an effective knowledge source is found through searching the database.

Persona-Response Prompt The input length for the LLM has a limit, and it is currently difficult to incorporate all past dialogues as input. However, maintaining memory of past dialogues is crucial to achieving consistent conversations with the user. Our system is designed to maintain memories during dialogues by storing and utilizing the personas of the user and the system itself. After each utterance by both speakers, persona sentences are obtained using a *Persona Extractor* module which is also implemented with the LLM. These persona sentences are stored in a vector database and utilized in the prompt during the response generation.

2.2 Dialogue System: Editing & Filtering

While the responses generated by the LLM is fluent, they sometimes lead to dull or stagnant conversation, or even prematurely end the dialogue, resulting in a loss of user interest. To circumvent these issues, we have implemented a *Boring Re-*

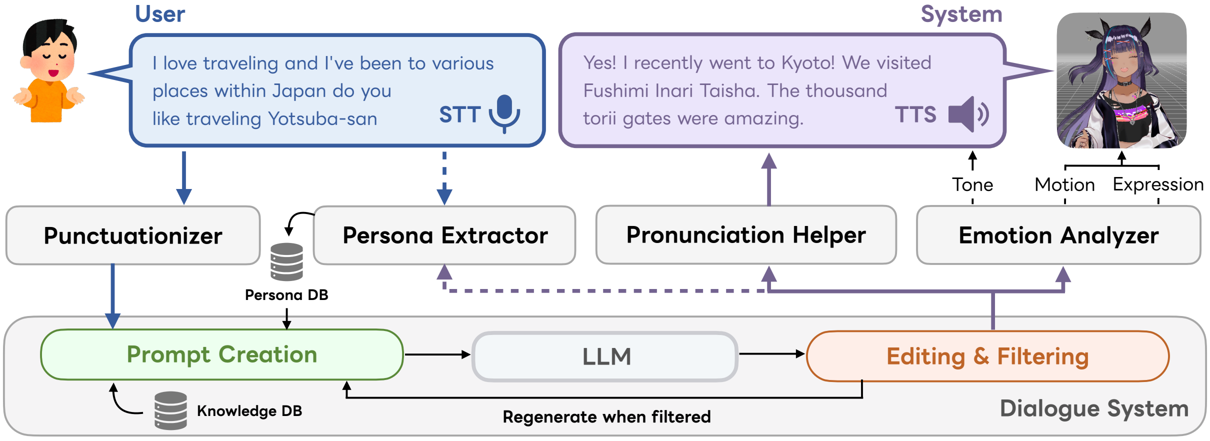

Figure 2: System Overview

```
System: What have you been interested in lately?
User: Well, I've actually started playing various
  games recently.
System: That's exciting! I'm a big fan of games
  myself. What kind of games are you into?
User: ...
System: (Hmm, I didn't quite catch the user's
  response...) Can you tell me about some other
  games you enjoy playing?
```

Figure 3: Example shot of an STT Error Recovery Prompt. The actual response is generated after the parentheses of the last utterance.

sponse Filter and a *Conversation-Closure Filter*. Both filters operate by identifying responses that are similar to manually collected boring and closing expressions. If either of these filters flags a response, the system will either add sentences using the LLM to enrich the content or revert back to the Prompt Creation stage for regeneration.

There's also a risk of generating ethically inappropriate responses, making it unsafe to directly provide the LLM's outputs to the user. In order to achieve communication that is both secure and respectful, we implemented *Toxic-Response Filter*. It utilizes a classifier that has been fine-tuned with BERT using a Japanese harmful expression dataset (Kobayashi et al., 2023). In case where these filters flag a response, the system adds suitable instructions to the prompt (e.g. "respond gently" or "expand on the topic") and again reverts back to the Prompt Creation to regenerate.

2.3 Response Timing

In spoken dialogues, promptly signaling understanding of the user's speech is considered crucial for enabling a comfortable conversation. However, the LLMs necessitate significant computational time, requiring approximately 2 seconds for each generation in case of our system. This overhead can lead to response delays, especially when regeneration is necessary.

To mitigate these potential sources of discomfort, our system employs concurrent operation of multiple modules. For instance, we execute persona extraction while the system is speaking, enabling it to expedite response times. Additionally, we have integrated the use of conversational fillers and animated motions during these waiting periods. By incorporating these elements, we aim to make the delay less noticeable and align with the natural flow of human conversation.

2.4 Emotion Analyzer and Avatar Control

In avatar dialogue systems, it is important to control the tone of the synthesized voice, as well as the avatar's facial expressions and motions, in accordance with the content of the utterance. Our system performs emotion analysis on the responses generated by the Dialogue System to manage these aspects. The analyzer employs a fine-tuned BERT trained on the WRIME dataset (Kajiwara et al., 2021). It recognizes eight types of emotions, namely joy, sadness, anticipation, surprise, anger, fear, disgust, and trust, on a strength scale ranging from 0 to 3.

Our in-house developed TTS system called *Coharis* can assign emotions such as "Happy" and "Sad", and these are mapped with the output of the emotion analyzer. Similarly, the expressions and motions of the avatar is also controlled based on the results of emotion analyzer. The avatar displayed in our demonstration is a sample model provided by the *VRoid Hub* [1] service from pixiv Inc. This sample model comes with a variety of facial expressions such as "happy," "sad," "angry," and so forth. Moreover, *Mixamo* [2] by Adobe Inc. offers a wide variety of avatar motions, which also is utilized for expressing emotions.

