Biomimetic Intelligence and Robotics 3 (2023) 100131

Contents lists available at ScienceDirect

Biomimetic Intelligence and Robotics

journal homepage: www.elsevier.com/locate/birob

Review

Large language models for human–robot interaction: A review

Ceng Zhang [a], Junxin Chen [b], Jiatong Li [b], Yanhong Peng [c,*], Zebing Mao [d,*]

[a] Department of Mechanical Engineering, National University of Singapore, Singapore 119077, Singapore
[b] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing 100044, China
[c] Department of Information and Communication Engineering, Graduate School of Engineering, Nagoya University, Nagoya 464-8603, Japan
[d] Department of Mechanical Engineering, Tokyo Institute of Technology, Tokyo 152-8550, Japan

ARTICLE INFO

Article history:
Received 17 September 2023
Revised 19 October 2023
Accepted 24 October 2023
Available online 28 October 2023

Keywords:
Large language models
Human–robot interaction
Task completion
Considerations and challenges

ABSTRACT

The fusion of large language models and robotic systems has introduced a transformative paradigm in human–robot interaction, offering unparalleled capabilities in natural language understanding and task execution. This review paper offers a comprehensive analysis of this nascent but rapidly evolving domain, spotlighting the recent advances of Large Language Models (LLMs) in enhancing their structures and performances, particularly in terms of multimodal input handling, high-level reasoning, and plan generation. Moreover, it probes the current methodologies that integrate LLMs into robotic systems for complex task completion, from traditional probabilistic models to the utilization of value functions and metrics for optimal decision-making. Despite these advancements, the paper also reveals the formidable challenges that confront the field, such as contextual understanding, data privacy and ethical considerations. To our best knowledge, this is the first study to comprehensively analyze the advances and considerations of LLMs in Human–Robot Interaction (HRI) based on recent progress, which provides potential avenues for further research.

1. Introduction

The past decade has witnessed remarkable progress in Human–Robot Interaction (HRI), which is now commonplace in everyday life. From robots helping with household tasks to humans and robots working together in production lines, it has a significant impact on human society. The key to research in this area is understanding the nature of interactivity and social behavior between robots and humans [1], which aims at endowing robots with the ability to estimate the intent of humans and to complete tasks more effectively that meet the requirements of users.

The emergence of LLMs has opened the potential to address long-standing challenges through their remarkable capacity for reasoning and generation. Initially developed for text-based tasks such as translation, they are now being applied in robotics. Equipped with common sense, they are capable of interpreting ambiguous human language into concrete responses, as well as processing multimodal input and generating the desired output. The techniques applied are essential for further progress [2], and many recent studies have focused on improving the structure of the model to improve robustness and efficiency. Exploring the use of language models in robot planning is a rapidly growing field of research. Studies have demonstrated that the effectiveness of these models in text tasks can be extended to embodied control [3–5]. New advances are being made every once in a while, leading to the development of intelligent HRI and allowing robots to be more easily integrated into human society.

To our knowledge, this is the first work to give a comprehensive review on recent advances of large language models in human–robot interaction, and our contributions can be summarized as follows:

- We provide an overview of the recent advances in techniques used in HRI and discuss how they are being applied in various fields and contexts, giving an insight of the progress made in different areas of HRI.
- We introduce mainstream LLMs in various categories and training approaches, while the concepts of base models are elucidated through mathematical analysis, making it easier to comprehend their principles and the key point for improvement.
- The cutting-edge applications of LLMs in HRI are presented, with a focus on the various areas and the core techniques being examined and evaluated.
- The current issues and boundaries of LLMs are outlined, and from this, we suggest potential avenues for further exploration in this area.

The arrangement of content is as follows: first a brief taxon of HRI and mainstream techniques used in LLMs are introduced,

* Corresponding authors.
E-mail addresses: yhpeng@nagoya-u.jp (Y. Peng), mao.z.aa@m.titech.ac.jp (Z. Mao).

https://doi.org/10.1016/j.birob.2023.100131

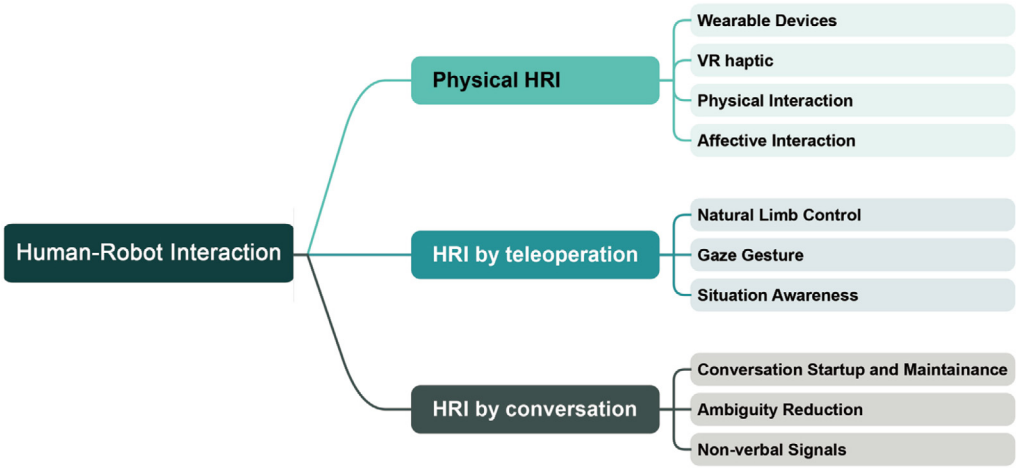

Fig. 1. The intuitive list of the directions of work in HRI and the criteria for selection.

then we list recent breakthroughs of LLMs applied in HRI according to three different domains. In the end, we point out existing challenges faced by this technology and potential directions for future research.

2. Brief and methods of human–robot interaction

In recent years, HRI, which previously adopted traditional control techniques, is being redefined due to the rise of artificial intelligence. Currently, it has a multifaceted scope of research that contains considerations from basic techniques to high-level ethics; as a result, it has applications in various fields, as mentioned in [6]. These areas include but are not limited to human's monitoring and control of robots in routine tasks, work in dangerous and confined areas with robots and socialization by robots with humans to provide assistance for those in need. In addition to different applications, the taxonomy of HRI can be defined according to various approaches differing in interactions with human as shown in Fig. 1.

2.1. Physical human–robot interaction

Interactions without a medium, which is referred to as physical HRI [7], is a direct and intuitive approach that allows humans and robots to communicate effectively and provides a wealth of tactile information, which is illustrated in Fig. 2, and major works in this field are concluded in Table 1. In the early stages of robotics, when robots lacked tactile senses, traditional haptic HRI was typically achieved through wearable devices that enabled the perception of force between humans and robots. Kim et al. in [8] developed a kind of exoskeleton masterarms that can detect the torque applied by the user and its direction so that the device allows free or restricted movement and provides feedback forces in a multimodal contact. Another reason for wearable devices is that haptics can be overloaded when direct physical contact between human and robots leads them to be used for both performing tasks and communications [9]. Hence, a good example of this cleavage is [10] in which a bracelet with vibrating motors is proposed to complete robot follower formation tasks led by the human user's trajectories. Based on this, the development of virtual reality (VR) techniques has adapted these devices into virtual environments. Several motion generation algorithms are presented in [11] that allow a robot to find the optimal path in virtual with obstacles giving haptic feedback for human perception. Coupled with the application of VR display, Wang et al. [12] presented a system using hidden Markov models that allows a human user to physically shake hands with a virtual partner similar to a real one via a tactile interface giving multimodal feedback signals.

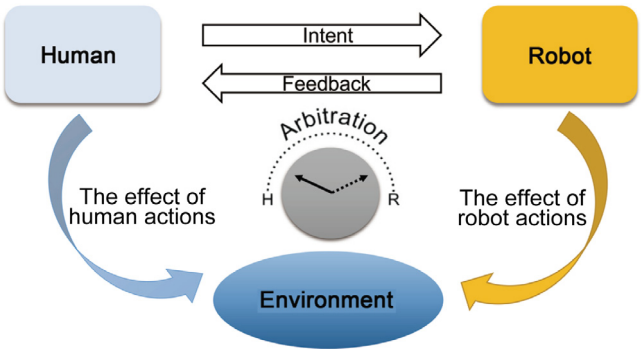

Fig. 2. Human–robot interaction. The information exchange between human and robot is achieved by interactions with environment depending on which side the arbitration leans towards human or robots [13].

