Overall Design and Physical Validation of Voice Interaction based on the ChatGPT Humanoid Robot Brain

Zhenzhong Li
CATARC
Automotive Test Center (Ningbo) Co., LTD
Ningbo, China
18858236336@163.com

Bing Li
CATARC
Automotive Test Center (Ningbo) Co., LTD
Ningbo, China
libing@catarc.ac.cn

Jie Tang
Trial and Testing Division of the Research and Development Center
East Shineray Group Co., LTD
Chongqing, China
466326287@qq.com

Liang Yan
Research and Development Laboratory
Amileyuan Intelligent Technology (Beijing) Co., LTD
Beijing, China
amileyuan@163.com

Jingwen Li
Yangtze Delta Region Academy of Beijing Institute of Technology,
Jiaxing, China
15753885032@163.com

Zhiyuan Yu
High-Speed Silicon-Based System-on-Chip Research Laboratory
BIT Chongqing Institute of Microelectronics and Microsystems
Chongqing, China
18243773695@163.com

Yibo He
Test Center
DONGFENG LIUZHOU MOTOR CO., LTD
Liuzhou, China
heyb@dflzm.com

***Abstract*—Aimed to the challenge of robotic voice interaction, this study leverages ChatGPT(Chat Generative Pre-trained Transformer) technology to develop a humanoid robot solution with a central processing system. Employing a top-down design approach, the solution encompasses the design of voice, video, and motion streams between users and the robot, enabling voice communication and expression output. By integrating hardware devices and a central control system, the entire humanoid robot system achieves a twofold purpose. On one hand, it combines data from conversational context, user tone and emotion, and user facial expressions to appropriately exhibit expressions. On the other hand, it formulates reasonable voice responses in conjunction with extracted statement content and emotional cues. Lastly, two physical prototypes of the humanoid robot are constructed. Experimental trials are conducted to assess the voice conversation and expression output capabilities of the humanoid robot, thereby confirming the rationality and effectiveness of the proposed solution. (*Abstract*)**

Keywords—Humanoid Robot; ChatGPT; Voice Interaction; Expression Output (*key words****)***

I. INTRODUCTION

With the rapid advancement of artificial intelligence technology, intelligent robots have gradually become a focal point of research in the realm of technology [1-3]. Humanoid robots, as a significant branch of intelligent robotics, hold the potential to be applied in various fields such as manufacturing, healthcare, education, and even as domestic service robots, facilitating intelligent interaction and service with humans [4].

The head of a humanoid robot plays a pivotal role in achieving voice interaction and facial expressions. The design of head-based voice interaction stands as a crucial element in realizing intelligent dialogues and human-robot interactions. Over the past few decades, head-based voice interaction design has been a prevailing area of research [5]. Researchers, through continuous exploration and innovation, have developed an array of diverse techniques for head-based voice interaction. With the evolution of natural language processing technology, the design of head-based voice interaction driven by the ChatGPT brain [6] has progressively garnered attention in the realm of humanoid robot research.

A. Current State of Traditional Head-Based Voice Interaction Design

Conventional head-based voice interaction design for humanoid robots predominantly relied on rule-based speech recognition and template-based speech synthesis techniques [7-8]. However, these approaches possess limitations and struggle to adapt to variations in users' speech characteristics and expressions, necessitating improvements in dialogue quality. With the ongoing development of natural language processing technology, machine learning and deep learning-based voice interaction techniques have found increasingly widespread application.

B. Current State of Deep Learning-Based Head-Based Voice Interaction Design

Deep learning-based head-based voice interaction design employs neural network models for speech recognition and synthesis. This approach can be trained and optimized with substantial speech data, enhancing dialogue accuracy and naturalness. Common deep learning techniques [9-10] encompass Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer networks. Among these, Transformer networks have gained extensive usage in the field of natural language processing, with voice interaction models based on Transformers emerging as a research focus.

Science and technology innovation:2025 key proiec(Ningbo key science and technology research project) (GrantNo.2022Z074) $2000000.00

C. Current State of Head-Based Voice Interaction Design with the ChatGPT Brain

ChatGPT represents one of the widely applied natural language processing technologies [11-12], boasting a powerful generative language model capable of producing dialogue expressions akin to human conversations. Utilizing the ChatGPT brain for head-based voice interaction design effectively enhances the naturalness and fluency of conversations, enabling humanoid robots to better comprehend and respond to users' linguistic expressions. Furthermore, the ChatGPT brain continuously learns and optimizes, thereby improving dialogue accuracy and adaptability.

