Using a LLM-Based Conversational Agent in the Social Robot Mini

Iván Esteban-Lozano[1,2], Álvaro Castro-González[1](✉), and Paloma Martínez[2]

[1] Robotics Lab, Systems Engineering and Automation Department, Universidad Carlos III de Madrid, Avenue de la Universidad 30, Leganés 28911, Spain
100383779@alumnos.uc3m.es, acgonzal@ing.uc3m.es

[2] Computer Science Department, Universidad Carlos III de Madrid, Avenue de la Universidad 30, Leganés 28911, Spain
pmf@inf.uc3m.es

Abstract. Natural Language Processing has witnessed significant growth in recent years. In particular, conversational agents have improved significantly thanks to the proliferation of the Large Language Models (LLM). Conversational agents have already been integrated with smartphones, smart speakers, or social robots (SRs). Unlike the mentioned electronic devices, a social robot allows more active and closer user engagement due to the presence of a physical object with a lifelike appearance that is able to express emotions. Therefore, SRs represent an appealing platform for deploying a conversational agent. In the field of social robotics, the ability of robots to interact with humans has traditionally been limited by their verbal skills. Until recently, robots could only understand a limited set of human utterances using specific rules, and the utterances of the robots were pre-defined sentences crafted offline. These restrictions, on many occasions, lead to repetitive interactions, which could cause users to lose interest during prolonged engagement with the robot. In this paper, we propose to integrate into our social robot Mini a conversational agent based on LLM. We present a new robot skill that can maintain a natural and seamless conversation with the user on any desired topic. The obtained results show a high usability of the skill and a high-quality interaction.

Keywords: Social Robots · Large-Language Models · chatbot · Conversational Assistants · Conversational Agents

1 Introduction

Conversational agents have already been integrated into smartphones, smart speakers, or social robots (SRs). Unlike the mentioned electronic devices, a social robot allows more active and closer user engagement due to the presence of a physical *object* that can be seen and touched. Moreover, the lifelike appearance of an SR eases users to establish a more realistic and enduring connection. Also, its ability to interpret and express emotions through gestures and movement

H. Degen and S. Ntoa (Eds.): HCII 2024, LNAI 14736, pp. 15–26, 2024.
https://doi.org/10.1007/978-3-031-60615-1_2

provides a more complete and natural experience. Therefore, SRs represent an appealing platform for deploying a conversational agent.

A conversational agent, or chatbot, is defined as a software system created for natural language interaction with users [18]. In particular, conversational agents have improved significantly thanks to the recent proliferation of the Large Language Models (LLM from now on). These models can analyze and generate human-like text, making them versatile tools for a wide range of applications (machine translation, assistants and interactive conversational bots, sentiment analysis, and text classification, among others). LLMs can process and generate text that appears coherent and right to people, but it does not mean that LLMs have consciousness or understanding. The performance of most LLMs is due to transformer architecture [17] that has revolutionized language technologies together with the availability of large training datasets to build and adapt these models to different applications.

In the field of social robotics, the ability of robots to interact with humans has traditionally been limited by their verbal skills. Until recently, robots could only *understand* a limited set of human utterances using specific rules or grammar [4]. Additionally, the utterances of the robots were pre-defined sentences that were crafted offline. These restrictions, on many occasions, lead to repetitive interactions which could cause users to lose interest during prolonged engagement with the robot [15].

The proposal is to integrate into our social robot Mini a conversational agent based on a LLM. We aim to create a new robot skill that is capable of maintaining a natural and seamless conversation with the user on any desired topic. Thanks to it, we expect to achieve more natural and friendly interactions that help to engage the user in the robot interaction [7].

Mini is a social robot designed to support and accompany seniors in their daily lives. We expect that using the robot as a conversational agent will provide companionship to elders helping them combat loneliness. In addition, we believe that social robots able to interact in a human-like manner allow their users to train their memory and mental agility and even have a fun and joyful time with the robot every day.

The rest of the paper is structured as follows. In Sect. 2, we describe and present the most relevant concepts of Language Models. After, the proposed conversational agent is presented in Sect. 3. The integration of the conversational agent in the robot Mini comes next (Sect. 4). In the last part of the paper, Sect. 5 presents the results that have led us to the conclusions of Sect. 6.

2 A Short History of Language Models

Large Language Models are neural networks with millions of parameters that represent adjustable weights in the network that are optimized during training to predict the next word in a sequence of words. During training, understanding context, as well as the relationship of tokens in a text, is considered to pay attention to specific parts of the input that are relevant in making predictions.