While adoption of VR considerably circumvents the complexity of modeling and relieves limitations existing in reality, challenges like device rendering speed can delay the arrival of agents to certain targets, causing discrepancies between user's visual and tactile information that degrade the experience. To address this, motion strategies were presented in [14] using the eye gaze direction and user's hands for motion prediction, which greatly reduces the detection time of the input. Rather than relying on wearable devices, another successful approach is to directly interact with mobile robots. A common use of this technology is the collaboration between humans and robots to complete a task, with the main focus being the interpretation of the partner's intentions. Aydin et al. [15] took moving the table as an example and proposed controllers for the robot to detect the intent of the human partner and automatically adjust its contribution to the task. And more interestingly, by [16] in the cooperated wart, the human body dynamics is approximated by an inverted pendulum and the robot is able to estimate its state by an extended Kalman filter (EKF) so that it can perform dance according to the human rhythm. The effectiveness of this approach was also demonstrated in another review by Losey et al. [13], in which the time-dependent Kalman filtering technique was proven to outperform the state-of-the-art RBF method in both speed and robustness.

Table 1
Adoption of methods and techniques in recent works related to physical HRI.

Year	Author	Technique	Perception modality	Achievement
2005	Kim et al. [8]	Wearable devices	Proprioceptive	An exoskeleton-type masterarm with actuators for force reflection.
2022	Gutierrez et al. [11]	VR	Haptic and vision	A motion scheme with obstacle avoidance in VR environment.
2014	Aydin et al. [15]	Physical interaction	Haptic	A controller for the robotic partner that cooperates with human.
2016	Bianchi et al. [17]	Affective interaction	Interoceptive	Discrimination between different caressing stimuli through the analysis of physiological signals.
2012	Scheggi et al. [10]	Wearable devices	Haptic	A haptic bracelet which helps the human to perform leader–follower for robots.
2012	Yohanan and MacLean [20]	Affective interaction	Haptic and interoceptive	Investigation on the human"s intent and expectations when displaying emotional touch by a haptic robot.
2012	Wang and Kosuge [16]	Physical interaction	Haptic	A robot follower that can dance with the human leader.
2011	Wang et al. [12]	VR	Haptic and Vision	Development of a VR system allowing a human user to handshakes with a partner through a haptic interface.
2022	Mugisha et al. [14]	VR	Haptic and proprioceptive	Motion strategies based on the hand motion and eye-gaze direction to determine the point of user interaction in a virtual environment.
2008	Yohanan and MacLean [19]	Affective interaction	Haptic and interoceptive	Development of an affect display through touch in the context of social interaction between human and robot.
2023	Peng et al. [26]	Wearable devices	Proprioceptive	A human muscle configuration inspired wearable assist device to induce natural kinesthetic perception by torso.

Based on intention estimation, researchers have extended this field to affective interaction to gain a deep understanding of emotional changes of human during the process. Early studies [17,18] have demonstrated that a haptic device that is stroked can have an effect on the autonomic nervous system, which is a key factor in the production of emotions in humans. Similarly, work such as [19,20] designed a haptic creature and presented a dictionary that associates different caress motions with the corresponding intended emotions, thus facilitating the robot to respond appropriately to meet user expectations. In this sense, non-verbal HRI can affect humans in several aspects, from cognition, emotions, behaviors [21], thus it has great potential to be applied to other areas such as education and medical treatment.

2.2. HRI by teleoperation

This approach takes advantage of the concept of controlling robots indirectly, usually by sending commands remotely through a console with functional buttons, which is beneficial as it allows robots to work in hazardous areas and tight spaces that are not accessible to humans. However, its effectiveness largely depends on the mastery and expertise of the operator; more importantly, this method of operation does not comfort with human habits, as it lacks the natural body language. To address this, Tsetserukou et al. [22] designed a robot arm that can teleoperate by movement of the human limb. Also, combined with a joint impedance algorithm, it is capable of detecting the contact of the arm with an object; therefore, the manipulator can interact safely with an unstructured environment. Similarly, Peppoloni et al. [23] proposed an integrated interface in ROS that allows the user to remotely control robots by hand motion and adjust the autonomy of the robot between different levels online. But as mentioned in [24], this also brings up a problem: the slave robot must be integrated into the operator's sensory-motor environment as if it were an extension of their body, however, the size of the two spaces can be vastly different, which significantly limits its practical application. To this end, Zhao et al. [25] applied electromyography (EMG) signals in teleoperation to change the space drag and impedance, which solves the spatial constraints of natural limb control while maintaining the user's posture.

While the works mentioned above greatly simplify teleoperation by natural limb motion, it remains a challenge for the handicapped who have difficulty moving their hands to issue signals to robots. As an effective measure, Yu et al. [27] proposed gaze gestures as an object selection strategy into Human–Robot Interaction for drone teleoperation. Based on this, Duenser et al. [30] compared teleoperation with eye tracking-based interfaces with traditional mouse and touchscreen interfaces with the results showing it outperforms especially in reaching tasks when the target is far from the initial point, but also pointed out its suffers from accuracy compared with the approach by touching, and how to improve this is still worthy of research. Besides that, teleoperation assisted with advanced display has broadened its field of application, Su et al. [31] introduces a tele-manipulation paradigm that is easy to use, based on Mixed Reality Subspace (MRS). Additionally, it provides an evaluation of various control and visual feedback HRI modes on a robotic arm-hand platform. More practically, an improved human–robot collaborative control (IHRCC) scheme is proposed in [32] to teleoperate Minimally Invasive Surgery (MIS) based on a hierarchical operational space formulation of a redundant robot.

Investigations have demonstrated the potential of teleoperation in a variety of applications shown in Table 2. Furthermore, situational awareness (AW) has recently become a focus of attention when creating new control systems. It refers to how humans perceive, understand, and respond effectively to one"s situation, especially important in this case since it directly affects the performance of the robot being manipulated. In this regard, Gatsoulis et al. [28] conducted a comparison of various techniques for measuring the SA of a person who is teleoperating a robot and determined which have the greatest potential to produce more reliable and precise results. Based on that, a structure was proposed in [29] to design interfaces that can improve the user's SA with the help of augmented reality, providing visual feedback on the robot's camera capabilities. Therefore, with the continuous rise of other techniques, the method of robot teleoperation is being extended to more fields with its unique advantages.

2.3. HRI by conversation

Dialogue acts as a natural way of human communication, applying it to human–robot interaction is conducive to the integration of robots into human society. However, the challenges are not limited to the need for a linguistic understanding of how conversations work and practical knowledge of how to collaborate on tasks, but also the robot must be able to interpret and generate behaviors that demonstrate the intention to keep the

Table 2
Adoption of methods and techniques in recent works related to HRI by teleoperation.

Year	Author	Technique	Perception modality	Achievement
2007	Tsetseruko et al. [22]	Natural limb control	Haptic and proprioceptive	A teleoperated robot arm enabling torque measurement in each joint.
2014	Yu et al. [27]	Gaze gesture	Vision	Gaze gestures as an object selection strategy into HRI for drone teleoperation
2010	Gatsoulis et al. [28]	Situation awareness	Interoceptive	Analysis on main aspects in developing reliable SA measurement methods.
2018	Hedayati et al. [29]	Situation awareness	Vision and haptic	Interface designs that provide users with augmented reality feedback while teleoperating an aerial robot.
2015	Duense et al. [30]	Gaze gesture	Vision	Comparison between manual interaction interfaces with eye-tracking-based techniques
2015	Peppoloni et al. [23]	Natural limb control	Proprioceptive	An integrated interface that allows the user to teleoperate the robot using hands motion.
2019	Zhao et al. [25]	Natural limb control	Proprioceptive	Human bioelectrical information used in virtual reality interactions.

Table 3
Adoption of methods and techniques in recent works related to HRI by conversation.