II. Overall Architectural Design of the Humanoid Robot

A. Overall Design Process of the Head based on the ChatGPT Brain

The design of the head module is primarily aimed at the three main participants within the system: the user, ChatGPT, and the robot. The entire process involves three key stages: dialogue interaction, expression recognition, and speech synthesis, as illustrated in Figure 1.

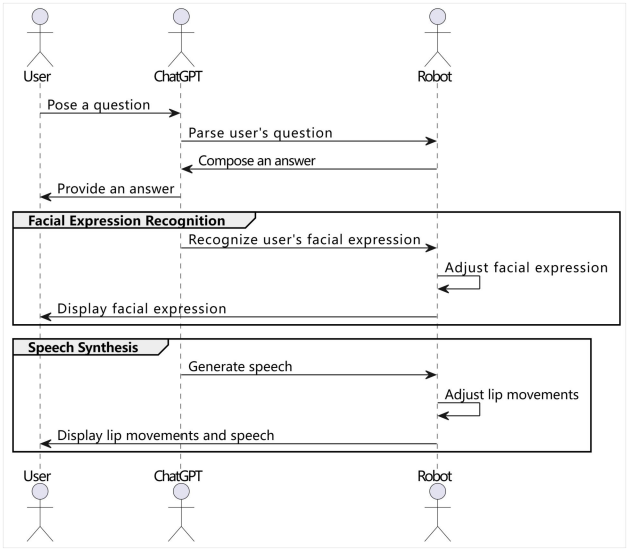

Fig. 1. Process Flow of Head Design based on the ChatGPT Brain

This activity diagram illustrates the various steps and roles within the "Humanoid Robot Head Design with ChatGPT Brain-Based Voice Interaction" system.

Initially, the user engages in voice interaction with ChatGPT by posing questions. ChatGPT interprets the user's queries and collaborates with the robot to formulate responses. In terms of expression recognition, ChatGPT analyzes the user's facial expressions and communicates this information to the robot. The robot adjusts its facial expressions based on this information and presents facial expressions to the user. For speech synthesis, ChatGPT generates a voice response, and the robot synchronizes its lip movements with the generated speech, showcasing lip movements and the spoken response to the user.

By integrating various technologies, the entire system achieves the functionality of head-based voice interaction for the humanoid robot, delivering a more humane and natural interactive experience.

B. Design of User-Robot Voice Stream

The voice interaction between the user and the robot takes place primarily between several nodes, including the microphone, ASURE, ChatGPT, and loudspeaker. The process is illustrated in Figure 2.

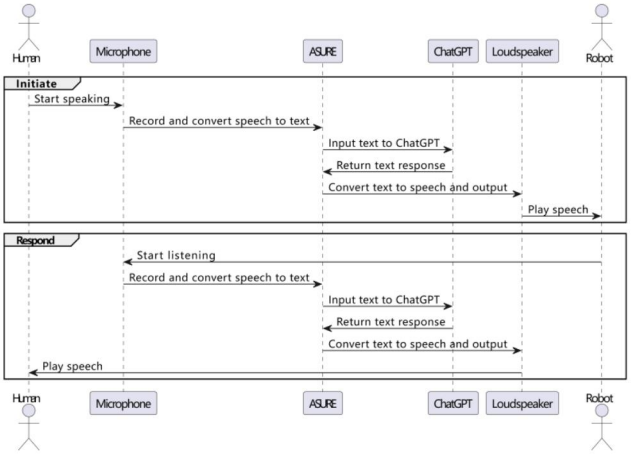

Fig. 2. Voice Communication Process based on the ChatGPT Brain

The process depicted in Figure 2 primarily accomplishes the voice communication between human users and the robot. It is described from both the user's perspective and the robot's perspective.

From the user's perspective: The human user begins by inputting voice through the microphone. The voice is then converted to text by ASURE, and this text is sent to ChatGPT. ChatGPT processes the text through natural language processing and produces a textual answer. This answer text is transformed into speech by ASURE and played back on either the human user's or the robot's loudspeaker.

From the robot's perspective: The system continuously monitors whether the microphone receives voice information from the human user. Once valid information is identified, and after processing through ASURE and ChatGPT to generate output data, it is played through the loudspeaker.

