



# Emotional Features of Short Conversations between a Generative AI-Controlled Virtual Communication Trainer and Four Children

Kenji Ito,<sup>1</sup> Tsubasa Abe,<sup>1</sup> Masanori Hariyama,<sup>2</sup> Hayato Sakurai,<sup>3</sup> Mamiko Koshiba<sup>1,2,3\*</sup>

<sup>1</sup>Yamaguchi University, Ube, Yamaguchi, 755-8611, Japan

<sup>2</sup>Tohoku University, Sendai, Miyagi, 980-0845, Japan

<sup>3</sup>Saitama Medical University, Moroyama, Saitama, 350-0495, Japan

## Abstract

It has been reported that the speed of real-time comprehension and responding to the native emotional language depends on one's personality and age. The increase in the number of child psychiatric cases with developmental disorders and other communication disorders as major symptoms has become a social problem. While the internet society may be a possible cause of communication disorders, chat-generating AI, which is emerging with the development of machine learning technology, may contribute to communication training, depending on its design. To approach it, we developed a prototype tablet PC system that shows the input/output to/from open AI Chat GPT audio-visually through the speech with boy's voice, real-time viewed conversational texts with time, and facial expressions of a 3D boy avatar of the free temporary service "VRoid Studio". In an event to introduce regional children about novel digital educational materials, four children who voluntarily participated in the event attempted a few minutes of brief interaction with the avatar. To evaluate this prototype, we conducted a qualitative evaluation of emotions, which should be emphasized as a psychological factor in communication, using Google BERT (Bidirectional Encoder Representations from Transformers), a deep learning model that uses natural language processing to understand the meaning and context of the sentences in both each of the children and the avatar's conversation. The preliminary prototype of the chat-generating AI avatar responded many times faster than the child, and the response time tended to increase with the number of characters expressed by the children themselves and the avatar itself. Furthermore, the contents of neutral than positive emotion seemed to require time for children's response. These notes would be suggested to find the next specifications to develop an AI avatar-child interaction system that nurtures a deep sense of socioemotional humanity.

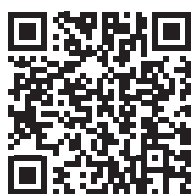
**Keywords:** Generative AI, Digital communication, Cloud internet, Mental health, Brain development

## Introduction

In recent years, the development of Internet and IT/AI technologies has continued to dramatically change the information and communication environment of society. These technologies had brought about a historic industrial revolution worldwide and violent and diverse changes in people's lifestyles and concepts. Some problems may have arisen in the social communication formation of children affected by systematic changes in the social environment.<sup>1,2</sup> In Japan, the increase in the number of child psychiatric cases such as developmental disorders in the general educational environment may imply that the society is becoming more difficult for children

to grow up in.<sup>3</sup> In cases of social dysfunction through language, such as autism spectrum disabilities, there is often a co-occurring emotional disorder, and a problem with understanding words that have emotional language meaning, with significantly more slow response.<sup>4</sup> The emotional disorders has been pointed out that there is a link with the transformation of experiences of emotional social interaction during the developmental period.<sup>5</sup> It has been reported that attention to and comprehension of the emotional vocabulary of one's native language, as well as response speed, depend on factors such as personality and age.<sup>6</sup> The speed of understanding and responding to language that includes emotional meaning may be influenced by similar factors across cultures.<sup>7</sup>

Quick Response Code:



**\*Corresponding author:** Mamiko Koshiba, Graduate School of Sciences and Technology for Innovation/Yamaguchi University, Ube, Yamaguchi, 755-8611, Japan

**Received:** 08 October, 2024

**Revised:** 02 January, 2025

**Accepted:** 11 February, 2025

**Published:** 14 February, 2025

**Citation:** Kenji Ito, Tsubasa Abe, Masanori Hariyama, Hayato Sakurai, Mamiko Koshiba. Emotional Features of Short Conversations between a Generative AI-Controlled Virtual Communication Trainer and Four Children: Research Article. *SOJ Eng Info*. 2025;3(1):1-11. DOI: [10.53902/SOJEI.2025.03.000504](https://doi.org/10.53902/SOJEI.2025.03.000504)

New styles of communication through digital among children, such as internet games, have characteristics unique to digital technology that younger children can easily absorb, and while they have new functionality to connect people, they may also cause children to lose the learning of complex social functions that have evolved through the interaction of humanistic information since ancient times.<sup>8</sup> The delay in conversation response may be a factor of change in a language communication society that is greatly different from the past due to the increased frequency of online digital communication. The explosion of OpenAI's ChatGPT, a conversational generative AI, has transformed the conventional lifestyle of human social information gathering. The interactive AIs like the Chat-type generative one are known about the following key technologies to achieve natural conversations. The first one is Transformer architecture based on a neural network utilizing a self-attention mechanism to efficiently learn the relationships between different parts of the input words.<sup>9</sup> The improvement brought abilities to understand context and generate appropriate responses. The second one is pre-training of large amounts of text data to learn the basic structure and patterns of language and fine-tuning for specific tasks and domains to generate responses with higher accuracy even in specific contexts.<sup>10</sup> And the third one is Natural Language Processing (NLP) Technology to analyze user input and generate appropriate responses. This includes morphological analysis, grammatical analysis, semantic analysis.<sup>11</sup> This new tool has been continued to learn from user feedback widely and is evolving day by day, combining both overall commonality and regional characteristics, and has become integrated into all areas, from general life to education.<sup>12-15</sup> As a result of achieving these key technologies, this generative AI generally realize natural and expected communication for a wide range of users caused by the three abilities; to understand the context of a conversation and generate appropriate responses, to respond to and answer the diverse topics raised by users, and to learn continuously based on new data and feedback to improve the quality of responses.<sup>16</sup>