Year	Author	Technique	Perception modality	Achievement
2003	Sidner et al. [33]	Conversation startup	Auditory and Visual	An architecture of engagement in human–robot collaborative conversations.
2015	Shi et al. [34]	Conversation startup	Proprioceptive and auditory	A conversation startup model involving the participation state and spatial formation.
2018	Papaioannou et al. [35]	Ambiguity reduction	Auditory	Design of an architecture that supports multi-threaded task planning dialogues.
2012	Han et al. [36]	Nonverbal signals	Visual and proprioceptive	Investigation on use of Non-Verbal Cues from sensors with the Nao platform.
2022	Dogan et al. [37]	Ambiguity reduction	Auditory	An interactive system asking for follow-up clarifications to disambiguate the described objects using accessible information.
2006	Sidner et al. [38]	Nonverbal signals	Visual and proprioceptive	Investigation on feedback by head nods from human to the robot in conversations.
2019	Vega et al. [39]	Conversation startup	Auditory	A navigation planning method which allow the robot to navigate in crowded environments in a socially acceptable way.
2019	Li et al. [40]	Nonverbal signals	Interoceptive	An emotion processing module that consists of a user emotion recognition function and a reactive emotion expression function.
2009	Mutlu et al. [41]	Nonverbal signals	Visual and proprioceptive	Design of a set of gaze behaviors for robot to signal different kinds of participant roles.

conversation going or make a termination [33]. To this end, it is important for robots to understand conventional manners to make the conversation smooth and acceptable, Shi et al. [34] developed a participation state model to measure communication participation and methods for structuring a robot's behavior to initiate a conversation at a proper distance as well as maintaining. A more practical application is that Vega et al. [39] proposed a planning domain that allows robots to perform navigation tasks in a crowded environment in an acceptable manner, for which the robot can request permission when the path is blocked. During the conversation, a valuable study in [42] investigated the role of robots' engagement in hosting activities, in which humans naturally and frequently interact with one another and this type of interaction can be replicated in human–robot interactions.

However, a problem arises when people do not interpret specifications in a concrete manner, which can cause confusion for robots if there are similar entities in the environment and thus have a serious impact on robots' ability to complete tasks. To address this, Papaioannou et al. [35] developed a flexible and expandable system that can account for the various dialogue acts that may occur during the execution of the task, thus allowing the handling of all potential variations. An alternate approach to address ambiguities is suggested in [37], the system involves an interactive process that seeks additional information to distinguish the described objects, which is done utilizing the knowledge that the robot can comprehend from the request and the objects in the environment that the robot is aware of.

In addition to the words spoken, humans also use nonverbal signals during communication, and the robot needs to comprehend their significance in order to react in a way that is satisfactory to humans. Previous research has explored techniques to explore the use of nonverbal cues in human–robot interaction with the Nao platform, which is composed of a variety of sensors to collect various information from human [36]. Specifically, on the one hand, the work of [38] investigated the role of head nods in communication and showed that robot nodding in response to human words leads to more engaging conversations in HRI; On the other hand, Mutlu et al. [41] investigated how a robot can identify the roles of its conversational partners, whether addressee, bystander, or overhearer, by using gaze cues. In addition to the subtle expressions that can provide a great deal of information about the participants, other signals, such as the assessment of emotions from prosody and sentiment analysis [40], can be utilized by robots to respond appropriately. It is essential to take into account the ability to comprehend signals when building robots for future human–robot collaborations (see Table 3).

Despite the numerous international conferences and professional scientists who are continuously creating new ideas in this area like achievements shown in Table 3, there are still a lot of issues that have yet to be resolved. A key of these is the extent to which robot takes over the task, for simple handling operations in warehouse, robots can achieve complete automation, but human intervention is essential in fields that require high precision such as aerospace and surgery, and there are still considerable risks like in the safety of self-driving vehicles. It is worth noting from the research in recent years that to minimize the labor-intensive operations by human, communication through chatting is an effective way to exchange information in HRI, where a challenge lies that how the robot transforms our languages into representations that it can understand. Thanks to the rise of large language models, there are more possibilities for problem-solving ideas, which are to be introduced in the following of this article.

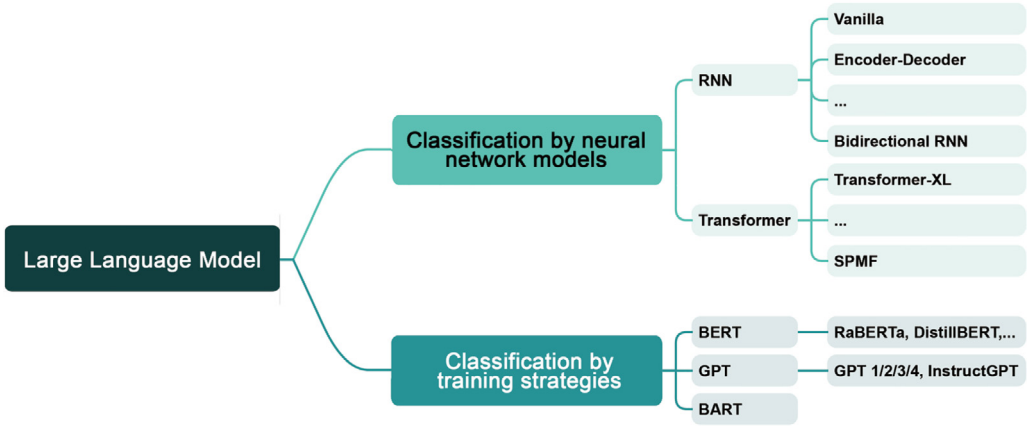

Fig. 3. The intuitive list of the directions of work in LLMs and the criteria for selection.

3. Recent advances in large language models

In the past decade, Natural Language Processing has been revolutionized by the emergence of LLMs. This has caused a surge of progress and creativity in the field. These cutting-edge models, powered by massive amounts of data and intricate neural network architectures once applied in other areas of modeling [43,44], have demonstrated unprecedented capabilities in understanding, generating, and manipulating human language. From the renowned GPT to its successors, these massive language models have demonstrated their capabilities in a variety of tasks, such as text production, translation, sentiment analysis, and question-answering. As researchers and developers continue to expand the limits of what is achievable, the influence of large language models on various industries and everyday life is becoming more and more significant. In the following content of this section, the recent advances of mainstream language models are investigated with their taxonomies in Fig. 3 according to multiple criteria to provide a brief understanding of different LLMs that are commonly applied in design interactive robots.

3.1. Classification by neural network models

3.1.1. Recurrent neural network based language models

Before the appearance of Transformer, RNN has always been an optimal choice of the design of the language model. When comparing traditional feedforward neural network with RNN, the former suffers from the limitation that the history is represented only by the specific number of predecessors, in contrast, the latter can possibly acquire knowledge of context dependencies that span more than a predetermined number of words that come before [45], thus theoretically having a context of unlimited length. As a result, the Recurrent Neural Network Language Model (RNNLM) proposed in [46] outperforms the one based on the Feedforward Neural Network by forming short-term memory to better deal with position invariance.

The strength of RNNML comes from its network structure, as shown in Fig. 4(a), it consists of an input layer x, a hidden layer h and an output layer y. With the corresponding weight matrices between the connected layers, the variables in the network can be calculated as

$$h(t) = f(Ux(t) + Wh(t-1)) \tag{1}$$

$$y(t) = g(Vh(t)) \tag{2}$$

where $h(t)$ represents the state of hidden layer while $f(x)$ is sigmoid function

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

and $g(x)$ is softmax function

$$g(x_k) = \frac{e^{x_k}}{\sum_i e^{x_i}} \tag{4}$$

Then the cross entropy is applied to generate error backpropagated to hidden layer for weight update. On the basis of this, to address the problem that information of history steps may not be maintained in the hidden layer, Mikolov et al. employed backpropagation through time (BPTT) in their subsequent work [47], allowing the network to learn to retain information for multiple time steps. In addition, under the assumption that words can be mapped onto different classes, by factoring the output layer, the probability of a predicted word given its history can be broken down into two components:

$$P(w_i|w_{i-1}, \ldots, w_1) = P(a_i|h(t))P(w_i|a_i, h(t)) \tag{5}$$

where the left side denotes the probability of the ith word conditioned on the history, the first term on the right side estimates a probability distribution over the classes and for the second a distribution over the words from a single class, a is the class which can be computed by

$$a(t) = g(Zh(t)). \tag{6}$$

This is illustrated in the RNN net structure as in Fig. 4(a), that is, the network still utilizes the entire hidden layer to calculate a potential full probability distribution across the entire vocabulary, while factorization allows it to evaluate only a portion of the output layer for training, thus significantly decreasing the computation load. In parallel, the RNNLM is enhanced in [48] by giving each word a related real-valued input vector, Fig. 4(a) illustrates the extra component of the network and computations of the values in (1) and (2) are updated as

$$h(t) = f(Ux(t) + Wh(t-1) + Ef(t)) \tag{7}$$

$$y(t) = g(Uh(t) + Cf(t)) \tag{8}$$

where $f(t)$ denotes the feature vector, which can be any form of information about the current word. This approach helps the model to understand the context and to more accurately predict the possibility of the next word.

