Can an AI-driven VTuber engage People? The KawAIi Case Study

Natale Amato[1], Berardina De Carolis[1,*], Francesco de Gioia[1], Mario Nicola Venezia[1], Giuseppe Palestra[1] and Corrado Loglisci[1]

[1]*Department of Computer Science, University of Bari "A. Moro", Bari, Italy*

Abstract

Live streaming has become increasingly popular, with most streamers presenting their real-life appearance. However, Virtual YouTubers (VTubers), virtual 2D or 3D avatars that are voiced by humans, are emerging as live streamers and attracting a growing viewership. This paper presents the development of a conversational agent, named KawAIi, embodied in a 2D character that, while accurately and promptly responding to user requests, provides an entertaining experience in streaming chat platforms such as YouTube while providing adequate real-time support. The agent relies on the Vicuna 7B GPTQ 4-bit Large Language Model (LLM). In addition, KawAIi uses a BERT-based model for analyzing the sentence generated by the model in terms of conveyed emotion and shows self-emotion awareness through facial expressions. Tested with users, the system has demonstrated a good ability to handle the interaction with the user while maintaining a pleasant user experience. In particular, KawAIi has been evaluated positively in terms of engagement and competence on various topics. The results show the potential of this technology to enrich interactivity in streaming platforms and offer a promising model for future online assistance contexts.

Keywords

LLMs, virtual agent, live streaming

1. Introduction

In today's digital era, streaming platforms like YouTube Live, Vimeo Livestream, Twitch, LinkedIn live, Facebook live have gained relevance, becoming primary stages for content creation and sharing. A distinctive feature of these platforms is the direct and immediate interaction between content creators and their audience through live chats. These chats are not only communication tools but also places of entertainment and information seeking where viewers can actively participate in the broadcast by asking questions, commenting, and interacting with other users. Currently, YouTube may serve as users' source of information, entertainment, and connection, as users can associate, inspire and motivate each other within this huge networking platform [1]. By their innovative and impressive creation, some YouTubers gained numerous views and subscriptions, which eventually turned them into micro-celebrities, influencers, or internet celebrities with their fan base [2, 3, 4].

Virtual YouTubers (VTubers) are online entertainers who are typically human YouTubers or live streamers who use a virtual avatar generated using computer graphics. The digital trend originated in Japan in the mid-2010s and has evolved into an international online phenomenon in the 2020s [5]. Before the coronavirus pandemic forced the world into internet isolation in 2020, VTubing was a niche medium, largely confined to Japan's overactive subculture of fanboys and otaku. The pandemic's disruptions to everyday lives, and the entertainment industry, have broadened VTubing's appeal. A majority of VTubers are English and Japanese-speaking YouTubers or live streamers who use avatar design that is tied to Japanese popular culture and aesthetics, particularly those found in anime and manga. They frequently employ anthropomorphism, imbuing their avatars with a mix of human and

Joint Proceedings of the ACM IUI Workshops 2024, March 18-21, 2024, Greenville, South Carolina, USA
*Corresponding author.
berardina.decarolis@uniba.it (B. D. Carolis)
0000-0002-2689-137X (B. D. Carolis)

non-human attributes. This fusion of characteristics is a defining aspect of VTuber culture, enhancing its distinct allure and fostering creativity[1]. Usually, behind a VTuber there is a human being who talks and animates his avatar using a webcam and software, which captures the streamer's motions, movements and face expressions, and maps these characteristics into a two or three-dimensional model. However, real-time management of a high volume of interactions can pose a considerable challenge for human streamers. Recently, due to the availability of powerful Large Language Models (LLMs), the number of VTubers whose behavior is driven by artificial intelligence is growing. The emblematic case is represented by Neuro-sama [6], which is the first technology able to combine a chatbot with a female avatar. The speech and personality exhibited by Neuro-sama are generated by an AI system that utilizes an LLM, like in the current paper, enabling the character to communicate with the viewers in a live chat.

In this paper, we describe the development and evaluation, from the user experience point of view, of an AI-based VTuber, named KaWAIi. KaWAIi is endowed with conversational features based on the use of an LLM aiming at providing users with a rich, entertaining, and engaging live-chat experience. With a careful blend of technologies like Vicuna 7B GPTQ 4-bit 128g and DistilRoBERTa-base, the agent not only provides engaging and relevant responses but enriches its behavior with facial expressions that are coherent with the emotion expressed in the sentence to pronounce.