C. Design of User-Robot Video Stream

The video communication between human users and the robot primarily involves external cameras, OPENCV image processing module, Azure Face API, and the CHATGPT module. The process is illustrated in Figure 3.

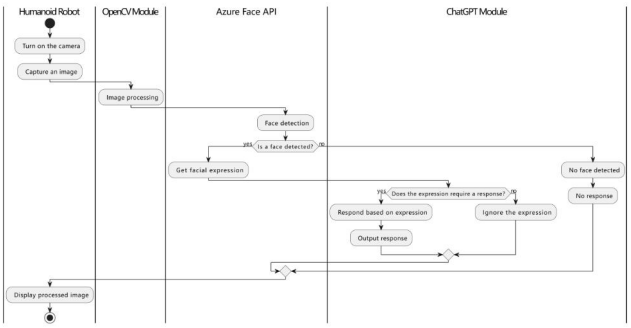

Fig. 3. Humanoid Robot Facial Expression Recognition Process

Figure 3 illustrates the workflow of a facial expression recognition system based on a humanoid robot. This system captures users' facial expressions through an built-in camera

in the robot's eyes and employs computer vision and machine learning techniques for analysis, identifying users' emotions and expressions. The overall architecture of this system comprises three main parts: Facial Expression Recognition Module, Interaction Module, and Response Module.

Within the Facial Expression Recognition Module, the built-in camera of the humanoid robot continuously captures users' facial expressions. Computer vision techniques process and analyze the captured images, sending the results to the Interaction Module.

In the Interaction Module, the analysis results are fed into a machine learning model to further identify users' emotions and expressions. This module then sends the analysis results to the Response Module to generate appropriate responses based on users' emotions and expressions.

In the Response Module, the humanoid robot responds accordingly to users' emotions and expressions, facilitating interaction. For instance, when the robot detects a happy expression, it might perform cheerful actions or play joyful music to enhance the user's positive experience. Conversely, if the robot identifies a sad expression, it could play comforting music or execute suitable comforting gestures to alleviate the user's negative mood.

D. Design of User-Robot Motion Stream

The execution of interaction actions between users and the robot is primarily manifested in the robot's facial expressions. The implementation process is depicted in Figure 4.

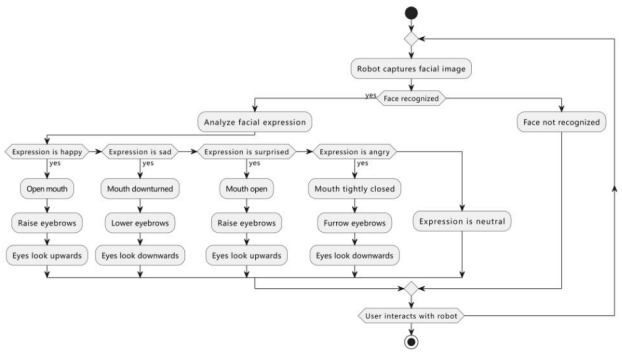

Fig. 4. User-Robot Interaction Motion Execution Process

Figure 4 demonstrates how the robot recognizes human facial expressions through visual cues and responds accordingly. During the execution process, the robot constantly reads information from its built-in camera. If no face is detected, it executes predefined internal expressions. However, when a human facial expression is captured, computer vision techniques are employed to analyze the expression, yielding relevant facial expression parameters.

Based on the recognized expression parameters, the robot controls the movements of its facial organs, including eyelid position, eyeball position, eyebrow position, and mouth opening extent. These actions of the robot are executed by the hardware system after the expression parameters have been parsed.

III. HARDWARE DESIGN OF THE HUMANOID ROBOT

In this section, we will delve into the hardware structure and control system of the humanoid robot's head. The aim is to enable the robot to perceive human expressions and generate corresponding head movements. The hardware of the humanoid robot can be divided into three main parts based on functionality: input devices, central control system, and output devices. The specific composition is illustrated in Figure 5.