Another recent development in this language model is the RNN encoder–decoder, which was introduced in [49]. This model

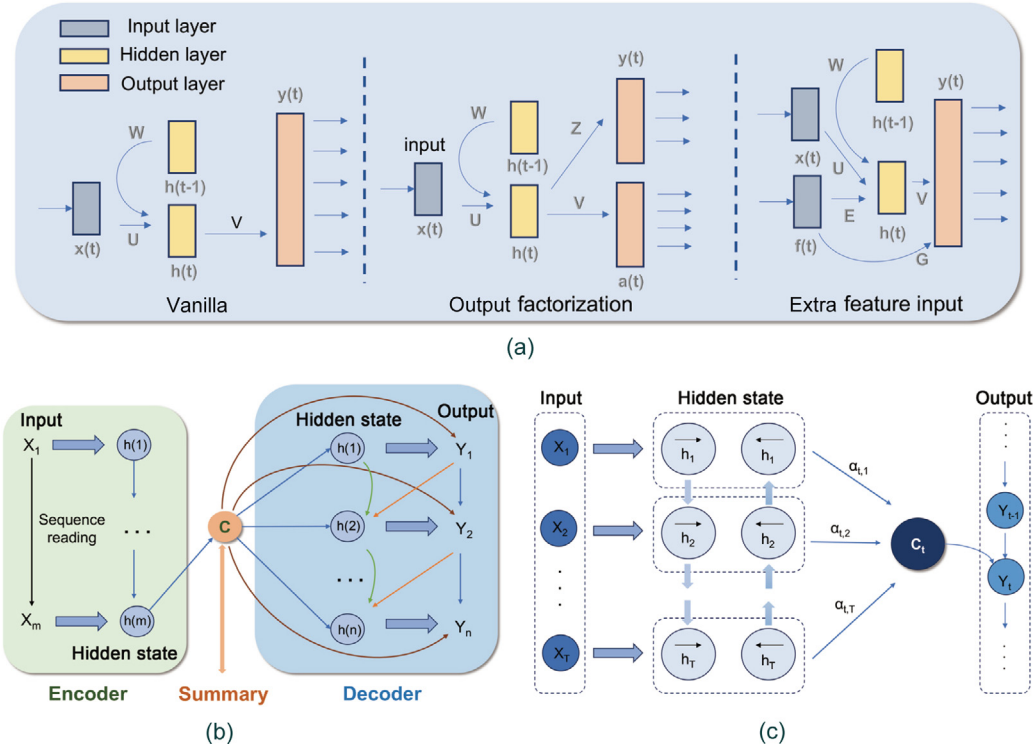

Fig. 4. (a) Three kinds of RNN models: Vanilla RNN, RNN with output factorization and RNN with extra feature input. (b) Encoder–Decoder structure. Two RNN in the model used for encoding and decoding respectively, c denotes the hidden state as a summary after traversing to the end of the sequence. (c) Bidirectional RNN used for capturing context dependencies in two directions. One traverses the input from the first to the end, and the other reverses backwards from the last token.

utilizes a Recurrent Neural Network (RNN) to transform a sequence of symbols into a fixed-length vector representation, and a second RNN to decode the vector into a different sequence of symbols. In Fig. 4(b), the completely updated hidden state of the RNN is a summary c of the entire input sequence.

Hence, the hidden state of the encoder at time t can be calculated as (1). Take the context into consideration, for the decoder:

$$h(t) = f(h(t-1), y(t-1), c) \tag{9}$$

then the goal is to maximize the conditional log-likelihood:

$$\max_{\alpha} \frac{1}{k} \sum_{i=1}^{k} logP_{\alpha}(y_i|x_i) \tag{10}$$

By comparison, this model is capable of effectively capturing linguistic patterns in phrase pairs, and the RNN Encoder-Decoder is capable of generating appropriate target phrases. Based on it, Bahdanau et al. [50] removed the restriction of fixed-length vectors, enabling the model to automatically identify parts of the source sentence that are relevant to predicting a target word, without having to explicitly define these parts as a rigid segment. In this approach, the summary c (also called context vector) depends on a sequence of hidden states of the encoder, which is computed as

$$C_n = \sum_{m=1}^{T} \alpha_{nm} h_m \tag{11}$$

where α is the weight of each state and calculated according to an alignment model, which can be trained with the translation model jointly.

Moreover, the encoder in this framework uses a bidirectional RNN, as illustrated in Fig. 4(c). This allows each word's annotation to take into account not only the words that come before it, but also those that come after. As a result, this approach has a major positive impact on the neural machine translation system's ability to produce satisfactory results for longer sentences, compared to the previous one.

3.1.2. Transformer based language models

In the ever-evolving landscape of natural language processing and machine learning, few innovations have made as profound an impact as the Transformer. Introduced in the groundbreaking work by Vaswani et al. in [51], the Transformer architecture revolutionized the way of approaching tasks like language translation, text generation, and even image synthesis. Its strength comes from an encoder–decoder framework shown in Fig. 5, which introduces an attention mechanism that can be computed as a score in the following form:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax}(\frac{QK^T}{\sqrt{d}})V \\ &= \text{softmax}(\frac{(W_Q X)(W_K X)^T}{\sqrt{d}})(W_V X) \end{aligned} \tag{12}$$

where X is the embedding matrix of input segments and W denotes the weights of each component at dimension d. And it also adapts parallel processing capabilities with multi-head attention mechanism computed as

$$\text{MultiHead}(Q, K, V) = \text{Concatenate}(h_1, h_2, \ldots, h_i)W^H \tag{13}$$

where h_i = Attention(Q_i, K_i, V_i). These heads with each of its learned projection matrices help enhance the model's ability to focus on different aspects of the data and learn complex relationships.

The vanilla Transformer has achieved remarkable results in a variety of applications. However, its ability to learn longer-term dependencies is limited by the fixed-length context of language modeling. This can lead to problems that

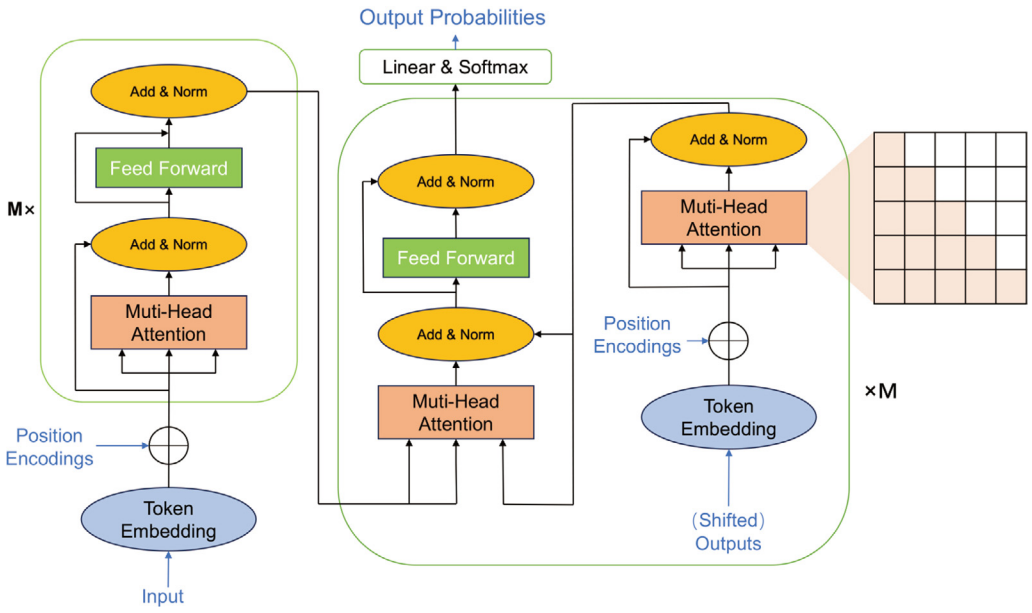

Fig. 5. The architecture of transformer. The framework consists of an encoder and a decoder, both of which have several multi-head attention mechanisms.