2.5 User Interface

The user interface operates on a web browser. We use the WebSpeechAPI of the browser for STT. The avatar is displayed by controlling WebGL through a library called three.js[3]. We also provide a VR interface using WebXR, allowing it to be displayed through a browser inside a VR headset.

3 Evaluation

We present the results of our original system obtained during the final round of DSLC5. It is important to note that the avatar and TTS used in the competition were provided by the organizers, and thus differ from the demonstration presented here. Furthermore, the evaluated system does not incorporate avatar control based on the emotion analyzer, so the evaluation results should be considered as reference values. For the LLM, we utilized HyperCLOVA 82B (Kim et al., 2021), which was trained on a Japanese corpus.

In the competition, the system engaged in a four-minute casual conversation in Japanese with an avatar about two topics, after which the dialogue was evaluated through a post-conversation survey. The survey was conducted based on four aspects: "Natural: whether the dialogue was natural", "Follow: whether the system could respond appropriately to the topic", "Providing: whether the system could provide new information related to the topic", and "Style: whether the way of speaking, demonstrated through voice, gestures, and facial expressions, was natural". These evaluation criteria were set by the competition organizers as metrics for assessing "human-level conversational ability". A five-point Likert scale was used for the evaluation.

Table 1 presents a comparison of our system with the second-ranked system, which is a dialogue system leveraging a 1B-params language model (Moriya et al., 2022). Our system significantly outperforms the second-ranked system in three metrics: Natural, Follow, and Providing, indicating high performance in dialogue content. As seen in the dialogue examples shown in Table 2, our system successfully follows and expands on topics. However, the Style score is notably lower than the other metrics, indicating the need for further enhancements in terms of avatar motions. We expect improvements with the Emotion Analyzer, as indicated by the displayed output results in Table 2. The emotion labels are accurately assigned, suggesting that they can be applied effectively to the avatar's facial expressions and motions.

	Natural	Follow	Providing	Style	Total
2nd	2.60	2.52	2.87	3.06	2.76
Ours	**4.00**	**4.00**	**3.89**	**3.54**	**3.86**

Table 1: Human evaluation results of the finals of DSLC5. "2nd" indicates the second place team.

Speaker	Utterance
User	Oh I see, you had a Djungarian hamster. I have a golden hamster.
System	Golden hamsters are also cute. I can't resist their big eyes. (Joy: 2)
User	Yes, their eyes are really cute.
System	They're so adorable, it makes me want to have one again. By the way, have you named it? (Joy: 1 / Ant.: 1)
User	Its name is Uchiko.
System	Oh, Uchiko, huh!? That's a cool name! How do your hamsters usually spend their time? (Joy: 1 / Sur.: 1)
User	They usually live in their cage.
System	I see, indeed, it would be full of dangers if they went outside. I wonder if they don't get stressed just staying inside the house. (Fear: 3)

Table 2: A sample dialogue translated from Japanese during the preliminary stage, with the output of the Emotion Analyzer in the parentheses. "Ant." and "Sur." corresponds to anticipation and surprise, respectively. The number represents the strength of the emotion (0-3).

4 Conclusion

In conclusion, we developed an open-domain avatar chatbot in a VR environment, leveraging a large language model (LLM). While encountering challenges in multimodal integration, such as addressing slow generation speed and controlling the avatar, our system demonstrated promising results in Dialogue System Live Competition 5. Addition-

[1] https://hub.vroid.com/
[2] https://www.mixamo.com/
[3] https://threejs.org/

ally, we attempted to improve the unnaturalness in the avatar's style of speaking, which was discovered after the competition, by using the emotion analyzer. We anticipate that this work will initiate meaningful discussions among the research community regarding the potential and challenges of integrating LLM into multimodal dialogue systems.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ryuichiro Higashinaka, Kotaro Funakoshi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets and evaluation metrics. In *Proc. of The Tenth International Conference on Language Resources and Evaluation*.

Ryuichiro Higashinaka, Tetsuro Takahashi, et al. 2022. Dialogue system live competition 5 (in japanese). In *JSAI SIG-SLUD, 96th Meeting*, page 19. The Japanese Society for Artificial Intelligence.

Jennifer Hyde, Elizabeth J Carter, et al. 2015. Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1719–1728.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2095–2104, Online. Association for Computational Linguistics.

Boseop Kim, HyoungSeok Kim, et al. 2021. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koga Kobayashi, Takato Yamazaki, et al. 2023. Proposal and evaluation of japanese harmful expression schema (in japanese). In *Proceedings of the 29th Annual Conference of the Association for Natural Language Processing*. Association for Natural Language Processing.

Shoji Moriya, Daiki Shiono, et al. 2022. aoba_v3 bot: A multi-modal chit-chat dialogue system integrating diverse response generation models and rule-based approaches (in japanese). In *JSAI SIG-SLUD, 96th Meeting*. Japanese Society for Artificial Intelligence.

OpenAI. 2022. Introducing ChatGPT. `https://openai.com/blog/chatgpt`.

Hugo Touvron, Thibaut Lavril, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Takato Yamazaki, Toshiki Kawamoto, et al. 2022. An open-domain spoken dialogue system using hyperclova (in japanese). In *JSAI SIG-SLUD, 96th Meeting*. Japanese Society for Artificial Intelligence.