These models arise from the field of Natural Language Processing (NLP) and are used for the purpose of understanding and producing natural language text. [12].

Language models have evolved greatly from the earliest to the present day. The first ones were statistical models, and their operation consisted of predicting words through the use of different statistical techniques. These had major limitations due to the limited predictive power of the statistical techniques used, the small size of the datasets used, and the limited computing power available in the 1990 s. At the turn of the century, neural networks began to be used for the purpose of making these predictions on text sequences. The learning of these networks was limited by the same factors as previous statistical models, the small dataset used, and the scarcity of computational resources, achieving unremarkable results. In the decade that followed, another type of neural network began to be used that was more suitable for this task: the Long short-term memory (LSTM) network. This type of network belongs to the Recurrent Neural Networks (RNN) and is focused on the processing of sequences [3]. Moreover, thanks to the constant development of computer hardware featuring a greater amount of computational resources and the emergence of the cloud, models were trained with larger datasets, and more robust models were created. The main problem that arose then was the memory limitation of these networks, causing failures in the processing of long texts and relating the context with previous fragments.

The latest breakthrough in NLP is the Transformer Networks used by the state-of-the-art LLMs. These neural networks have a large long-term memory, and this characteristic makes it possible to analyse longer text strings, unlike the neural networks previously used in this type of generative models. Moreover, due to the way these networks process the data, a better analysis of the relationships between words, sentences and even fragments at different levels is achieved. This feature allows the language model to improve the interpretation of the text, understanding the context and making it more similar to the interpretation of the human being himself [17].

There are three types of Transformer architectures:

- (a) encoder-decoder [17], where the self-attention mechanism tunes each token weigh depending on the context of the entire sequence, capturing relationships and dependencies between the different parts of the input. Encoder-decoder architectures are suitable for applications like machine translation, text summarization and question-answering systems. An example of this architecture is T5 [14].
- (b) the encoder-only model outputs vectors generated by the encoder that are used as input to a classifier to make predictions. Some applications are text classification and sentiment analysis; BERT architecture is an encoder-only model [8].
- (c) decoder-only models are used in text generation. They only take into account the previous tokens to predict the next token in a sequence; Open AI GPT (Generative Pretraining models) has adopted a decoder-only model

[13]. All these models can be tailored to specific tasks and domains. Recently, Zhao et al. have presented an extensive review of current LLMs [19].

LLMs are pre-trained models that can be used "off the shelf", but they might require further fine-tuning to enhance their capacity. Fine-tuning could be performed in two ways: training on additional data for adapting to different domains or using prompt engineering to improve the output of the model. Prompts are instructions that provide guidance during the text generation process that include detail and context to the input. There are several strategies like zero-shot, one-shot, and few-shot prompting depending on the number of examples (input-output) the user provides to the model. Moreover, some instructions or constraints can be provided to improve the prompt. Min et al. describe in detail several works applying pre-training and then fine-tuning, prompting, and text generation approaches [11]. Fine-tuning on additional data might require many annotated examples, could be expensive in computing resources, and it can increase the carbon footprint.

3 The Proposed System

The conversation agent proposed in this work is integrated into a social robot as part of one of the functionalities that this robot offers to its users. The robot interacts verbally with its users thanks to two modules: the Automatic Speech Recognition (ASR) module and the Text-to-Speech (TTS) module. Both modules are used as the interface between the user and the LLM that implements the conversational agent. The conversational flow starts either with the user's utterance, if the user takes the initiative, or with the robot's utterance, in case it takes the initiative. In case the user takes the first step, the robot collects this audio fragment through its microphones, and it is sent to the ASR engine. The ASR translates the audio signal into the corresponding text transcription, and this is input into an LLM. After processing the input, the LLM generates the response as a text that the TTS engine synthesizes, producing the robot's speech. This is a cyclical process that is repeated over time until the conversation is over. This process is shown in Fig. 1.

For the conversational agent, we have considered several LLM. After evaluating their performance, their inference time, and their cost, we opted to use GPT-3.5, in particular gpt-3.5-turbo. We set the *temperature* parameter to 0.3, which corresponds to a low degree of randomness in the output of the model, and used the OpenAI API to access this model online.