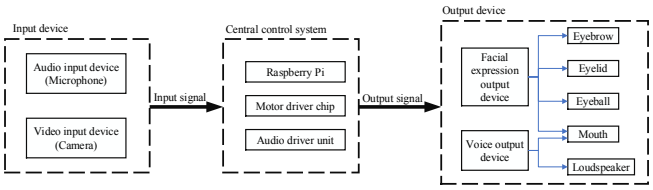

Fig 5. Hardware Composition of the Humanoid Robot

A. Robot's Input and Output Devices

The robot's input devices are primarily used to receive information from human users, including audio input devices and image input devices. The audio input device consists of microphones placed within the robot's ear canals, functioning to real-time capture users' voice input. A microphone is situated in each ear to mimic the reception of external sounds as closely as possible. The image input device consists of cameras placed within the robot's pupils, facilitating real-time capture of users' facial expressions. Each eye is equipped with a camera to simulate the way humans observe external stimuli.

The robot's output devices are employed for expression display and dialogue output, encompassing audio output devices and expression output devices. Audio output devices include the mouth and loudspeaker. The mouth can be adjusted to correspondingly open and close according to the audio output. The loudspeaker is positioned within the head shell, playing back processed audio information. Expression output devices consist of four components: eyebrows, eyelids, eyeballs, and the mouth. Each component can move independently. Eyebrows and eyelids have two degrees of freedom each, while eyeballs have a total of four degrees of freedom. The mouth possesses one degree of freedom, as illustrated in Figure 6.

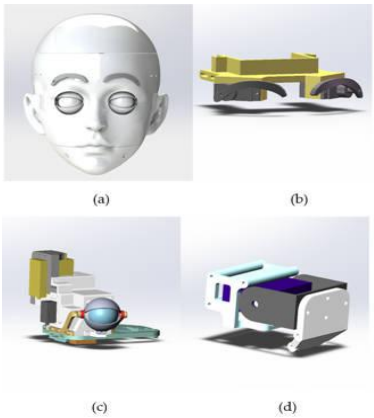

Fig. 6. Robot Head Design and Articulated Organs: (a) Robot Head Design; (b) Eyebrow Mechanical Structure; (c) Eyeball and Eyelid Mechanical Structure; (d) Eyeball and Eyelid Mechanical Structure.

Figure 6 illustrates the overall appearance of the robot's head and schematic representations of articulated parts such as eyebrows, eyeballs, eyelids, and the mouth. These degrees of freedom can be controlled through motor drives to achieve a variety of head movements. The central control system can

manage the control of all motors and configure motor movements into different sets of facial expressions.

The robot's central control system needs to be capable of realizing various sets of facial expression movements, which can be represented as a multi-degree-of-freedom motion control problem. Within the central control system, each motor has its distinct control parameters that can be calculated to achieve various sets of expression movements. The angles of each motor correspond to the movements of different parts of the robot's head, and the control parameters within the central control system can be represented as vectors:

$$P = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} \quad (1)$$

Where n is the number of control parameters, and p_i represents the value of the i-th control parameter. For each type of expression, a corresponding control parameter vector P can be obtained. As a result, the action group matrix for all expressions can be obtained:

$$M = (P_1 P_2 \cdots\cdots P_m) \quad (2)$$

Where m is the number of expressions. By utilizing the control matrix through the central control system, the manipulation of facial expressions and actions in the robot can be achieved. Specifically, the central control system receives instructions from the Raspberry Pi and governs the actions of multiple internal motors within the robot's head, controlling various movements.

The two degrees of freedom in the eyebrows can adjust the eyebrow shapes as needed. The two degrees of freedom in the eyelids control the extent to which the eyes open or close. The four degrees of freedom in the eyeball section coordinate eye movement in vertical and horizontal directions. The degree of freedom in the mouth controls its opening and closing.

The central control system can combine the actions of each motor into different sets of facial expression movements, thus achieving precise control over the robot's facial expressions. By coordinating the values of eyelid and eyebrow control parameters, the robot can portray a series of expressions, such as furrowing brows, raising brows, glaring, and squinting. Through varying control parameter values for the eyes, expressions like mischievousness, contemplation, and disdain can be conveyed. Similarly, the mouth's control parameter values, adapting with changes in language content, enable actions of speaking, in addition to expressions of surprise, indifference, and plainness.

B. Central Control System

The central control system, as the core component enabling various functionalities of the robot, is responsible for analyzing and processing the collected sensory data, computing the required movements for each organ of the head, and controlling motor rotations to achieve facial expressions and language output. This control system employs a Raspberry Pi-based control solution, facilitating efficient data transmission and computational capabilities. Specifically, the central control system comprises hardware components such as the Raspberry Pi, driver chips, and audio driver units. By utilizing AI technologies like ASR (Automatic Speech Recognition) and Chatbot, it realizes the robot's intelligent interaction capabilities.