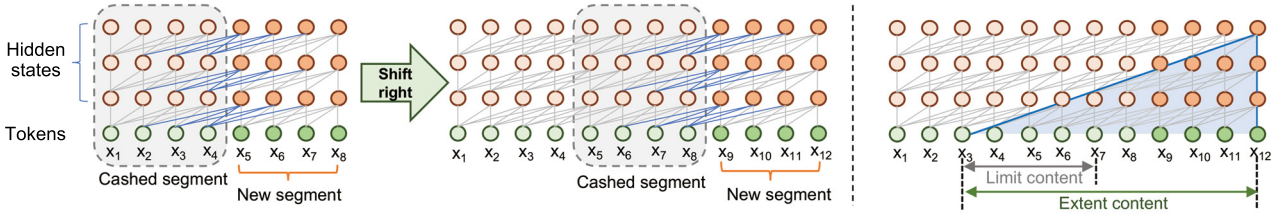

Fig. 6. Sharing of hidden states between segments in Transformer-XL. In the left-side training phase, there is no gradient in the previous segment and its hidden states are cached for connection with the new segment. In the left-side evaluation phase, by sharing hidden states the context is extended to better understand a long sequence of tokens.

(1) The model cannot deal with beginning tokens when there is few context.
(2) During the evaluation phase, new segments must be completely reprocessed when it is shifted one bit to the right, which requires a significant amount of computation.
(3) Consequently, it is unable to obtain dependencies for sentences that are longer than a fixed context length.

In order to tackle the issues, Dai et al. [52] presented *Transformer-XL*, which allows learning dependencies to go beyond a fixed length without compromising temporal consistency. It uses an RNN-like mechanism to recur between segments to gain an extended context, that is, to reuse the cached hidden states between segments, as shown in Fig. 6.

Based on this, Adaptive Attention Span is proposed in [53] to automatically learn the optimal attention span, which is achieved by introducing a soft mask function that project a distance into [0,1] by

$$m_t(x) = \text{clamp}(\frac{1}{\alpha}(\alpha + t - x), 0, 1) \tag{14}$$

where α is a hyperparameter and t is to be trained. By applying it to the *softmax* part of (12), the attention weight becomes

$$a_{pq} = \frac{m_t(p-q)exp(S_{pq})}{\sum_{k=p-S}^{p-1} m_t(p-k)exp(S_{pk})} \tag{15}$$

By doing this, the optimization of spans can be done independently for each head, which helps to reduce the amount of computing and memory resources needed and allows for a longer maximum context length for the model. In addition to context extension, subsequent studies have achieved considerable progress on the vanilla model to enhance its efficiency, Al-Rfou et al. [54] proposed the use of Auxiliary Losses to train a deep Transformer model in character-level language modeling, and the results demonstrate that the model is more effective than Long Short-Term Memory (LSTM). Other variants include Sparse Attention Matrix Factorization (SPMF) to train hundreds of layers of dense attention network [55], Reformer proposed by [56] reduces time complexity to a large extent, and even a Universal Transformer [57] combining self-attention with recurrence mechanisms in RNNs, which allows benefiting from both the Transformer's long-term global receptive field and the RNN's learning inductive bias.

The model's performance has been improved, which has enabled Transformer to be used in a variety of different disciplines. Starting from the initial purpose of text translation, it is now capable of generating images by associations between pixels by the virtue of Image Transformer [58], even Gated Transformer-XL [59] has broken the barrier of gradient explosion that RNNs were suffering from, and it has been successfully used in Reinforcement Learning, leading to a well-trained agent. These breakthroughs broadened the possibilities for its application in a wide variety of fields, and in the rapidly growing realm of human–robot interaction, research has also demonstrated that Transformer has a role to play.

Table 4
Recent language models and variants.

Year	Model	Parameters	Features
2018	BERT (base/large) [60]	110 M/340 M	Bidirectional pretraining for language representations
2019	ALBERT (base/large) [61]	12 M/18 M	Factorized embedding parameterization and cross-layer parameter sharing
2019	DistilBERT [62]	66 M	Size reduction by knowledge distillation
2019	RoBERTa (large) [63]	355 M	Optimizations with design choices and training strategies
2020	SpanBERT (large) [64]	340 M	Masked contiguous random spans and prediction by span boundary representations
2018	GPT [65]	1.17 B	Generative pretraining and discriminative fine-tuning
2019	GPT-2 [66]	15 B	Zero-shot unsupervised Multi-task learning
2020	GPT-3 [67]	175B	Few-shot learning with vast model size
2022	InstructGPT [68]	1.3 B	RL with reward feedback by manual rankings
2019	BART [69]	10% more than BERT[a]	Bidirectional encoder and left-to-right decoder

The number of parameters may vary from the above depending on the downstream tasks.
[a] Without an exact digit in the original text, here is a comparison with the BERT model of the same scale.

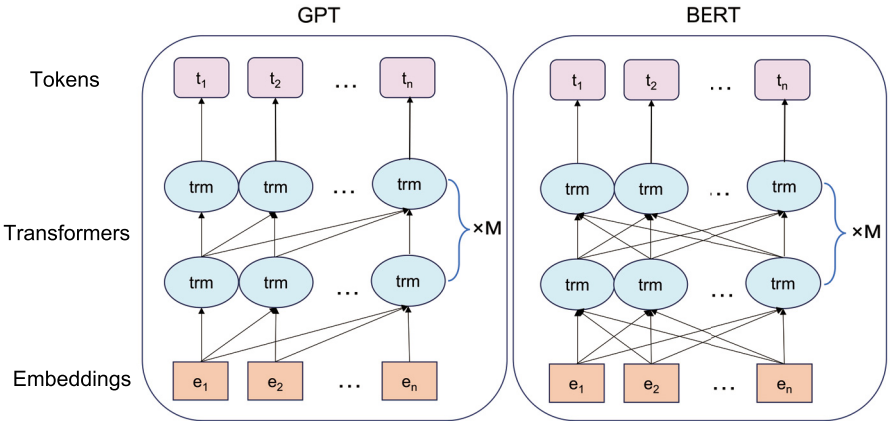

Fig. 7. A comparison in structures of GPT and BERT, while GPT utilizes a left-to-right transformer, BERT employs a bidirectional Transformer in its architecture.

3.2. Classification by training strategies

In addition to classifying language models according to their network structures, Transformer-based versions are widely used in many areas to meet users' requirements. This section provides a brief overview of some of the most popular models that have emerged in recent times, which can be traced back to the following two basic versions, and the different sizes and applications of these variants are summarized in Table 4.

Beginning with **GPT** (Generative Pre-training), Openai presented it as an autoregressive language model [65], and it has demonstrated remarkable potential in a variety of tasks such as text generation and creative writing. Through self-attention mechanisms in multiple stacked transformers, GPT-1 is able to recognize long-term associations and to be processed in parallel efficiently when dealing with natural languages. In addition, this approach utilizes a two-phase training procedure, beginning with pre-training the model on a large collection of unlabeled text and then refining it to meet the needs of specific downstream tasks. Based on this, GPT-2 [66] took a major step towards a few-shot and even a zero-shot transfer through unsupervised learning, it is also scaled to a large model size and can be trained on a gigantic corpus, leading to remarkable progress in multiple NLP tasks. And this point is comprehensively illustrated in GPT-3 proposed by Brown et al. [67], which adapts a model one hundred times the size of GPT-2 to generate realistic, coherent, and creative text, proving the scaling law that larger is better [70]. So far, the GPT model has demonstrated its superiority over traditional natural language processing models, showcasing the potential and promise of unsupervised learning and transfer learning in the realm of natural language processing.