As has been the case with newly developed revolutionary IT tools, it is impossible to deny the supposedly existence of negative effects, such as the degeneration of brain functions that had been activated by the time and effort spent on them, while bringing benefits to humankind.<sup>17</sup> To explore the possibility of supporting communication among the new generation by utilizing the positive aspects of the new IT tools that are more compatible with overcoming risks.<sup>13,14</sup> we prepared a tablet PC that can express voice and typical emotion-dependent facial expressions of an avatar friend who can understand the user's speaking contexts of neutral, positive and negative emotions give advice and exchange information to support their mental development in a virtual space. The PC can display facial expressions with chat voice. To further support communication and monitoring of the growth process, a function to display and record conversations in real time was prepared on the prototype machine. This study

is to report on the possibility of contributing to the formation of communication functions and presenting requirements for improvement by reviewing the contents of the prototype, which was shown at events introducing new IT tools to children in several locations in our local area, Yamaguchi Prefecture and played with for a few minutes by four participated children. To examine the mutual interaction of avatars and children, we focused on the comparison of delayed response times and attempted to evaluate the correlation between the number of letters and emotional categories under hypothetic prediction of the delay caused by the quantitative processing of received and uttered sentences, or the different processing times affected by some emotional quality. By capturing the conversational response delay time, a simple indicator detectable easily for practical use in actual educational and personal settings, and avoiding the difficult task of evaluating emotions, we explored whether there was any correlation with text that could be considered as one of the signals of emotion.

## Data and Methods

### The cloud chat AI 3D-Avatar system with voice, emotional face and real-time transliteration

The fuller chart of this system design, the outer UI/UX images, and the subsequent flowchart of the analysis are described in Figure 1. Each module, Speech Recognition (Browser Edge), conversation generation (Azure OpenAI Service (GPT 3.5)), Speech Recognition and web Speech API (Browser Edge) and 3D avatar model (three javascript@pixiv/three-VRM: free examples of VRM models available for commercial use (Fumituya Sakurada (Male model))) through a pixiv account were integrated for this communication avatar.

### VRM Model: Controlling the facial expressions of 3D models

VRM (Virtual Reality Model) is a 3D avatar file format for VR applications. Originating in Japan, this format is designed to emphasize interoperability between different platforms. VRM is characterized by a 3D model format that can be used in a variety of applications. This allows you to create an avatar once and use it on multiple platforms. VRM files are stored in "vrM" extension and is based on glTF 2.0 VRM data can be set with rights information for commercial use and redistribution. In order to control the VRM model in pixiv, we mainly use the three-VRM library. This library is combined with Three.js to display and control VRM models on the web.

### Basic commands

The following setting information is about the basic JavaScript commands for controlling a VRM model using three vrM.

**Avatar's character personalization settings:** In order to have an affinity for a user with the characters, it is possible to set the avatar's expression of the individuality by setting each profile item

of Table 1, such as name, age, gender, hobbies, special skills, and self-introduction.

**Changing facial expressions through conversation in emotional aspects:** Using sentiment analysis of Chat GPT, the technology to identify specific emotions from text as positive, negative, or neutral controlled the avatar's facial expression. The JavaScript commands enabled to change of the avatar's facial expressions according to the user's speaking in four emotional aspects, "Neutral", "Positive (joy, smile)", "Negative (sorrow) and Negative (angry) shown in Table 2. The three main categories of emotion were neutral, positive, and negative. In the avatar's facial expression settings, sadness and anger were set separately in the negative emotions although both were together regarded as the negative emotion in the response text analysis due to the small amount of data.

For instance, if the user says "I'm happy", the following command runs:

```
[[JavaScript]
if (userEmotion === 'happy') {
    vrm.blendShapeProxy.setValue('joy', 1.0);
}
}
```

Similarly, if the user says "sad", the following command runs:

```
[[JavaScript]
if (userEmotion === 'sad') {
    vrm.blendShapeProxy.setValue('sorrow', 1.0);
}
```

By coordinating the expressions of emotion in conversation with the facial expressions of the characters, a sense of realism with the conversational target could be achieved.

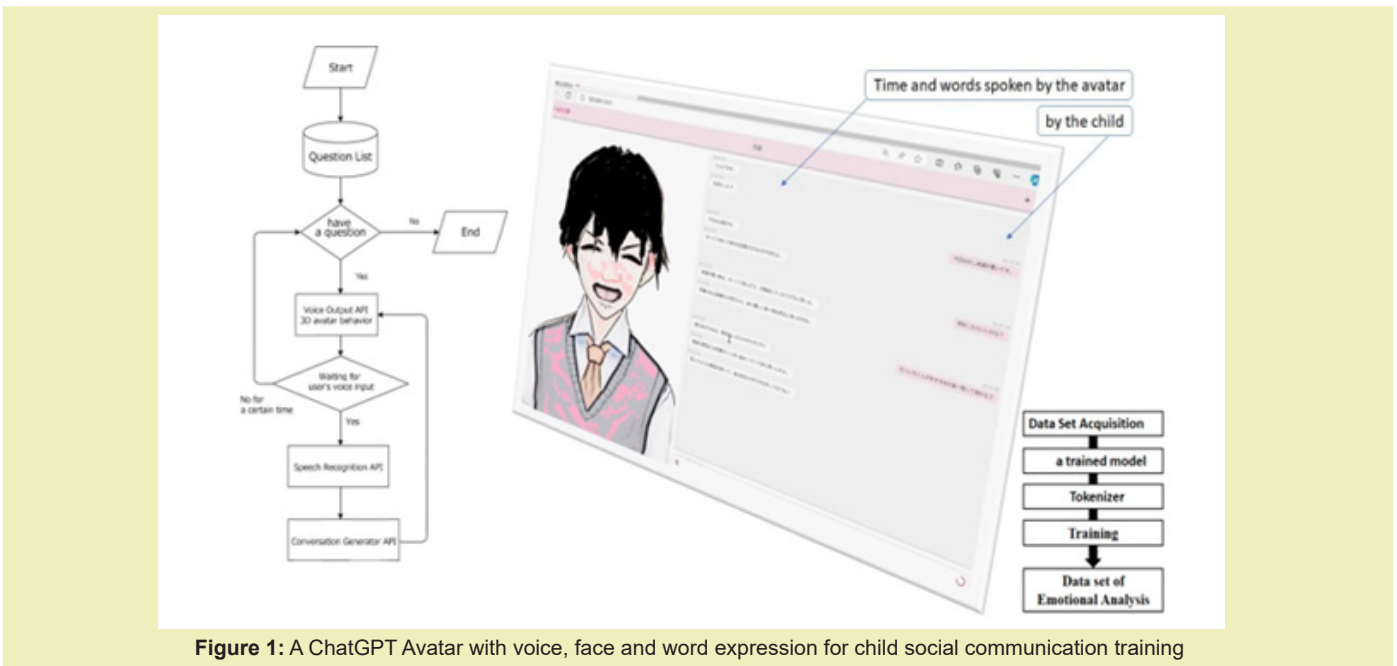
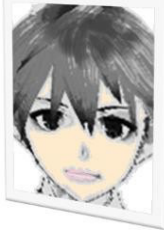
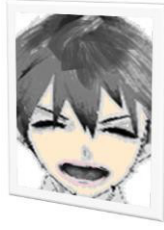
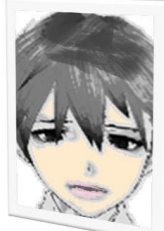