3.1 Prompting

In the context of Artificial Intelligence (AI), a prompt is an instruction that triggers a response from an AI model, allowing interaction between the user and the AI system. These models do not perform actions on their own but respond to instructions (prompts) provided by users. Prompting engineering, or simply

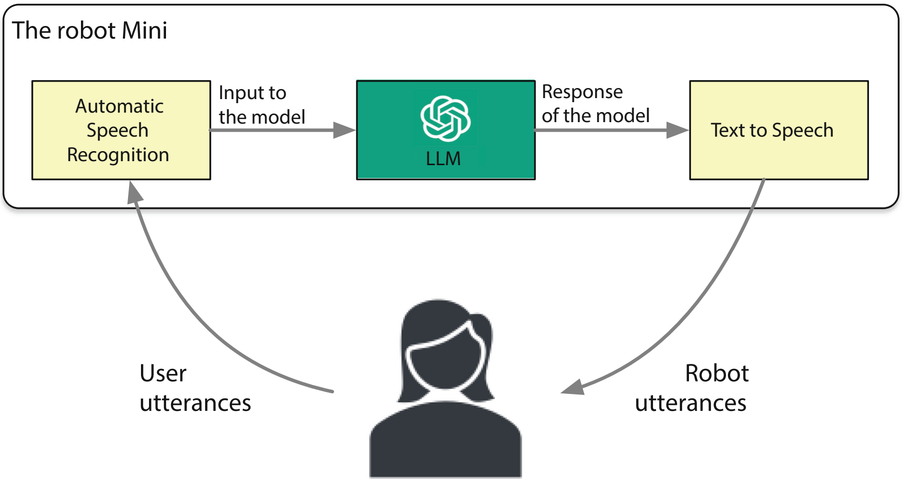

Fig. 1. Overview of the proposed conversational agent in the robot Mini

prompting, refers to the process of providing prompts to the generative AI system and is crucial to obtaining appropriate responses. The quality of the prompt can significantly influence the response generated by the AI. The quality of these prompts is critical for successful results.

Therefor, defining the right prompt for the task at hand is crucial. In our case, the prompt needs to clearly describe the task, i.e., operating as a conversational assistant, and the context, i.e., interacting with elderly people. Some general considerations for designing prompts include: describing tasks and context, generating short responses to speed up communication, adopting a familiar tone, and taking the initiative in the conversation [6].

In this work, we have followed an iterative process during the design of the final prompt for the model used in the conversational agent. After each iteration, the result was tested and compared with the model's responses to the same questions with the previous prompts. In this way, the prompt was adjusted after each iteration and tested again so that it was aligned with the objectives proposed for the assistant. The final prompt consists of three blocks included in Tables 1, 2, and 3:

1. Context block: Establishes the purpose of the assistant and its focus on older people, underlining the importance of natural and friendly conversations (see Table 1).
2. Instructions Block: Details specific instructions, from starting the conversation to answering questions and maintaining an appropriate tone (see Table 2).
3. Additional Information Block: Provides details about the attendee's purpose, priorities and general guidelines for positive interactions (see Table 3).

Table 1. Prompt Context block.

BLOCK 1: CONTEXT
Your name is Chatbot and you are a conversational support service for the elderly. \
The first sentence of the prompt is intended to enable the model to get a sense of the task for which it is intended.
Your task is to maintain a conversation on a selected topic with the user. \
With this second interaction, the model is able to know the context of the situation and adapt to it.

Table 2. Prompt Instructions block.

BLOCK 2: MAIN INSTRUCTIONS
First you greet the user, then you ask them how they are, then you ask if they want to start a conversation on a topic of their choice. And you tell them that if they want to change the subject at any time, all they have to do is let you know. \
After this brief introduction to the task to be carried out, the assistant is told how to carry out the first interaction with the user. The user is greeted with a first greeting, followed by a first wildcard question about his or her personal status in order to have a first topic to talk about. Subsequently, it is proposed to start a conversation on a topic and the user is told that he/she can change the topic at any time he/she wishes.
You wait for his response, and begin to engage him in a conversation through questions and answers on the subject.\
This indication provided to the model is the instruction for the model to start and maintain a conversation with the user on the topic proposed by the user, providing questions and answers on the topic.
Don't just wait for a response from the user, take the initiative in the conversation. \
The next instruction is of great importance as the robot has to be active in the conversation and with the user, in order to prolong the conversation so that it does not reach a point where it stops. The model has to achieve an interactive conversation on both sides.
You respond in a very brief, conversational, approachable and friendly style. \
The fifth proposal instructs the model to use a response style focused on achieving fluency and closeness in the conversation, which is of vital importance so that the conversation does not become tiresome for the user.
Don't focus on providing a lot of scientific data, but more on an informal and close conversation.\
Continuing with this trend of making the conversation fluid, the following instruction is included to emphasise to the model the type of interaction that is desired with the user and the importance of the conversation being informal and close so that the user feels comfortable and safe interacting with the robot.
Limit your answers to three sentences. \
The maximum length of the response is defined at the end of the instructions concerning the structure of the conversation. In this way, it is possible to make the conversation even more fluid, preventing the model's response from being too long and containing too much information that could saturate the user.