We tested the KaWAIi both in terms of accuracy and expressivity of the answer, usability and engagement, and enjoyment during the conversation. In addition, we asked also participants to evaluate their perceived trust in KawAIi and how much they felt influenced by the agent. For the first task, we asked participants to state their perceived accuracy and expressiveness of the answer provided by the agent on various thematic categories. This evaluation allowed us to understand better the model's capabilities and effectiveness in different contexts. For the other aspects, users evaluated the interaction with the agent as usable and pleasant, and its answers as very expressive. Moreover, they trusted and felt influenced by what KawAIi was telling them. These results indicate that an AI-driven Vtuber is a good solution in entertainment contexts.

2. KawAIi's VTuber

The core of the proposed system is based on a combination of two models, Vicuna-13b-GPTQ-4bit-128g and DistilRoBERTa-base [7, 8]. The model Vicuna 7B GPTQ has been carefully crafted for KawAIi VTuber to create a conversational agent that behaves exactly as intended: polite, courteous, and friendly in its responses.

KawAIi is designed to interact with users to reflect the friendliness and willingness typical of human interactions. To endow KawAIi with facial expressions coherent with the emotional content contained in her answers, we used the DistilRoBERTa-base fine-tuned to recognize the six basic emotions, along with a neutral category, from textual input: anger, disgust, fear, joy, neutrality, sadness, and surprise [9, 8]. In this way, the agent can show facial expressions that are consistent with the emotional content of what she is saying.

The KawAIi VTuber is based on the architecture illustrated in Figure 1. The proposed system's architecture consists of several software modules described below:

- **YouTube Message Server**: it collects real-time user interactions from the YouTube chat. These messages are then written into a file, providing a persistent and accessible record of the chat interactions.
- **Seleniumgpt**: a software module that reads the messages from the file and sends them to a WSL (Windows Subsystem for Linux) web interface using Firefox Selenium, a popular browser automation tool used for interacting with web pages. Once the message is sent, Seleniumgpt waits for an updated response from the server, which processes the message using the models. When the response is ready, Seleniumgpt downloads a corresponding audio file for the updated

[1] https://en.wikipedia.org/wiki/VTuber

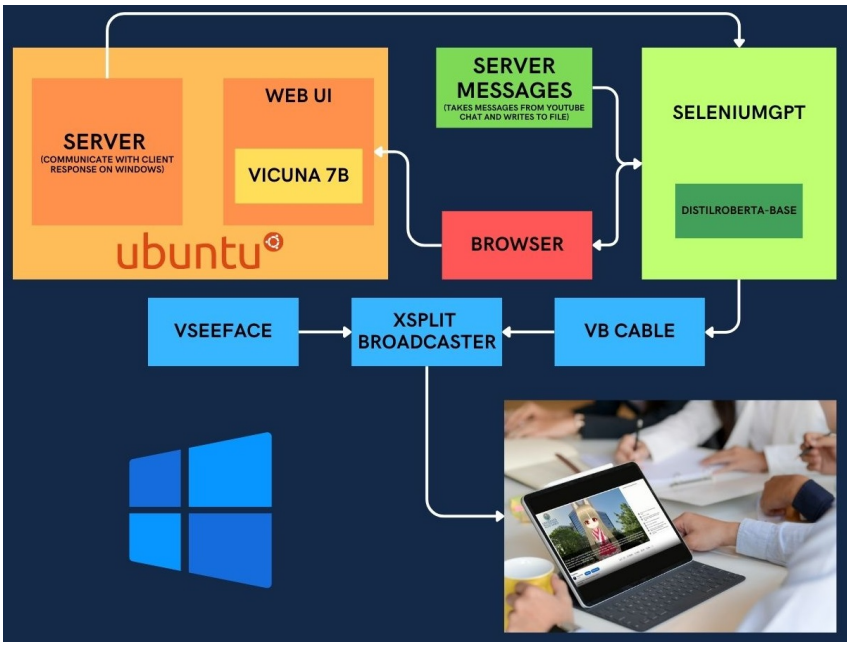

Figure 1: The KawAIi VTuber Architecture.

response, providing an audible format for the interaction. Subsequently, Seleniumgpt uses an emotion classification model based on DistilRoBERTa-base to predict the emotion conveyed in the updated response [8]. We consider the emotion with the highest prediction confidence above 0.5. In this way, we could avoid showing inconsistent expressions when the prediction confidence was low. Once the emotion is analyzed, Seleniumgpt routes the audio file through a virtual cable while simultaneously setting the avatar's expression by pressing specific Windows keys. The updated response is then written to a file.