Figure 1: A ChatGPT Avatar with voice, face and word expression for child social communication training

Table 1: Commands for personalizing avatar's character

Profile Item	Command	Function
Name	SetName("Character Name")Sets	The name of the character
Age	SetAgeSets	The age of the character
Gender	SetGender("Gender")Sets	The gender of the character
Height	SetHeight	Sets the height of the character
Weight	SetWeightSets	The weight of the character
Hobbies	SetHobbies(["Hobbies1", "Hobbies2", ..])	Set the character's hobbies
Feats	SetSkills(["feat1", "feat2", ...])	Set the character's feats
Self-introduction	SetIntroduction("Self-introduction")	Set the character's self-introduction

**Table 2:** The commands of the Avatar's emotional expression

Emotion	Command	Avatar's Expression
Neutral	<code>vrm.BlendShapeProxy.SetValue('neutral', 1.0);</code>	
Positive joy smile	<code>vrm.blendShapeProxy.setValue('joy', 1.0);</code> <code>vrm.blendShapeProxy.setValue('smile', 1.0);</code>	
Negative confused sorrow	<code>vrm.blendShapeProxy.setValue('confused', 1.0);</code> <code>vrm.blendShapeProxy.setValue('sorrow', 1.0);</code>	
Angry	<code>vrm.blendShapeProxy.setValue('angry', 1.0);</code>	

#### Four children's playing trials at regional IT educational instruction events

This study protocol was approved by the Yamaguchi University Review Committee for Non-Medical Research Involving Human Participants (2023-064). All study members complied with the approved protocol. The researchers carefully followed the approved protocol at all times. Two times of public open events to introduce novel IT educational tools were held in 2023 fall in Yamaguchi prefecture. At the public event, the tablet-type avatar controlled by a chat-type generation AI was introduced on the spot without any prior publicity. Four primary school boys in grades 2 through 4 spontaneously attended with their parents at the area and played the conversational games with the prototype for a few minutes or less in the presence of each teaching assistant at a classroom desk.

#### Delayed response time with evaluation of processing load volume dependence

Conversation time and content information were automatically extracted by programming codes from a JSON file containing conversations between a generated AI avatar and four children, and converted to an Excel file. From this data set, the average and standard deviation of the delayed response times and numbers of conversation characters of the Avatar and the children were calculated. Pearson's product-moment correlation coefficient, intercept using linear regression, and probability were used for evaluation whether there was correlation between the amount of speech produced by avatars and children and the delay response time. If significance is found, the relationship between the physical load involved in cognitive processing and sentence formation may be hypothetically more suggestive than the influence of the emotional content of speech.

### Delayed response time comparison among emotional groups classified by BERT

With Bidirectional Encoder Representations from Transformers (BERT) or natural language processing, an interactive unsupervised language model to be released by Google in 2018, a small amount of supervised data can be used for various tasks such as sentence comprehension and sentiment analysis through fine tuning.<sup>18</sup> The sentences were tokenized and for the teacher data, "tyqiangz/multilingual-sentiment-datasets" from Hugging face, an open source platform for sharing and using Artificial Intelligence (AI) models and data, was used to represent the naturalness of the sentences with probability through neural language models. We used a BERT model trained on the Japanese version of Wikipedia, "cl-tohoku/bert-japanese", pretrained on texts in the Japanese language, released by the Natural Language Processing Research Group at Tohoku University (<https://www.nlp.ecei.tohoku.ac.jp/research/open-resources/>). The model architecture is the same as the original BERT base model; 12 layers, 768 dimensions of hidden states, and 12 attention heads. The trained data was 8000, test data was 200, the batch and epoch sizes were 8 and 1. The performance of the fine-tuned trained models was evaluated using normalized confusion matrices to represent. Using the obtained model, each set of conversation sentences between the four children and the avatar were divided into three emotional groups, positive, neutral, and negative. And the delayed response time of the avatar or the children were compared among three different emotional groups, positive, neutral and negative in own or the other speaking groups to seek any hypothetic emotion-dependency.

### Results

#### Avatar's delayed response time was approximately 10 times faster than children's one

With the aim of estimating the features related to any emotions in short conversations with four children and a chat-type AI-controlled virtual trainer, the speech response delay time of both the children and the virtual trainer was targeted as the assumed indicator. As the initial outline. Table 3 summarizes the average and standard deviation of the delayed response time of children and avatars, respectively, indicating that avatars' response time was much faster than that of children. The averages and standard deviations of delayed response mean values of these four children or the Avatar was  $9.5 \pm 1.0$  seconds or  $1.0 \pm 0.2$ , respectively. The result of the statistical comparison using the Student's t-test was a probability value of  $0.0003^{***}$ . This meant that the response delayed time of the avatar controlled by the chat-type generative AI was about 9.5 times shorter than the response delayed time of children, significantly. In addition, a comparison of the amount of Japanese word letters between four children and the Avatar was summarized the averages and standard deviation as  $8.4 \pm 5.2$  uttered by four children and  $32.6 \pm 21.7$  uttered by the Avatar.