Table 3. Additional Prompt Instructions block.

BLOCK 3: ADDITIONAL INSTRUCTIONS
Some of the topics you can suggest to talk about with the user are: History, geography, art, literature, sports, talking about your life, practising languages. \
Finally, there are some topics that the assistant can talk about with the user, the central theme has to focus on talking about the user's life and the topics of the user's choice. Even so, there are some general topics about which the assistant can suggest a conversation to the user in case the user does not take the initiative.
Play a game proposed by you or by the user, some games can be: Guess the song, trivia questions or continue with the proverbs. \
Also, it is of great value the possibility for the wizard to propose the user to play a game to pass the time, such as those mentioned in the last lines of the code.

4 Integration into Mini

In this work, we have considered the social robot Mini (see Fig. 2). Mini is a desktop robot intended for seniors to accompany, support, and assist them in their daily activities [16]. It has been designed and built by the Robotics Lab, from Universidad Carlos III de Madrid. Mini is equipped with touch sensors to detect when and how it is touched, a microphone to capture its users' voice and other audio signals from the environment, and a tablet to extend the interaction capabilities of the robot. Besides, Mini can move its head, arms, and body, can change the color of its cheeks and beating heart, has a vumeter-like mouth, and has a pair of screen-based animated eyes.

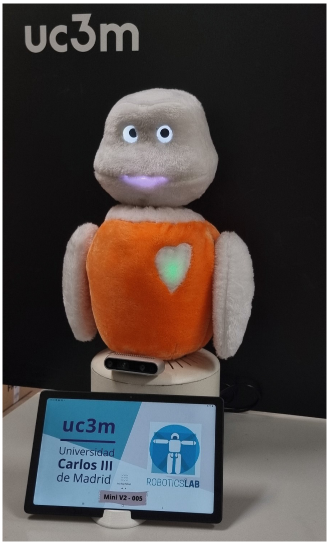

Fig. 2. The social robot Mini

In terms of the software architecture, the robot Mini has five main elements (see Fig. 3): (i) the skills represent the repertoire of functionalities that are offered to Mini users; (ii) the Decision-Making System (DMS) selects the skill that needs to be activated at each moment depending on external (e.g. user preferences) and internal (e.g. internal motivations) events [10]; the HRI Manager orchestrates the multimodal human-robot interaction using Communicative Acts (CAs) that handle the exchange of information between Mini and a user [4]; (iv) the Perception System manages the low-level communication and configuration of the different sensors; and (v) the Expression System orchestrates the robot actuators to communicate a coherent multimodal message in a timely manner [5].

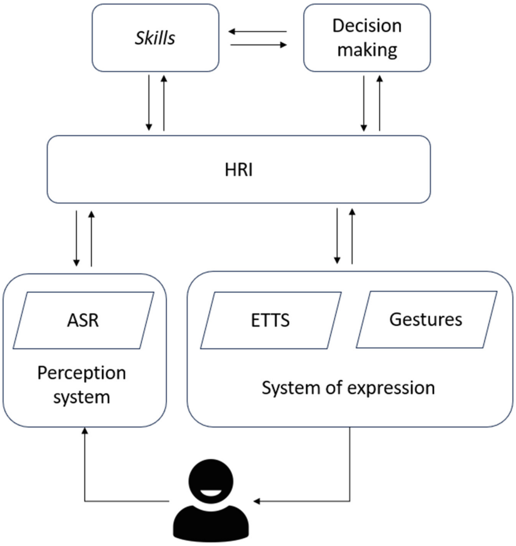

Fig. 3. Diagram of the Mini software architecture.

For this work, we have created a new skill for allowing Mini to operate as a conversational agent using verbal communication. This skill will be activated by the DMS and will request the HRI Manager to execute the proper CAs to give information to the user, to ask questions, or when waiting from some ipnut from the user.

Focusing on the Perception System, we have integrated the ASR engine from Google[1]. This is a grammar-free voice-to-text tool that is executed in the Google Cloud. This service provides a literal transcription of the user speech that will feed to our LLM.