BERT (Bidirectional Encoder Representations from Transformers), as another transformer-based model from the same era, has also demonstrated striking results in tasks such as text classification, sentiment analysis, and question answering using a distinct training approach. For unidirectional predictions, the full meaning of the sentence cannot be grasped. To address this, BERT uses a bidirectional prediction approach [60], combining the bidirectional predictions from the beginning to the end and vice versa to form the final result. The exceptional performance of BERT is also due to two novel tasks used in pre-training stage, namely Masked Language Model (MLM) and Next Sentence Prediction (NSP), such training and subsequent fine-tuning enable BERT to even outperform humans in multiple language processing tasks. Further studies have shown that increasing the size of models can enhance their performance on downstream tasks, however, this also puts a strain on hardware and necessitates a long training period. To release the limit, ALBERT [61] reduces the number of parameters to increase the training speed by factorized embedding parameterization and cross-layer parameter sharing without performance degradation. Similarly, Liu et al. [63] proposed RoBERTa, which optimizes certain design decisions during the pretraining process, demonstrating that the base model can be significantly improved without making major changes to its structure. While there are other strategies to minimize the size of the model and speed up training, such as DistilBERT [62] and TinyBERT [71], a major advancement is SpanBERT [64], which masks adjacent random spans instead of tokens, thus allowing predictions of entire masked spans without having to rely on individual token predictions. Moreover, many new variants of BERT have recently been presented, demonstrating that there is still a great deal of potential for its use in a wider range of applications.

In addition to the well-known models mentioned above and their differences in Fig. 7, there are other studies that push the boundary of language models to interdisciplinary fields, like

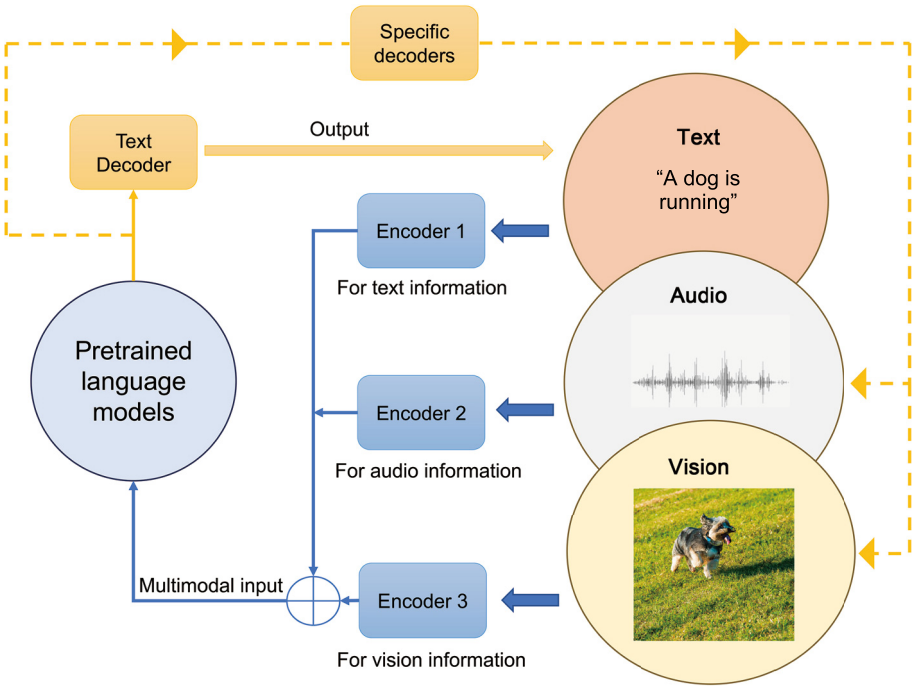

Fig. 8. Applications in inquiry answering. Multi-modal inquiries are first input into the pre-trained model through their specific encoders, and the output of the model is passed through the decoder to generate text, or if necessary, special decoders are used to generate information in various forms.

the new version GPT4 [72] with unprecedented performance in interpretation, InstructGPT [68] that introduces reinforcement learning from human feedback and BART [69] combining the merits of GPT and BERT, thus proposing a universe model to be applied to two different domains of natural language tasks. On the other hand, robotic learning, a popular area of research in the field of human–robot interaction, has long been impeded by the difficulty of translating vague instructions from humans into low-level control signals. Consequently, the development of large language models offers the potential for robots to be designed with increased intelligence.

4. Applications of LLMs in HRI

The remarkable interpretive capabilities of large language models have enabled them to be applied in a variety of disciplines, from everyday housework [3] to industrial production [73,74]. In this section, the applications listed are classified into three categories to provide a comprehensive overview of the potential of LLMs in different HRI cases.

4.1. Inquiry answering by multi-modal generation

Since the advent of LLMs, their ability of mapping inquiries to responses has been widely applied in search engines, providing users with access to knowledge bases. As a type of robots are designed to respond to user inquiries, an example of this is machine translation, which is a form of Question & Answer, Brants et al. [75] studied the advantages of large-scale statistical language modeling trained on trillions of tokens and the results compete with those of intricate methods when the language model's size is increased. And Hermann et al. developed deep neural networks combined with attention mechanisms [76], enabling them to read documents and answer intricate questions without requiring extensive prior knowledge of the structure of the language.

Although this text-text form can be also applied to lengthy story generation [77], challenges remain that the nature of human languages makes prompts ambiguous and lack of details, and it is easier for standard language model pretraining approaches to acquire certain kinds of factual information than others [78], thus this would lead to bias from the prompt and result in unwanted or even hazard content. To solve this, Keskar et al. proposed Ctrl [79], a transformer-based model with control codes derived from the structure that coexist with raw text to give more explicit control over text generation; and pretrained models combined with attribute classifiers for a backward pass are applied in PPLM [80], achieving controllable text generation. Additionally, interactive query rewriting [81] is a useful approach to reduce ambiguity, transforming original queries into readable and specified natural languages. In terms of optimization, KenLm [82] offers a data structure design that is effective for language model queries, resulting in both time and memory savings.

On the basis of that, later researches have explored more possibilities. Not confined to text, [83,84] pretrain models on large image datasets with captions from the Web, eliminating the need for a large amount of manual labeling. An extension VideoCLIP is proposed by Xu et al. [85], which connects videos and text to enable zero-shot transfer learning, and Audioclip [86] and Imagebind [87] further improved this by introducing audio models into the CLIP framework, thus allowing for the association of text, images and audio. As shown in Fig. 8, these developments alter the way of data exchange from one form to various modalities, thus Human–Robot Interaction can be enhanced by having a clearer understanding of the other's purpose.

Although LLM-based frameworks have shown the ability to deal with multimodal generation, there are emerging latest studies that push the boundaries even further. As extensions to text generation, Codex [88] and Palm [89] compete in code writing, which have demonstrated impressive fluency and effectiveness in the field of coding. On certain benchmarks, they have outperformed human coders in terms of accuracy and cost. More visually, MotionGPT [90] which uses text prompts as input, can generate motions of avatars by training a motion tokenizer and a motion-aware language model. In addition to this, UDE [91] has an audio module which allows humans in simulation to create

rhythmic movements in response to music. And Yi et al. [92] demonstrated the ability to create complete human motions from a single piece of speech, and the result maintains great coherence between body language and speech content.

As modality and structure have been enhanced, language models are now able to generate responses in different forms to inquiries posed by users. And models such as ChatGPT have been shown to have the potential to be applied in other fields, such as medical education, where they demonstrate a high degree of agreement and understanding in their explanations [93]. At the same time, we should be aware of the potential danger posed by text that has been artificially created and looks genuine, as this kind of text can be taken as accurate without any person or organization being held accountable for it [94]. Therefore, only by rationally viewing their generated results can humans maintain the dominance in the interaction with language models.

4.2. Social robots with commonsense

The integration of robots into human society has been a popular topic of research. By using the most natural way, dialogue, social robots are capable of assisting people in different areas without causing discomfort. However, progress in this area has been hindered by the robots' lack of reasoning ability, which makes it difficult for them to gain commonsense, leading to potential hazard in some cases. The introduction of LLMs greatly alleviated this challenge, providing robots with pretrained models that can generalize and act as vast repositories of knowledge to answer inquiries that require skills in different domains.