During operation, the control system continuously receives audio and video signals from the input devices. Simultaneously, it analyzes and filters the information, activating data fusion and processing upon detecting valid signals, such as conversational audio or human images. On one hand, it analyzes the content of spoken audio, identifying the speaker's tone and emotional content. Simultaneously, it retrieves response content through ChatGPT and commands audio output devices. On the other hand, it interprets human expressions from the provided images. Ultimately, by integrating previous emotional information from the conversation (context), current dialogue information, and facial expression data, the system computes the expression output information, orchestrating actions through the facial expression output devices. The operational flow is illustrated in Figure 7.

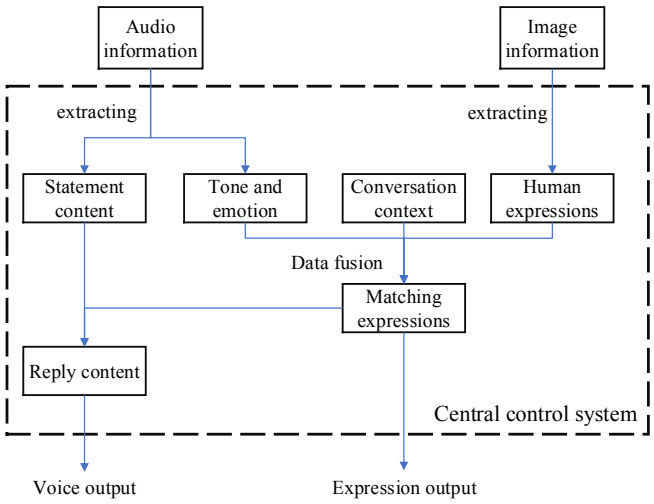

Fig. 7. Workflow of the Central Control System.

IV. Experiments of the Humanoid Robot System

The assembly of the robot's headshell and hardware components is depicted in Figure 8.

Fig. 8. Process of Robot Headshell and Silicone Fabrication.

To achieve more natural facial expressions, the humanoid robot employs platinum silicone skin, as depicted in Figure 9.

Fig. 9. Robot Silicone Production Process.

In order to validate the system's effectiveness, two physical humanoid robot heads were fabricated for experimentation, specifically testing the robot's speech dialogue functionality and facial expression capabilities.

A. Verification of Robot's Speech Dialogue Functionality

The first humanoid robot head was integrated into the complete robot body system, simulating real human phone conversations, as illustrated in Figure 10.

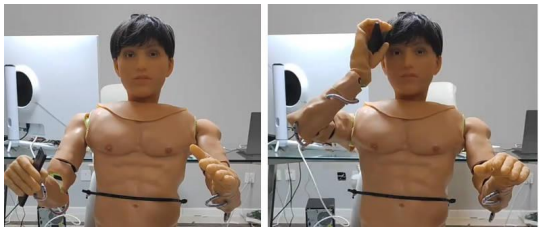

Fig. 10. Actual Scenario of Robot Making a Phone Call.

Upon receiving the call, the robot's hand picks up the phone and activates the phone's speaker functionality for the conversation experiment. The robot uses the microphone in its ear to receive the incoming voice content and responds appropriately. The dialogue process is illustrated in Figure 11.

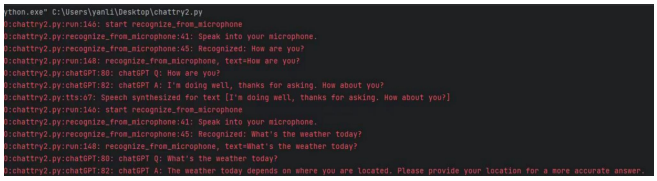

Fig. 11. Process of Robot Making a Phone Call.

Based on the displayed content, it is evident that the robot's speech dialogue functionality meets the design requirements.

B. Verification of Robot's Facial Expression Output Functionality

The second humanoid robot head was primarily employed to verify the facial expression output functionality. It relies on cameras placed in the eye sockets to capture human facial expressions and respond accordingly. Figure 12 illustrates the process of the camera recognizing the user's facial expressions.