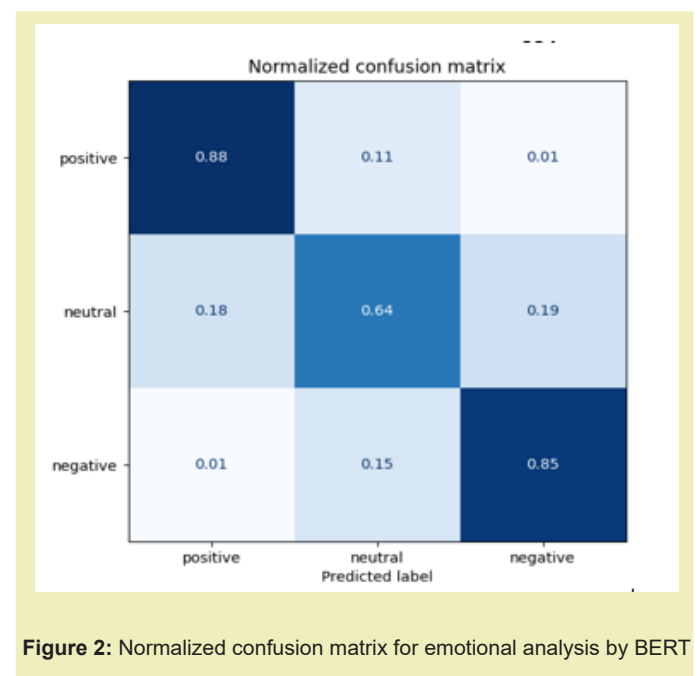
The probability calculated by student t-test between the children and the Avatar as  $2.32E-09^{***}$  determined significantly higher amount of letters that were uttered by the Avatar about 3.9 times more than the children.

**Table 3:** Delayed Response Time Average  $\pm$  Standard Deviation [s]

-	Children	Avatar
Child_1	$10.6 \pm 3.3$	$0.9 \pm 0.5$
Child_2	$8.5 \pm 1.1$	$0.9 \pm 0.4$
Child_3	$10.1 \pm 5.6$	$0.91 \pm 0.3$
Child_4	$8.8 \pm 2.1$	$1.3 \pm 0.4$

### Confusion matrix of model performance for text emotional analysis with BERT

The performance of the fine-tuned trained models with BERT by the Natural Language Processing Research Group at Tohoku University for emotional classification, based on the multilingual sentiments dataset, was evaluated each prediction accuracy of positive, neutral and negative emotion. The normalized confusion matrix of the performance of the fine-tuned trained model is shown in Figure 2. The discrimination between positive and negative is correct nearly 90% and false almost 0.01%, while the discrimination of neutral was generally correct nearly 64% and false from 0.11% to 0.19%. Compared to the higher accuracy rates for both negative and positive, so-called "polarities", the lower accuracy rate for neutral that means neither, was confirmed consistent with the imagined inherent difficulty of absolute emotion classification at certain boundaries in human sentiments of words, vocabulary related to emotions.



**Figure 2:** Normalized confusion matrix for emotional analysis by BERT



### Data set of words, emotional prediction and delayed response time between four children and the avatar

As shown in Table 4, an example of a pair data set along the time series of one child and the Avatar, we output each conversational sentences of full duration of several minutes between the four

children and the avatar, the time of delayed response, and the results of the emotional analysis of each set of multiple sentences by the BERT pre-trained model. These words were collected entirely from the logged conversation data. This example includes ten of conversing sessions between a child and the avatar, exchanged alternately.

**Table 4:** An example of a child and avatar data set

Avatar's Words	Child Response [s]		Child's Words	Avatar Response [s]	
	Emotion	Time [s]		Emotion	Time [s]
Hi, how are you doing?	positive	8.513	Hi, how are you doing?	positive	0.535
Hi, I'm doing well.	neutral	8.522	Usually. Yamaguchi	positive	0.603
A beautiful place. There are many famous tourist attractions and delicious food!	neutral	12.993	Fool?	negative	1.196
That's rude! It is not good to use such language! Respect the other person.	neutral	6.443	Fall.	neutral	0.873
Why do you use such language? Please refrain from being hurtful to others.	negative	10.686	What? I don't know. Good work.	neutral	0.678
Excuse me, are you in need of any help? I may be able to help.	neutral	19.262	Talk about something difficult.	neutral	1.503
For example, let's talk about quantum mechanics. It is the study of physical phenomena in the microscopic world, where it is difficult to make definitive predictions about the position and velocity of particles, etc. There is a probabilistic element involved.	neutral	10.604	How old?	neutral	0.622
I'm six years old, still in elementary school.	neutral	9.145	I agree.	neutral	0.901
Okay, so you agree. That is a fine opinion.	neutral	6.688	I kind agree.	neutral	1.074
That is a fine opinion. Why do you feel that way?	neutral	6.617	How old am I?	negative	1.176

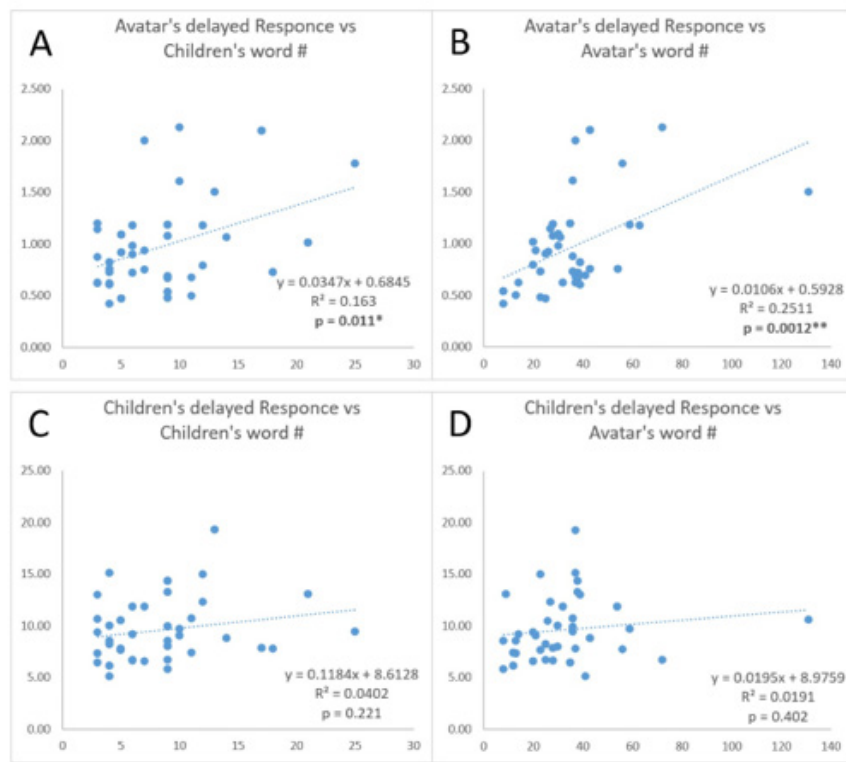
### Correlation evaluation of avatar and children's delayed response time with the number of letters