In the Expression System, the robot utterances are generated by the commercial TTS ReadSpeaker[2]. In combination with the TTS, Mini accompanies its

[1] https://cloud.google.com/speech-to-text/.

[2] https://www.readspeaker.com.

utterances with non-verbal gestures that have been defined offline. During the operation of the conversational agent skill, the robot performs smooth random movements of its arms, head, and body in order to give Mini a lively appearance.

4.1 Design of the *Conversational Agent Skill*

The *Conversational Agent Skill* has been modeled as a state machine with four states: *greeting*, *conversation*, *continuing*, and *ending*. The flowchart is shown in Fig. 4.

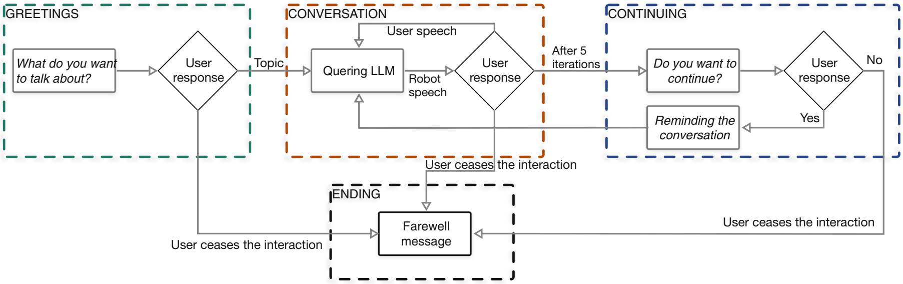

Fig. 4. Diagram of the Conversational Agent Skill.

When the DMS activates the skill, the skill initiates in the state *greeting*, where the robot asks the user what (s)he wants to talk about. After the user responds, the skill moves to *conversation* state. In this state, the transcription of the user's response provided by the ASR is input to the model, and the output is collected and synthesized by the TTS. Then, the user can continue the conversation with other response. This dialog continues until a certain number of turns, i.e. five turns, is reached. At this point, the skill advances to the *continuing* state, where the robot asks the user if (s)he wishes to continue the conversation. If the user answers no, it advances to the *ending* state, where the robot launches a farewell message, and the conversation ends. Otherwise, if the user replies affirmative, the robot reminds the user what they were talking about before the skill transits to the *conversation* state to continue the dialog.

It is important to mention that our skill can stop at any moment if the user does not want to continue or if the system does not receive a response (for example, if the user is gone).

5 Evaluation

We have evaluated the Conversational Agent Skill with 24 users (16 of them were male) interacting freely with the robot. Participants were faculty members and graduate students and 75% of them were between 23 and 27 years old.

Prior to the interaction with the robot, an experimenter presented Mini to the participants and they were informed of the skills and capabilities of Mini. They have also been informed of the data protection policy and have signed a consent form. Data collection was conducted in compliance with the Data Protection Regulations of Universidad Carlos III de Madrid.

During the evaluation, the participants were asked to keep a conversation with Mini about one topic that they decided following the interaction flow described in Fig. 4. Participants could stop the interaction at any moment.

After their interaction, participants completed the System Usability Scale questionnaire [2] to measure the perceived ease of use of the Conversational Agent Skill. The resulting SUS score was 78.5 points, which means good usability according to Bangor et al.'s adjective rating scale [1].

After that, participants were asked to rate the quality of the interaction with Mini, ranging from 0 (the lowest value) to 10 (the highest value). The obtained average value was 8.1, which represents a very high participants' satisfaction with our system.

Finally, participants completed two open questions about the positive and negative aspects of the interaction. The most frequent positive aspects were the ability to discuss any topic, Mini's spontaneity and coherence in its responses, and the generation of engaging conversations. On the other hand, the most frequent negative aspects were ASR failures, some unusual interactions, and delays in Mini's responses.

6 Conclusions

In this paper, we have presented a new skill that allows our social robot Mini to operate as a conversational agent. The core element of this skill is the GPT-3.5 LLM that has been integrated with the robot's software architecture. After evaluating the skill in real interactions, it has exceeded our initial expectations, standing out for the quality of conversations and its high usability. This implies a great opportunity to endow this kind of functionality in social robots that operate as companions for elderly people.