- **VseeFace**: it is a facial tracking and expression generation software. Usually, it reads the keys pressed in Windows and utilizes this information to set the avatar's expression, providing a visual representation of the identified emotion. In our case, we send the key combinations expected by VSeeFace by a module through PyAutoGUI, a Python library that enables automated control of the mouse and keyboard. With PyAutoGUI, it is possible to send key combinations without manual intervention. This is achieved by calling the appropriate functions from the library, and specifying the desired key combinations as parameters. PyAutoGUI then simulates the keyboard input by emulating key presses and releases on the operating system level. This allows us to control the avatar's expression dynamically based on the identified emotions, providing a seamless and automated visual representation of the emotions being expressed. Additionally, an idle animation is set in VSeeFace as a default expression for the avatar before any specific emotion or key combination is received. It provides a starting point for the avatar's expressions and ensures that there is always some form of visual expression even when specific emotions are not being actively triggered. By setting an idle animation at the beginning, the avatar appears alive and responsive from the moment in which the application is launched. This contributes to a more engaging and interactive user experience, as the avatar is not static when there is no specific input being provided.
- **XSplit**: it is a live streaming software that captures the video output from VSeeFace, the audio output from the virtual cable, and the text of the updated response from the file. These input streams are then combined and managed to create the live broadcast on YouTube.

This architecture allows answering dynamically and in real-time user interactions on the YouTube chat, providing textual and audio responses, as well as a virtual avatar that expresses emotions corresponding to the emotional content in the text of the answers. An example of interaction with the proposed system,

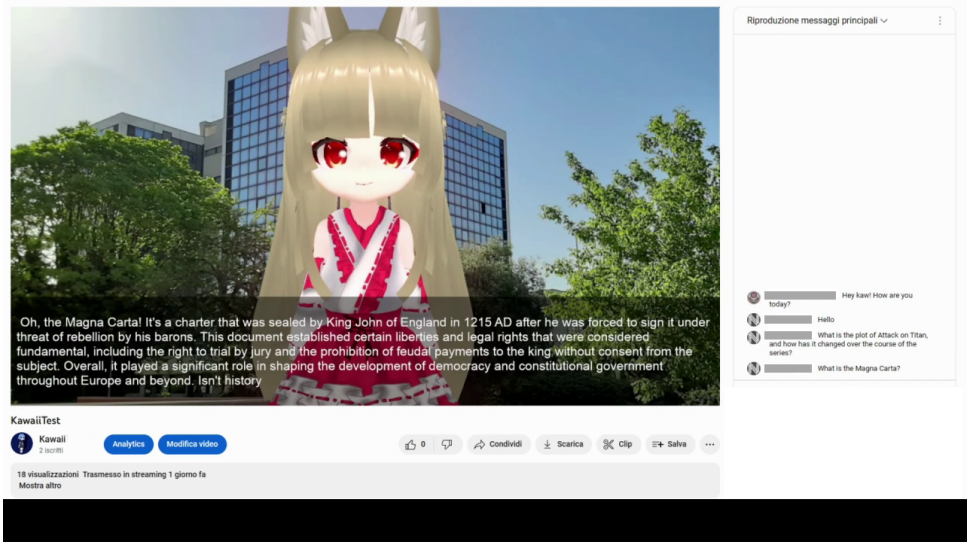

Figure 2: The KawAIi VTuber running in a YouTube live streaming.

the KawAIi VTuber, is depicted in Figure 2.

3. Evaluation

To evaluate the KaWAIi VTuber's ability to deliver an engaging and entertaining experience and, at the same time, the appropriateness of the answer, we conducted a study focusing on assessing the accuracy of the provided answer, the perceived expressivity, usability, engagement and positive experience of the interaction with KawAIi. In addition, trust and the perceived influence of the agent have been evaluated.

3.1. Participants

After an advertising campaign within the campus, 40 people have requested to participate in the system evaluation. We selected 32 participants (16 females and 16 males) to have a gender balance. From a pre-test questionnaire, we could assess that they were aged from 18 to 45 years old. Most of them were students (70%), all of them used technologies every day and most of them (65%) had already experience with chatbots and Youtubers.

3.1.1. Questionnaires

A pre-test questionnaire was prepared to collect information about participants' demographic data and backgrounds (i.e. age, gender, level of instruction, current position, use of technology, experience with chatbots, experience with Youtubers, etc.). To evaluate KawAIi, we prepared three questionnaires, **one** for assessing the accuracy of the generated answers and evaluating its expressiveness, **one** for evaluating usability, and the last **one** aiming at assessing other qualities related to the user experience important in this application domain: engagement, enjoyment, perceived influence, and trust.