In order to examine whether the processing time for avatars and children to receive and respond to statements depends on the amount of text, we performed a correlation analysis between the amount of text received and sentences uttered with each of the delayed response time by Linear regression and Pearson product-moment correlation coefficients, intercepts, and probabilities Figure 3. Consequently, the correlation coefficient, R-squared value between the amount of letters received or uttered and the delayed response time by the children was much lower with lower probabilities as not significant in Figure 3C (Children's delayed response versus children's latter number:  $y = 0.1184x + 8.6128$ , R-squared = 0.0402,  $P = 0.221$ ) and Figure 3D (Children's delayed response versus children's latter number:  $y = 0.0195x + 8.9759$ , R-squared = 0.0191,  $P = 0.402$ ) than that of the avatars delayed response time which revealed higher correlation with the amounts of letters received, cognized, formed and output by machine processing with the generative AI as significant in Figure 3A (Avatar's delayed response versus children's latter number:  $y = 0.0347x + 0.6845$ , R-squared = 0.163,  $P = 0.011^*$ ) and Figure

3B (Avatar's delayed response versus Avatar's latter number:  $y = 0.0106x + 0.5928$ , R-squared = 0.2511,  $P = 0.0012^{**}$ ). This result suggests that the processing functions of the AI machine-controlled Avatar's word recognition and formation implied on a simpler, more uniformed mechanism, while the processing functions of the children's word recognition and formation rather showed diversified without converging.

### Exploring the emotional impact of conversations between avatars and children

Finally, to seek the possibility to indicate emotional dependence in delayed response time, an intergroup comparison was attempted among three emotional groups positive, neutral, and negative to see the effect of the words identified by the sentiment analysis by BERT on the delayed response time when uttered (Figure 4A: by Children, Figure 4B: by Avatar) or received (Figure 4C: by Children, Figure 4D: by Avatar). The results showed that the delayed response time of all four children was commonly shorter when they uttered Figure 4A or received Figure 4C positive words, and tended to be longer for neutral words (red arrows), but this tendency was not observed in the avatars either uttered Figure 4B or the received Figure 4D situation to any child of four.



**Figure 3:** Avatar and children's delayed response time with the numbers of each text Letters. Linear regression and Pearson product-moment correlation coefficients, intercepts, and probabilities less than 0.05 or 0.01 were visualized as significance with one or two asterisks

- A. Avatar's delayed response seconds from the child's last word versus the number of Children's last addressing words  
 B. Avatar's delayed response seconds from the child's last word versus the number of Avatar's addressing words  
 C. Children's delayed response seconds from the Avatar's last word versus the number of Children's addressing words  
 D. Children's delayed response seconds from the Avatar's last word versus the number of Avatar's addressing words

## Discussion

In this preliminary prototype evaluation study, we attempted to explore the requirements for development considerations related to conversations between chat-generating AI avatars and ordinary children, for seeking educational function but to remove the risk that active use of IT/AI may cause communication disorders,<sup>19,20</sup> which have become a social problem. On the contrary, this novel AI tool would be rather needed to open up new directions to actively enhance humanity.<sup>21</sup> Because of the complex neural basis that has been identified in biomedical neuroscience, this unknown generative AI/IT technology could have a serious impact on children who are forming the fundamental psychosomatic neural networks for their lifetime during their high-sensitivity period,<sup>5,22-28</sup> when exposed to this technological environment without sufficient care. The average of the delayed response time of avatars in the default setting was around ten times faster than that of children, as shown in Table 1. Furthermore, it was found that the amount of text produced by the Avatar was significantly higher than that of children, at around 3.9 times higher. Such large difference in the performance of the two speakers who are responsible for their dialogues each other was clearly expected to be far from the

ideal state of communication between people of peers. Since the synchronous rhythm of both communicators in a conversation should be considered as an important factor, the optimality of the reaction time setting would be required for future investigation.<sup>29,30</sup>

The correlation of avatar and children's delayed response time with the number of speech letters in Figure 3 was aimed to study for seeking any quantitative features and the results answered that the machine creation, avatar dominated by Chat GPT seemed significantly more dependent on the number of letters received and recognized (Figure 3A,  $P = 0.011^*$ ) and furthermore formed by the Avatar (Figure 3B,  $P = 0.0012^{**}$ ), supposedly related to the amount of processing, which was assumed as the processing load physically while the children were not significantly. This independence of children's delayed response time with the number of letters of the child-self or the Avatar might imply their diversified mental activities explainable the complexity in psychological biology,<sup>31</sup> not derived from stereotyped mechanisms of the machine processing, which might be reflected on the Avatar's significant correlation in Figure 3A,3B. This may paradoxically indicate the existence of emotions, as the delay time before speaking in response to the amount of words used by oneself and the other person shows

a more diverse distribution, but further verification is needed. Although this study limitation of small size was derived from only four children we could fortuitously meet at the public events as natural situation, each of the children's words was handled carefully and we attempted to explore not only the physical processing load in their brains, but also the psychological impact in more depth, accompanied by quantitative analysis using deep learning BERT<sup>32,33</sup> under deep consideration for their mental health.<sup>34</sup> The amount of text as the numbers of letters in the conversational looked slightly more related to the delayed response time of the avatar as the aforementioned about Figure 3, while the delayed response time of the children was more related to the changes of shortening in positive than neutral emotional states as visualized in Figure 4A (emotions in the own words) and Figure 4C (emotions for others' words) with expression of all the four children commonly, suggesting that in this child-avatar interaction system, the avatar behaves like a machine, while the children showed human sensitivity between positive and neutral in the response of this system. In chat GPT, the avatar's character was preset as a boy of the similar generation, however the content of his comments was more like that of an adult guide who knows most of information very well. In this analysis, the delayed response time of the avatar appeared rather dispersion meaning independent from the mechanism of difference between emotional positive and neutral in the speech words. To consider about any further improvement with essential factors for peer social learning in such as sympathy with synchronization or latent intrinsic attachment,<sup>27</sup> the avatar's response speech condition might be set emotion-dependent delay time like the shorter in emotion positive than neutral. Of the types of emotions, there was no particular commonality between the four children and the avatar in the case of negative emotions, whereas in the case of positive and neutral emotions, which may have implied an emotion-dependent mechanism only in the four children but not in the Avatar. One hypothesis is that negative emotions may have caused a more diverse dispersion of psychological states. It is another thought that the cause may have been that the different emotions of anger and confusion were set to one negative mode in the avatar's facial expression settings, explained in Table 2. It was suggested that more detailed settings for joy, anger, sadness, and happiness,<sup>35</sup> as well as delayed response analysis, will need to be carefully developed in the future.