Although the positive results, we have observed some limitations. Using LLM in the cloud adds extra delays that might not be acceptable for a system interacting with humans in real time and requires a stable high-speed Internet connection. New research in reducing the size of LLM (small LLM) with fewer parameters offers new opportunities to get faster LLM and, additionally, with lower power consumption. Because of the extensive use of generative AI, recent small LLMs enhance ecological sustainability and reduce carbon footprint. Also, users have to be aware that the LLM output is not always correct and might suffer from "hallucinations". To cope with these problems, Retrieval-Augmented Generation (RAG) [9], a more sophisticated way of customizing generative LLMs, has emerged. RAG combines information retrieval with text generation, which helps to provide more accurate and contextually relevant answers. Concerning user interaction, some problems experienced by the participants were due to the limitations of the ASR in certain circumstances, such as very short sentences, noisy environments, or a limited range of operation of the robot's microphones.

Acknowledgments. This work was partially supported by ACCESS2MEET project (PID2020-116527RB-I0) supported by MCIN AEI/10.13039/501100011033/ and Madrid Regional Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17) and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation. This work has been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M ("Fostering Young Doctors Research", SMM4HRI-CM-UC3M), and in the context of the V PRICIT (Research and Technological Innovation Regional Programme). This work has been partially supported by the project "Robots sociales para mitigar la soledad y el aislamiento en mayores (SOROLI)", funded by Agencia Estatal de Investigación (AEI), Spanish Ministerio de Ciencia e Innovación (PID2021-123941OA-I00).

Disclosure of Interests. The authors have no competing interests.

References

1. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. **4**(3), 114–123 (2009)
2. Brooke, J.: SUS: a 'quick' and 'dirty' usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (eds.) Usability Evaluation in Industry, vol. 21, pp. 189–194. Taylor and Francis (1996)
3. Dupond, S.: A thorough review on the current advance of neural network structures. Annu. Rev. Control. **14**(14), 200–230 (2019)
4. Fernández-Rodicio, E., Castro-González, l., Alonso-Martín, F., Maroto-Gómez, M., Salichs, M.: Modelling multimodal dialogues for social robots using communicative acts. Sensors (Basel). **20**(12), 3440 (2020). https://doi.org/10.3390/s20123440
5. Fernández-Rodicio, E., Maroto-Gómez, M., Castro-González, l., Malfaz, M., Salichs, M.: Emotion and mood blending in embodied artificial agents: expressing affective states in the mini social robot. Int. J. Soc. Robot. **14**(8), 1841–1864 (2022). https://doi.org/10.1007/s12369-022-00915-9
6. Fulford, I., Ng, A.: ChatGPT prompt engineering for developers (2023). https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/
7. Hameed, I.: Using natural language processing (NLP) for designing socially intelligent robots. In: 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), pp. 268–269 (2016). https://doi.org/10.1109/DEVLRN.2016.7846830
8. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, vol. 1, p. 2 (2019)
9. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NIPS 2020: Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 9459–9474 (2020)
10. Maroto-Gómez, M., Castro-González, l., Castillo, J.C., Malfaz, M., Salichs, M.N.: An adaptive decision-making system supported on user preference predictions for human-robot interactive communication. User Model. User-Adap. Interact. **33**(2), 359–403 (2023). https://doi.org/10.1007/s11257-022-09321-2

11. Min, B., et al.: Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput. Surv. **56**(2), 140 (2023). https://doi.org/10.1145/3605943
12. Peng, B., et al.: RWKV: reinventing RNNS for the transformer era. arxiv preprint (2023)
13. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Technical report, OpenAI (2018)
14. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 1–67 (2020)
15. Salcedo, J.S., Martínez, S.C., Montoya, J.C.C., Castro-Gonzalez, A., Salichs, M.A.: Modelos de lenguaje natural para robots sociales. XLIII Jornadas de Automática (2022).https://doi.org/10.17979/spudc.9788497498418.0828
16. Salichs, M.A., et al.: Mini: a new social robot for the elderly. Int. J. Soc. Robot **12**(6), 1231–1249 (2020). https://doi.org/10.1007/s12369-020-00687-0
17. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) 31st Conference on Neural Information Processing Systems (NIPS 2017)**30**. Curran Associates, Inc. (2017)
18. Wahde, M., Virgolin, M.: Conversational agents: theory and applications. arXiv arXiv:2202.03164 (2022). https://api.semanticscholar.org/CorpusID:246634059
19. Zhao, P., Jin, Z., Cheng, N.: An in-depth survey of large language model-based artificial intelligence agents. arXiv preprint arXiv:2309.14365 (2023)