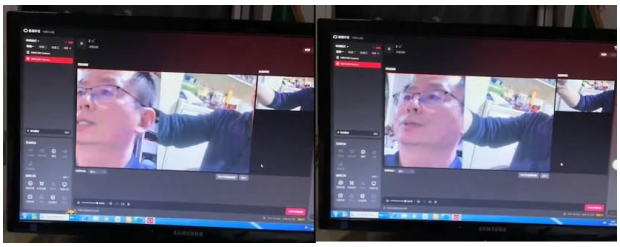

Fig. 12. Robot's Eye Cameras Capturing User's Image.

During the experimentation process, various expressions were simulated by the robot while engaging in conversation. The robot's performance is demonstrated in Figure 13.

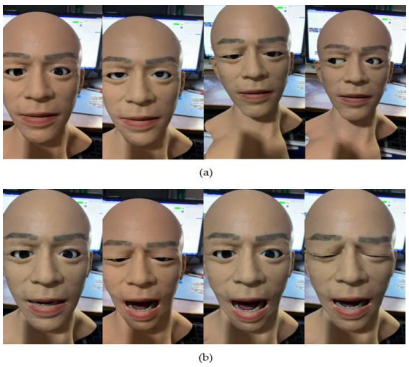

Fig. 13. Experimental Results of Robot's Facial Expression: (a) Different Expressions Displayed by the Robot; (b) Robot's Speaking State under Different Expressions.

In these figures, Figure 13(a) shows the robot's response after recognizing human expressions, while Figure 13(b) depicts the scenario where the robot recognizes human expressions and engages in conversation with them.

From this, it can be inferred that the humanoid robot head system developed based on the software and hardware design in this paper is capable of effectively responding to human facial expressions and language, accurately displaying facial reactions, and providing verbal answers. This validates the rationality of the system configuration and the correctness of the theoretical analysis.

V. CONCLUSIONS

This paper presents an overall design and physical validation method for voice interaction based on the ChatGPT humanoid robot brain. Firstly, the paper provides a detailed description of the overall architectural design of the humanoid robot, including the design process of the head based on the ChatGPT brain, the flow design of voice interaction between users and the robot, the design of video interaction, and the design of motion flow. It also demonstrates the interaction and connection of various components of the system. Subsequently, the paper introduces the hardware design of the humanoid robot, including input devices, central control systems, and output devices. Through the integration of software and hardware, the physical humanoid robot head is realized, enabling more intelligent interaction with users. Through two sets of physical experiments, the effectiveness of the proposed approach is validated. The experimental results indicate that the humanoid robot based on the ChatGPT brain possesses perceptual, processing, and expressive abilities during voice interactions, and can accurately respond to users, providing a natural and smooth interaction experience. This research

provides valuable references and insights for the development and application of humanoid robots, offering innovative ideas and practical methods for the field of intelligent interaction.

ACKNOWLEDGEMENTS

Author Contributions: Conceptualization, B.L. and Y.G.; methodology, B.L., Y.G. and Z.L.; software, X.W. and L.Y.; validation, Z.L. and L.Y.; formal analysis, Y.G. and X.W.; investigation, L.Y.; resources, B.L. and Y.G.; data curation, Z.L.; writing—original draft preparation, L.Y.; writing—review and editing, X.W.; visualization, B.L.; supervision, Y.G.; project administration, Z.L.; funding acquisition, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and technology innovation 2025 key project (Ningbo key science and technology research project), grant number 2022Z017.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusion of this article will be made available by the authors without undue reservation.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language Models are Unsupervised Multitask Learners. NeurIPS 2019, 33, 9771-9780.

[2] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. NeurIPS 2017, 30, 5998-6008.

[3] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI 2019.

[4] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ACL 2019, 1, 4171-4186.

[5] Vaswani, A.; Boving, A. A Comprehensive Survey of Transformers. arXiv 2021.

[6] ChatGPT - OpenAI's Language Model. https://openai.com/research/chatgpt (accessed on 10 May 2022).

[7] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A Robustly Optimized BERT Pretraining Approach. arXiv 2019.

[8] Wu, F.; Fan, A.; Baevski, A.; Auli, M.; Edunov, S. Pay Less Attention with Lightweight and Dynamic Convolutions. arXiv 2019.

[9] Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv 2019.

[10] Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language Models are Unsupervised Multitask Learners. NeurIPS 2019, 33, 9771-9780.

[11] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI 2019.

[12] Wu, F.; Fan, A.; Baevski, A.; Auli, M.; Edunov, S. Pay Less Attention with Lightweight and Dynamic Convolutions. arXiv 2019.