For the educational support system of communication learning, an emotional aspect is crucial to comprehend how both users and AI function each other. In order to improve the accuracy of emotion detection, a Knowledge-Enriched Transformer (KET)<sup>36</sup> and the peripheral applications<sup>37,38</sup> were developed, which analyzes emotions by utilizing context and common sense knowledge to gain a deeper understanding of the context of the conversation and send appropriate emotion signals to control the next programming step. As for the emotion analysis based on the text used in this report, it is necessary to consider a comparison of the two directions

that should be explored in the future. One is the dictionary-based approach: words and phrases related to specific emotions are registered in a dictionary, and emotions are identified based on this data base. The others are machine learning models. Architectures for classifying emotions from text include naïve Bayes, Support Vector Machines (SVM), and deep learning (neural networks),<sup>39,40</sup> are necessary to be observed, analyzed, and reflected on the direction of rapidly evolving due to the dominant influence of human creations, machines, which have never existed before. In the development of digital technology to support children's communication learning, there is a speech recognition method that identifies emotions by analyzing the characteristics of the speaker's voice tone, pitch, speed for extracting acoustic features<sup>41</sup> such as Mel Frequency Cepstral Coefficients (MFCC),<sup>42,43</sup> and training models to classify emotions based on these features.<sup>44,45</sup>

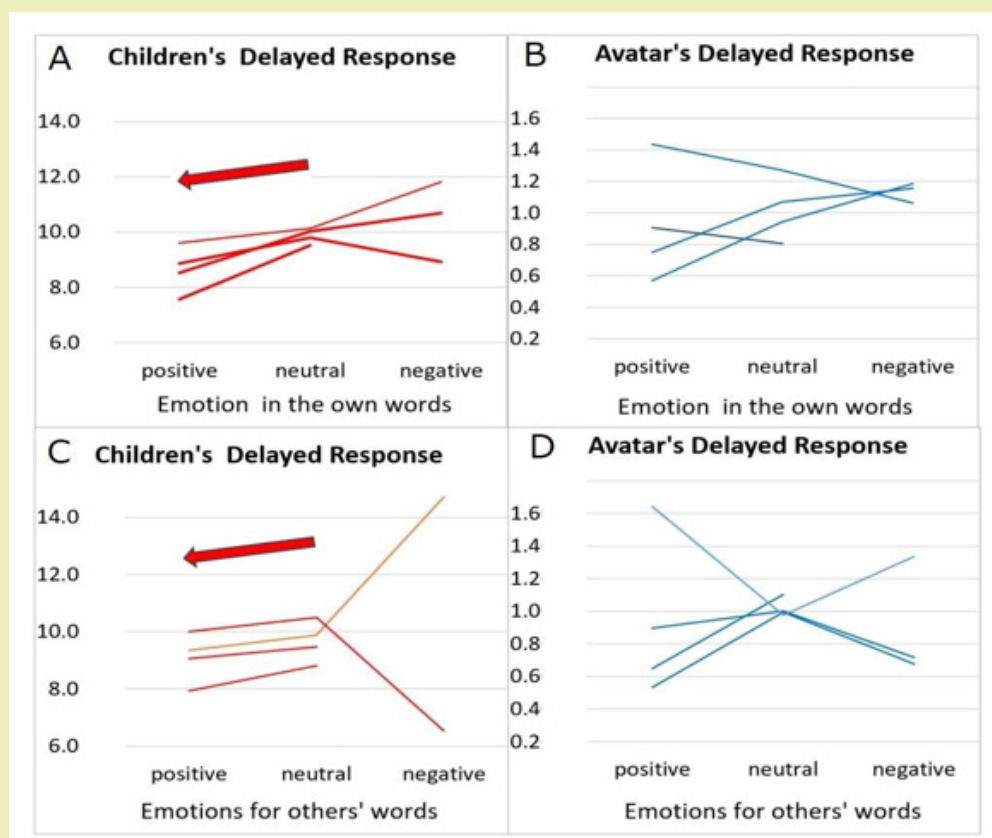
In addition, there are methods for classifying emotions using non-speech biological signals (heart rate, electrodermal activity, Electro Encephalo Graphy (EEG), Electrocardiogram (ECG), etc.) from the user. Furthermore, there is a method for identifying emotions by combining multiple sources such as text, speech, images, and biological signals, called multimodal emotion detection.<sup>36,46</sup> As another approach to develop interactive output media and platform, there may be the possibility that the network formation and adjustment methodology of human cognition and emotional psychology can be improved by combining multimodally with three dimension space control technology that provides a more immersive experience, such as metaverse<sup>47</sup> and larger scaled projection mapping,<sup>48</sup> which are being expanded and developed through IT currently. Children with learning disabilities and developmental disorders recently face many difficulties not only in their studies but also in their everyday interpersonal relationships. In order to alleviate these problems, it is necessary to provide training to develop social skills on a regular basis. To solve this issue, social skills training<sup>49</sup> has been developed and improved in various facilities such as schools, day service centers, and special needs schools.