One of the important applications areas of social robots is education, bringing robots into the class not only resolves the issue of overcrowding in schools with limited resources, but also provides the opportunity for more tailored curricula for students with varying requirements [95]. A Chatbot created by Crutzen et al. [96] based on a database is one of the first of this kind, offering adolescents advice and information on topics such as sex, drugs, and alcohol. Later Gordon et al. [97] proposed a learning technique for children to acquire second language skills by an autonomous robot companion, and the result shows a notable increase in valence due to personalized affective feedback. With similar target, a study is conducted in [98] that presented a framework utilizing Augmented Reality (AR), Voicebots, and ChatGPT for foreign language learning. By the impressive reasoning ability of large language models, chat robots are also able to generate coherent stories for children. The work by Nichols et al. [99] serves as a demonstration of a storytelling system, and it is capable of producing new phrases based on the narrative so far, guiding the plot in the direction desired by the user. At a more abstract level, LLMs can gain a profound understanding of content that is ambiguous, such as philosophy, which is not typically encountered in everyday conversation. Schwitzgebel et al. [100] creates a GPT-3-based chatbot for generating philosophical texts, which are comparable to the authentic works of philosophers and hard to discern by humans. Therefore, the introduction of language models strengthens the reasoning ability about commonsense, at deeper levels, they even have a potential to learn cognitive understanding beyond that of humans.

In addition to the field mentioned above, medical care is another area of research that has attracted a great deal of attention. Robots should be handled with extreme care in this field, as any oversight can have disastrous consequences, causing many studies to be discouraged. By virtue of deep learning techniques, research has recently been investigated in this direction. In a previous work, Li et al. [101] combined Bi-LSTM and attention mechanisms to develop a general Chinese chatbot for children with ASD (autism spectrum disorder), providing a novel HRI approach for people with special needs. Using a similar idea, research can be carried out to use the same concept on a physical Q-CHAT-NAO robot [102], where the robot would obtain data directly from the toddler instead of from an external source. Therapy bots have been developed to teach users how to replicate emotions and provide support to people with communication difficulties through various methods [105]. Studies have shown that ChatGPT has a high level of understanding when it comes to pathological diagnosis [93], to the point that it can easily pass a medical license exam. Since social robots based on language models can be introduced in many fields shown in Fig. 9, they are also worth exploring in other domains.

Besides the advances concluded in Table 5, the development of language models is stepping rapidly, such as recent LaMDA [106], which has demonstrated the ability to address the primary issues of LLM, including safety and factual grounding. This progress is likely to expand the scope of social robots to a wider range of applications. Research is being carried out to explore other areas, such as customer service systems in business [104] and university admissions [103]. Social robots are becoming increasingly common in public places, such as markets, where they can provide guidance and advice through conversation, and these interactions can even take place in multi-agent systems, with roles being played while still adhering to human intentions [107]. The immense capacity of social robots, which is derived from large language models, enables people engaging in Human–Robot Interaction to experience a dialogue as if they were conversing with a thinking, rational being.

4.3. Instruction following and task completion

While the development of large language models has enabled multimodal generation by agents in virtual settings, recent studies have further explored its potential in physical robot applications. This has opened up a variety of possibilities, from assisting people with daily routine tasks to automating assembly on the production line. With the powerful reasoning capability of LLMs, intelligent robots shown in Fig. 10 are able to generate high-level task and low-level motion plans in response to instructions, making it more convenient than traditional control methods, as it eliminates the need for extensive programming for non-expert users.

This field has been faced with a long-term difficulty: instructions provided by humans are usually vague and incomplete without thorough details. When a command is issued, such as "prepare breakfast", it is instinctive for humans to come up with a sequence of plans, such as first cleaning the table and then serving the dishes. However, for a robot, it needs to be able to perform commonsense reasoning in order to ground the natural language instructions into motion plans. To achieve this, Nyga et al. [108] adapted a probabilistic model with Markov Logic Network (MLN) to make use of prior knowledge to guess missing elements of the plan, which remains valid even when the workspace is partially known; this allows robots to look for the needed objects in possible locations that are invisible in the current state. This approach has a downside in that it necessitates a broad base of pre-existing knowledge that links objects to their affordances, and may not be able to generalize to new environments with objects outside the category since the expensive cost of data and the lack of efficient sampling impede the creation of flexible agents that can perform a variety of tasks and learn new ones quickly [109]. Therefore, many studies have used LLMs as the brains of robots with access to commonsense, allowing them to make decisions while robots act as physical arms and legs to carry out the plans generated by LLMs. Recent developments [5,110–114] have demonstrated the potential of using multimodal input, such as text and vision, to

Table 5
Application studies of social robots in different areas.

Ref.	Author and year	Aim	Techniques	Area
[96]	Crutzen et al. (2011)	A chat agent that answers physical and mental questions for adolescents	–	Education
[97]	Gordon et al. (2016)	A tutor robot for children learning second language	Intelligent Tutoring Systems (ITSs)	Education
[99]	Nichols et al. (2020)	Collaborative story generation in human's desire	GPT-2	Education
[100]	Schwitzgebel et al. (2023)	A chatbot capable of answering philosophical questions	GPT-3	Education
[98]	Topsakal and Topsakal	A framework helping little children learn foreign languages	ChatGPT and Augmented Reality (AR)	Education
[101]	Li et al. (2020)	A chat robot for children with ASD (autism spectrum disorder).	Bi-LSTM and attention mechanism	Medical care
[93]	Kung et al. (2023)	Evaluation of the performance of large language models on the United States Medical Licensing Exam (USMLE)	ChatGPT	Medical care
[102]	Romero-García et al. (2021)	An observation-based autism screening system powered by NAO robots	Machine learning models	Medical care
[103]	Day and Shaw (2021)	A customer service system applied in the university admissions	GPT-2	Business
[104]	Kushwaha and Kar (2020)	Chatbot for Business to Address Marketing and Selection of Products	Language models based on encoder and decoder	Business

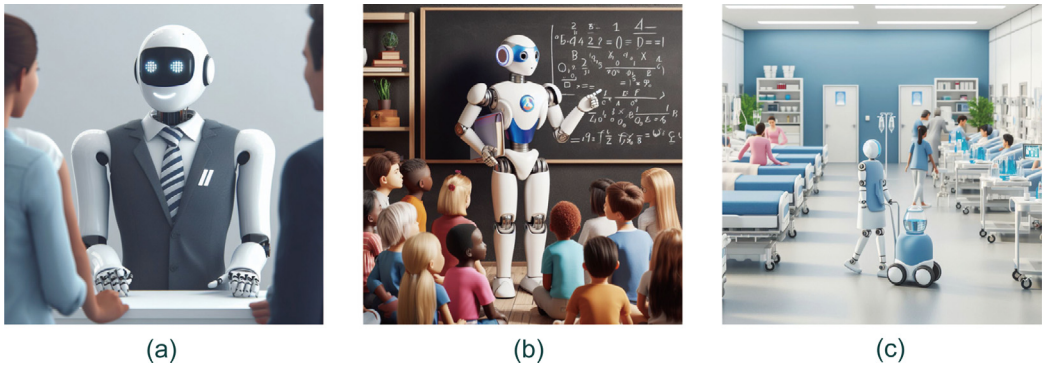

Fig. 9. Examples of social robots in (a) business reception (b) children education and (c) medical healthcare.

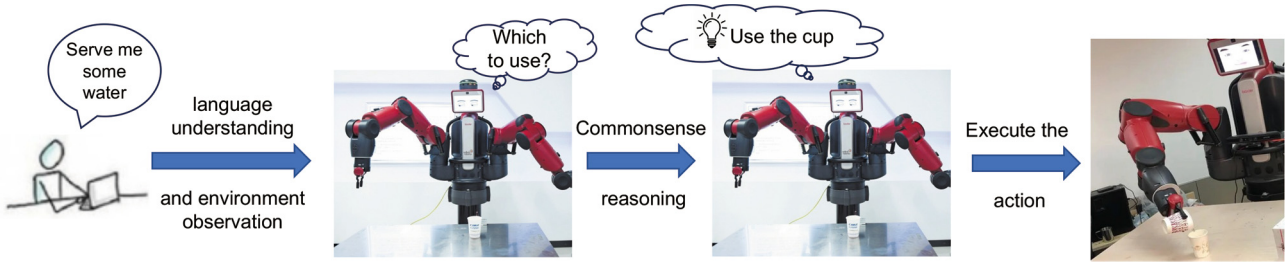

Fig. 10. Instruction following tasks by service robots. Given the command, large language models enable robot to reason with commonsense and look for possible tools in environment to accomplish the task.

reason about the environment and generate possible task plans, then by utilizing value functions [3] or specific metrics such as KAS [115], the most reasonable choice for the agent to perform can be determined.