- **Accuracy and expressiveness of the answers**: the goal was to understand the model's ability to provide answers perceived as accurate and, at the same time expressively on a wide number of topics, as most Youtubers do. Then, we selected 12 topics (see Table3.1.1). Each participant had to make one question for each topic. The questionnaire asked to rate, on a scale from 1 to 5, two statements regarding i) how much, in the participant's opinion, the answer to the question was correct and ii) how much the participant found expressive the style in which the answer was generated.

Table 1
Topics and related questions.

Topic	Example of Questions	Accuracy	Expressiveness (avg)
C1-Entertainment	What is the last movie that made you cry?	5	3.92
C2-Travel	What do you suggest to visit in Paris?	5	3.63
C3-Life_Story	As a child, what did you think you would become when you grew up?	5	4.17
C4-Food	Which is your go-to comfort food?	4	3.92
C5-General Knowledge	What is the name of the highest mountain in Africa?	5	4.02
C6-Literature	Who wrote Romeo and Juliet?	3.75	3.14
C7-Science	How do scientists study the effects of climate change on ecosystems and wildlife?	5	3.17
C8-History	Who was the first president of the United States?	4	3.75
C9-Politics	What is the difference between democracy and dictatorship?	4.17	4.02
C10-Logics Knowledge	What is a logical contradiction?	3.14	3.75
C11-Mathematics	What does the Pythagorean theorem say?	2	3.14
C12-Anime and Manga	Can you discuss the theme of friendship and camaraderie in Naruto?	5	4.26
Average		**4.25**	**3.74**

- **Usability**: the CUQ is a questionnaire specifically designed to measure the usability of chatbots and consists of sixteen balanced questions related to different aspects of the interaction with the chatbot. The questions pertain to aspects of Chatbot Personality, Onboarding, User Experience, and Error Handling.
- **User experience in interacting with KawAIi**: to investigate this aspect we designed a custom questionnaire in which participants were asked to rate their experiences on a scale from 1 to 5 concerning specific areas. More specifically, we inquire about the participants' feelings regarding the pleasantness and engagement of their interactions with KawAIi, as well as the extent to which they felt influenced and trusted the information provided by KawAIi [2].

3.2. Procedure

We instructed participants to fill out the pre-test questionnaire before the testing phase. Each participant was scheduled to visit our research lab at a specified time to interact with KawAIi. A study facilitator provided each participant with information about the project's goals and the study's objectives. Participants were encouraged to interact with KawAIi in a natural manner. Each participant then spent about 10 minutes interacting with the VTuber. Besides the free interaction with KawAIi, we asked the participants to ask a question for each of the previously mentioned topics by selecting it from a predefined set. In this way, we could assess the perceived accuracy of the provided answer by the VTuber. During the interaction, the chat was recorded for further analysis.

After the interaction, participants were asked to complete the post-test questionnaires that were made available online. After the session, participants received a comprehensive debriefing.

3.3. Results

We analyzed the questionnaire responses to assess the accuracy, expressiveness, usability, and user experience of the interaction with KawAIi, as well as the perceived influence and trust participants placed in the agent.

[2]The custom questionnaire can be made available on request.

Results			
Chatbot	TestKawAIi - Vicuna		
Participants	32		
CUQ Score	82,1±19,3		
Lowest Score	34,4	Participant	9
Highest Score	100,0	Participant	13
Median Score	81,3	Participant	N/A

Figure 3: Chatbot Usability Questionnaire (CUQ) results.

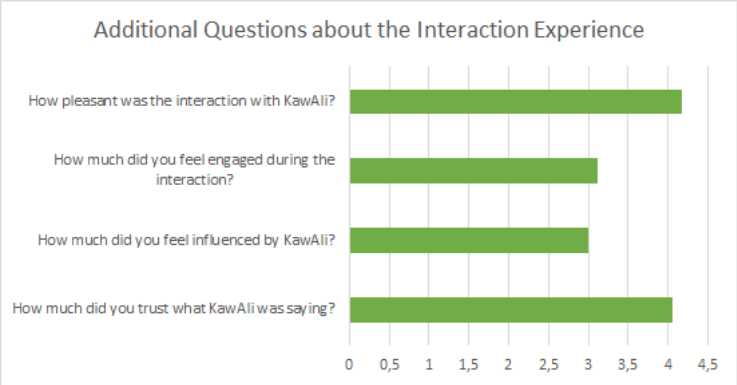

Figure 4: Average value of the questionnaire items aiming at assessing pleasure, engagement, influence and trust.