Social skills training, which improves the ability to think and act appropriately and effectively in interpersonal relationships and group living through play and in a fun way, is expected to allow children to actively engage in the process themselves. However, the general social skill training requires specialist knowledge and to take time to create an environment where they can engage in this on a daily basis, which is important to be realized. The purpose of this study is to examine the improvement of communication skills in children with developmental disorders for future, especially those with communication disorders, through conversations with avatars using generative AI. As this research is only a preliminary trial, it is necessary to carefully verify its potential for application to developmental disorders in the future.<sup>50</sup> By adding digital signals to imitate the complexity of humans, IT robotics, which have artificial intelligence that continues to learn autonomously,



are coming closer and closer to living organisms. As a result, the sense of distance that was previously organized by the great differences between living beings and machines may change, for example, as the unique productive activities of each individual begin to be replaced by machines, and human activities may be forced to evolve into activities that surpass machines.<sup>51</sup> On the other hand, digital products equipped with generative AI that can replace human activity are becoming increasingly capable of interacting with humans, and there are concerns that this could lead to the development of addiction. There are also many reports about the risk of children becoming increasingly dependent on digital games, which could lead to serious social maladjustment, such as not attending school. By using the brain's neural network to exchange digital signals, which are easier to process than analog

signals, the brain may become less adept at analog processing, which requires complex signal processing, and this may accelerate the onset of addiction.<sup>52</sup> With this in mind, we would like to sound the alarm and demand that the development of a new educational system that combines real and virtual spaces is an urgent task. We alternatively hear many current views that machine brains have improved cognitive levels but no or less mind at the current study report. However, is there any denying that the physicochemical systems that neuroscience has revealed are far more complex than current machines, but that machines may arrive at that complexity someday.<sup>53,54</sup> We would like to search for the wisdom to coexist and evolve with and enhance machines in life on Earth and in the universe.



**Figure 4:** Avatar and children's delayed response average time compared among emotional (positive, neutral and negative) word groups classified by BERT

- A. The delay time from the last word spoken by the avatar to the child's next speech, positive, neutral or negative  
 B. The delay time from the last word spoken by the child to the avatar's next speech, positive, neutral or negative  
 C. The child's delay time from the last avatar's speaking words, positive, neutral and negative  
 D. The Avatar's delay time from the last child's speaking words, positive, neutral and negative

## Conclusions

We developed a prototype tablet PC system using a chat-type generative AI that shows the input/output to/from open AI Chat GPT audio-visually through the speech with a male's voice, real-time viewed conversational texts with time, and facial expressions of a three dimension male avatar of the free temporary use

service "VRoid Studio". In an event to introduce regional children about novel digital educational materials, four children who voluntarily participated in the event attempted a few minutes of brief interaction with the avatar. To evaluate this prototype, we conducted a psychological evaluation of emotions, positive, neutral and negative in the communication words between each of the children and the avatar, using Google BERT, a deep learning model

of natural language processing. Consequently, it was found that the preliminary prototype of the chat-generating AI avatar responded many times faster and more than the children. The Avatar's response time tended to increase with the number of characters expressed by the children's words the Avatar received and recognized and the avatar's speech himself. Furthermore, the contents of neutral than positive emotion seemed to require time for the four children's response to make their sentences and to receive and recognize the Avatar's words. The response delay time of both speakers, the children and the Avatar might be useful as an index to predict their emotional states in the speech communication. These notes would be suggested to find the next specifications to develop an AI avatar-child interaction system for education of social humanity with cognition and emotion.

### Acknowledgements

We acknowledge the children and parents, and the citizens and children supports of Hagi city, Yamaguchi Prefecture. We also acknowledge great supports of Ms. Iris Zhao AIH, AccScience Publishing.

### Funding

This study was supported by Yamaguchi University, Tohoku University and Saitama Medical University.

### Conflicts of Interest

Regarding the publication of this article, the authors declare that they have no conflicts of interest.

### References

- Tomoto F, Iwashiro K, Ota M, et al. Human Motion Tracking AI Revealed That a Hand-Made Swing in Nature Led to the Emergence of Children's Cooperative Society. *Stress Brain and Behavior*. 2023;3:e023001.
- Masam M, Somei S, Yoshino Hayakawa Y. A cherry tree's stress and recovery by inclusive intervention. *Stress Brain and Behavior*. 2019;1:2-6.
- Hua Z, Tao T, Akita R, et al. Four Temporary Waterslide Designs Adapted to Different Slope Conditions to Encourage Child Socialization in Playgrounds. *J Vis Exp*. 2022;190.
- Lui M, So WC, Tsang YK. Neural evidence for reduced automaticity in processing emotional prosody among men with high levels of autistic traits. *Physiol Behav*. 2018;196:47-58.
- Koshiba M, Senoo A, Mimura K, et al. A cross-species socio-emotional behaviour development revealed by a multivariate analysis. *Sci Rep*. 2013;3:2630.
- Lai VT, Pfeifer V, Ku LC. Emotional language processing: An individual differences approach. *Psychology of Learning and Motivation*. 2024;73-104.
- Stivers T, Enfield NJ, Brown P, et al. Universals and Cultural Variation in Turn-Taking in Conversation. *International Computer Science Institute*. 2009;106(26):10587-10592.
- Tsui YY yau, Cheng C. Internet Gaming Disorder, Risky Online Behaviour, and Mental Health in Hong Kong Adolescents: The Beneficial Role of Psychological Resilience. *Front Psychiatry*. 2021;12:722353.
- Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *Computer Science*. 2017.
- Radford A, Narasimhan K, Salimans T, et al. Improving Language Understanding by Generative Pre-Training.
- Jurafsky D, Martin JH. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models 3<sup>rd</sup> (edn.) Draft Summary of Contents.
- Dwivedi YK, Kshetri N, Hughes L, et al. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manage*. 2023;71:102642.
- Kamalov F, Santandreu Calonge D, Gurrib I. New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability*. 2023;15(16):12451.
- Lv Z. Generative artificial intelligence in the metaverse era. *Cognitive Robotics*. 2023;3:208-217.
- Chat B, Bard E, Motlagh NY, et al. The Impact of Artificial Intelligence on the Evolution of Digital Education: A Comparative Study of OpenAI Text Generation Tools Including ChatGPT The Impact of Artificial Intelligence on the Evolution of Digital Education: A Comparative Study of OpenAI Text Generation Tools Including ChatGPT, Bing Chat, Bard, and Ernie. 2023.
- Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. 2020.
- Baldassarre MT, Caivano D, Nieto BF, et al. The Social Impact of Generative AI: An Analysis on ChatGPT. 2024.
- Devlin J, Chang MW, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 2018.
- Hohenstein J, Kizilcec RF, DiFranzo D, et al. Artificial intelligence in communication impacts language and social relationships. *Sci Rep*. 2023;13(1).
- Privitera AJ, Ng SHS, Kong AP, et al. AI and Aphasia in the Digital Age: A Critical Review. *Brain Sci*. 2024;14(4):383.
- Jo H, Park DH. Effects of ChatGPT's AI capabilities and human-like traits on spreading information in work environments. *Sci Rep*. 2024;14(1):7806.
- Mimura K, Mochizuki D, Nakamura S, et al. A Sensitive Period of Peer-Social Learning. *J Clin Toxicol*. 2013;3(2).
- Karino G, Shukuya M, Nakamura S, et al. Common marmosets develop age-specific peer social experiences that may affect their adult body weight adaptation to climate. *Stress, Brain and Behavior*. 2015;3:1-8.
- Koshiba M, Watarai Senoo A, Karino G, et al. A Susceptible Period of Photic Day-Night Rhythm Loss in Common Marmoset Social Behavior Development. *Front Behav Neurosci*. 2021;14:539411.
- Shirakawa Y, Nakamura S, Koshiba M. Peer-Social Network Development Revealed by the Brain Multivariate Correlation Map with 10 Monoamines and 11 Behaviors. *J Clin Toxicol*. 2013;3:2.
- Karino G, Senoo A, Kunikata T, et al. Inexpensive Home Infrared Living/Environment Sensor with Regional Thermal Information for Infant Physical and Psychological Development. *Int J Environ Res Public Health*. 2020;17(18):6844.
- Koshiba M, Karino G, Senoo A, et al. Peer attachment formation by systemic redox regulation with social training after a sensitive period. *Sci Rep*. 2013;3.
- Senoo A, Okuya T, Sugiura Y, et al. Effects of constant daylight exposure during early development on marmoset psychosocial behavior. *Prog Neuropsychopharmacol Biol Psychiatry*. 2011;35(6):1493-1498.
- Fujii S, Wan CY. The Role of Rhythm in Speech and Language Rehabilitation: The SEP Hypothesis. *Front Hum Neurosci*. 2014;8:777.
- Kotz SA, Ravignani A, Fitch WT. The Evolution of Rhythm Processing. *Trends Cogn Sci*. 2018;22(10):896-910.

31. Li Z, Wu X, Li H, et al. Complex interplay of neurodevelopmental disorders (NDDs), fractures, and osteoporosis: a mendelian randomization study. *BMC Psychiatry*. 2024;24(1):232.
32. Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.
33. He J, Hu H. MF-BERT: Multimodal Fusion in Pre-Trained BERT for Sentiment Analysis. *IEEE Signal Process Lett*. 2022;29:454-458.
34. Nakayama M, Hatanaka C, Suzuki Y, et al. Generational differences in the image of text-based online counseling: Text analysis with deep learning technology. *Psychologia*. 2023;65(1):100-129.
35. Becker C, Conduit R, Chouinard PA, et al. Can deepfakes be used to study emotion perception? A comparison of dynamic face stimuli. *Behav Res Methods*. 2024;56(7):7674-7690.
36. Zhong P, Wang D, Miao C. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. 2019.
37. Pereira P, Moniz H, Carvalho JP. Deep Emotion Recognition in Textual Conversations: A Survey. 2022.
38. Hao Y, Cao P, Chen Y, et al. Complex Event Schema Induction with Knowledge-Enriched Diffusion Model. *Association for Computational Linguistics*. 2023;pp.4809-4825.
39. Machová K, Szabóová M, Paralič J, et al. Detection of emotion by text analysis using machine learning. *Front Psychol*. 2023;14.
40. Chutia T, Baruah N. A review on emotion detection by using deep learning techniques. *ArtifIntell Rev*. 2024;57(203).
41. Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features, and methods. *Speech Commun*. 2006;48(9):1162-1181.
42. Ancilin J, Milton A. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*. 2021;179:108046.
43. Sato N, Obuchi Y. Emotion Recognition Using Mel-Frequency Cepstral Coefficients. *Journal of Natural Language Processing*. 2007;14(4):83-96.
44. Nigar N. Speech Emotion Recognition Using Convolutional Neural Network and Its Use Case in Digital Healthcare. 2024.
45. Yun HI, Park JS. End-to-end emotional speech recognition using acoustic model adaptation based on knowledge distillation. *Multimed Tools Appl*. 2023;82(15):22759-22776.
46. Udahemuka G, Djouani K, Kurien AM. Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. *Applied Sciences*. 2024;14(17):8071.
47. Maghaydah S, Al Emran M, Maheshwari P, et al. Factors affecting metaverse adoption in education: A systematic review, adoption framework, and future research agenda. *Heliyon*. 2024;10(7):e28602.
48. Tao T, Sato R, Matsuda Y, et al. Elderly Body Movement Alteration at 2nd Experience of Digital Art Installation with Cognitive and Motivation Scores. *J-Multidisciplinary Scientific Journal*. 2020;3(2):138-150.
49. Mirzaei SS, Pakdaman S, Alizadeh E, et al. A systematic review of program circumstances in training social skills to adolescents with high-functioning autism. *Int J Dev Disabil*. 2020;68(3):237-246.
50. Conde M, Rodríguez Sedano FJ. Is learning analytics applicable and applied to education of students with intellectual/developmental disabilities? A systematic literature review. *Comput Human Behav*. 2024;155:108184.
51. Chhetri B, Goyal LM, Mittal M. How machine learning is used to study addiction in digital healthcare: A systematic review. *International Journal of Information Management Data Insights*. 2023;3(2):100175.
52. Mohammad S, Jan RA, Alsaedi SL. Symptoms, Mechanisms, and Treatments of Video Game Addiction. *Cureus*. 2023;15(3):e36957.
53. Shuaib A, Arian H, Shuaib A. The Increasing Role of Artificial Intelligence in Health Care: Will Robots Replace Doctors in the Future?. *Int J Gen Med*. 2020;13:891-896.
54. Harzheim JA. What Does It Mean to Be Human Today?. *Cambridge Quarterly of Healthcare Ethics*. 2024.