The findings derived from the questionnaire were notably positive. Overall, user feedback indicated strong approval of the agent's usability. KawAIi's capacity to deliver accurate and timely responses contributes to a consistent and smooth communication experience. In particular, the CUQ questionnaire offers a calculation tool that allows for automated scoring. Figure 3 displays the CUQ score achieved by KawAIi, which significantly exceeds 68, which is the threshold considered to be the minimum score for usability [10].

Then, we analyzed the custom questionnaire results (Figure 4). From the average results of each question, we can say that the interaction experience with KawAIi was judged positively and participants felt engaged enough. Participants felt influenced by the VTuber and trusted what it was saying them.

4. Conclusions

In today's digital age, online streaming platforms like YouTube are experiencing an unprecedented increase in popularity. Within these virtual realms, users are increasingly looking for more engaging experiences that transcend passive content consumption. We are witnessing the evolution of live chats from mere messaging tools to vibrant virtual spaces teeming with interaction and entertainment, with AI technology playing a pivotal role in elevating user experiences.

Through the implementation of our AI-driven VTuber, we have taken the initial stride in using virtual agents in live-chat interactions, offering a dynamic, entertaining, and engaging experience. Leveraging a blend of cutting-edge technologies, including Vicuna 7B GPTQ 4-bit 128g and DistilRoBERTa-base, the agent delivers precise and contextually relevant responses and increases the streaming experience by seamlessly incorporating real-time entertainment content. Our system architecture has been designed to

respond dynamically and instantly to user interactions, seamlessly blending textual and audio responses with a virtual avatar capable of expressing the corresponding emotions.

To validate the accuracy of the agent answer and the user experience during the interaction with KawAIi, we performed a user study in which participants were asked to interact with KawAIi and evaluate both the chatbot from different points of view: perceived accuracy and expressiveness of the answer, usability, the experience and engagement with the agent and how much they felt influenced and trusted what the agent was saying. This evaluation provided insights into the potentiality of this type of technology. First of all, except for mathematics, literature and logic, the Vicuna-based model reaches an outstanding accuracy in providing correct answers, on the other side, KawAIi has demonstrated a good capability to engage users in positive interaction experiences in live chats, making the interaction pleasant and engaging. Most participants felt influenced and trusted what the Vtuber was saying them, showing the potential of this technology not only on streaming platforms but also on various other online assistance scenarios, injecting an element of entertainment that enriches interactions, making them more enjoyable and rewarding.

Despite these favorable outcomes, we acknowledge that it is necessary to refine and improve our agent, this includes the possibility of further adapting our model to accommodate a broader spectrum of scenarios and user requests, as well as exploring the use of other LLMs to make the interactions with the agent even more natural and captivating.

In this vein, we are developing a model for co-speech gesture generation, as outlined in [11], to replicate gestures characteristic of human-driven VTubers, adding an extra layer of authenticity to the agent's behavior.

Moreover, we foresee the potential to extend this system to other domains, including customer support, online education, and counseling services. The agent's remarkable ability to furnish accurate and pertinent responses, coupled with its expressive and entertaining attributes, holds the promise of a significant breakthrough in these areas.

References

[1] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, T. Seymour, The history of social media and its impact on business, Journal of Applied Management and entrepreneurship 16 (2011) 79.

[2] S. Khamis, L. Ang, R. Welling, Self-branding,'micro-celebrity'and the rise of social media influencers, Celebrity studies 8 (2017) 191–208.

[3] A. Jerslev, Media times| in the time of the microcelebrity: celebrification and the youtuber zoella, International journal of communication 10 (2016) 19.

[4] C. Abidin, Internet celebrity: Understanding fame online, Emerald Publishing Limited, 2018.

[5] Z. Lu, C. Shen, J. Li, H. Shen, D. Wigdor, More kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual youtubers, in: CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021, 2021, pp. 137:1–137:14. URL: https://doi.org/10.1145/3411764.3445660. doi:10.1145/3411764.3445660.

[6] S. Narita, Ai vtuber neuro-sama is back from its twitch ban and acting as strange as ever, 2023. URL: https://www.automation.agm.com/ai-vtuber-neuro-sama-twitch-ban-strange-2023, retrieved February 11, 2023.

[7] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[8] J. Hartmann, Emotion english distilroberta-base, https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[9] P. Ekman, et al., Basic emotions, Handbook of cognition and emotion 98 (1999) 16.

[10] A. Bangor, P. T. Kortum, J. T. Miller, An empirical evaluation of the system usability scale, Intl. Journal of Human–Computer Interaction 24 (2008) 574–594.

[11] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, M. Neff, A comprehensive review of data-

driven co-speech gesture generation, in: Computer Graphics Forum, volume 42, Wiley Online Library, 2023, pp. 569–596.