Although a plan may appear to be the best option when it is created, there are unpredictable elements that cannot be foreseen. As the agent's movements can cause changes in the environment, a static plan may not be suitable for all situations and could result in potential dangers. In such cases, the plan has to be refined online, which can be achieved by state feedback from the environment. By applying success detection and passive or active scene description, robots can use inner monologue [4] to evaluate whether a plan is suitable in the current state and to improve the successful execution of high-level instructions in various domains. When the robot is not sure about the state information of the environment, interaction is considered as an effective approach to reduce uncertainty. In the work by Zhao et al. [116], the robot agent is able to interact with objects by weighing or knocking to determine their inner properties that are not available by vision; and KNOWNO [117] enables robots to request assistance from human users when faced with the unknown and utilizes conformal prediction to guarantee the validity of the plan for task completion.

Language model techniques offer a range of opportunities for downstream tasks in various fields, researchers can design and train their own models, such as PIGINET [118], to generate the desired representation and output. Meanwhile, it is possible to use pre-trained models without any fine-tuning to complete tasks, as demonstrated by ProgPrompt [119] and CaP [70], which directly generate code of control policies for robots by putting instructions and planning samples into GPT models. Apart from that, RoCo [120] shows that the interaction is not limited to humans and robots, but can also occur between agents. When participating in role-playing conversations, multiple agents can

collaborate to accomplish extraordinary tasks. These latest progressions have enabled robot control to develop towards intelligence, making it easier for human–robot interaction and for robots to be seamlessly incorporated into human society.

5. Challenges and discussion

The sections above give a comprehensive overview of the applications of large language models in human–robot interaction, demonstrating their unprecedented capabilities of reasoning and generation. However, beneath the surface of this promising technology there are many formidable challenges that must be addressed to ensure safe, effective, and ethical interactions. This paper looks into the issues that researchers, developers, and practitioners face when attempting to utilize large language models in the realm of human–robot interaction. We investigate the intricacies and factors that must be taken into account when blending human and artificial intelligence, and for each type of difficulty, we suggest potential solutions and provide possible paths for further research in this area.

- Safety and ethical concerns

This concern revolves around the potential for unintended consequences when deploying powerful AI systems in real-world scenarios. Large language models, while capable of generating human-like responses, can also generate harmful, biased, or inappropriate content, posing risks to users and society at large. When it comes to security, there is a risk that private information could be revealed or language models could accurately guess confidential or other sensitive data due to the need for vast amounts of data in their training [121]. Another consideration is whether these models adhere to strict ethical guidelines, including principles of fairness, transparency, and accountability. As social robots are already being used in educational settings [96–99], biased or harmful information in the content taught to students could have disastrous consequences.

Consequently, it is imperative to create reliable safeguards to stop malicious use and to manage data privacy issues cautiously, especially when dealing with delicate user data. Additionally, further progress is needed to guarantee that the output is controllable and is seen as acceptable to people. As the field of human–robot interaction continues to progress, addressing these safety and ethical issues is an ongoing and essential task to take advantage of AI while reducing potential harms.

- Contextual understanding

Whether large language models are capable of understanding context remains a pivotal challenge in the realm of LLMs employed in human–robot interaction. These models, while proficient in processing and generating text, often struggle to grasp the intricate nuances of conversation and context. Despite advances made in the development of multimodal communication frameworks, which offer more ways to comprehend context, there is still a need for more effective methods to manage complex interactions and behaviors [7]. Misinterpretations of user queries or responses that lack alignment with the ongoing conversation can lead to frustrating and ineffective interactions. This is particularly essential in the case of instruction following and task completion. If the intent of the human is not estimated correctly, the reckless actions taken by the robot could lead to dangerous outcomes and put lives and property at risk.

Hence, accurate contextual understanding necessitates not only the recognition of the immediate context, but also the maintenance of a broader conversation history, the tracking of user intent, and the perception of nuances such as tone, mood, and urgency. Meanwhile, the system designers need to appreciate such a deficiency in understanding the context to explore more possibilities of the problem [6]. As we seek to create more natural, intuitive, and responsive human–robot interactions, addressing these challenges in contextual understanding emerges as a fundamental task, necessitating advancements in model's reasoning and inference capabilities to ensure that robots can engage in meaningful and contextually relevant conversations with users.

- Scalability and generalization

Although large language models have demonstrated remarkable capabilities in controlled environments, their effective adaptation to diverse contexts, languages, and user groups remains a complex undertaking. The architecture of the model is a major limitation when it comes to human–computer interaction through language models. As mentioned, ChatGPT is currently limited to processing a maximum of 5000 text tokens as input [122], which could prevent it from being able to generalize to more complex human–robot interaction scenarios, where users' prompts are converted into commands to control agents. Ensuring that these models can seamlessly scale to support a large number of concurrent interactions or accommodate a wide array of domains and languages without compromising performance is no small feat. Achieving robust generalization, where models can effectively comprehend and respond to unforeseen scenarios, conversational styles, or languages not encountered during training, is a crucial but complex objective. As we aspire to harness the full potential of large language models in human–robot interaction across diverse and evolving contexts, addressing these challenges in scalability and generalization becomes imperative to ensure the practicality and versatility of AI-driven interactions.

Therefore, despite the recent advances in large language models that have made great strides in human–robot interactions, the question of "How big is too big?" [94] and the challenges posed by these models remain. The database used for training the model, the structure design, and the pre-trained model fine tuning all have areas that can be improved. It is essential to find solutions to these problems to successfully integrate robots into human society and make interactions simpler for users.

6. Conclusion

In this review, we have ventured into the rapidly evolving domain of human–robot interaction (HRI) augmented by the capabilities of large language models (LLMs). The confluence of natural language processing and robotics is reshaping the paradigms of interaction, introducing a new dimension of fluidity and intuition to human–robot collaborations. LLMs, with their vast knowledge repositories and generation prowess, offer unparalleled potential in transforming rudimentary robot tasks to complex, contextually-aware endeavors.

The paper underscored the feasibility of LLMs in grounding ambiguous human intent into concrete responses. Classifying these models on different grounds, recent advances in language models are theoretically introduced with their practical applications in multidisciplinary domains. From multimodal generation to social robots with conversation interface that serve people in various cases, they illustrate great potential to complete tasks meeting the requirement of users. With extraordinary reasoning and generation ability, large language models are hopefully to be applied in a broader range of areas in the subsequent future.

However, with great power comes great responsibility. The integration of LLMs in HRI is fraught with challenges, not limited to, but including, safety, ethical considerations, and the model's genuine comprehension of context. As robots steadily permeate sensitive domains like education, the ramifications of biased or

inappropriate outputs from LLMs can be profound. Furthermore, the nuances of conversation, laden with cultural, emotional, and personal idiosyncrasies, demand a higher echelon of contextual understanding that remains a formidable challenge for LLMs. The scalability and adaptability of these models to diverse situations, languages, and user groups further accentuate the complexity of the integration process.

In conclusion, while the fusion of LLMs and robotics heralds an exciting epoch in HRI, it also mandates rigorous scrutiny, research, and interdisciplinary collaboration. The promise of this synergy, while immense, must be approached with caution, ensuring that as we move forward, the trajectory remains anchored in ethical, safe, and user-centric principles. As the frontier of LLM-enhanced HRI expands, the onus is on the academic and industrial communities to harness its potential responsibly, paving the way for a future where robots seamlessly integrate into the tapestry of human society.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Abbreviations

LLM	Large language model
HRI	Human–robot interaction
EKF	Kalman filter
EMG	Electromyography
MRS	Mixed reality subspace
IHRCC	Improved human–robot collaborative control
MIS	Minimally invasive surgery
SA	Situation awareness
RNNLM	Recurrent neural network language model
BPTT	Backpropagation through time
GPT	Generative Pre-training
BERT	Bidirectional encoder representations from transformers
MLM	Masked Language Model
NSP	Next Sentence Prediction
AR	Augmented reality
ASD	Autism spectrum disorder
MLN	Markov Logic Network
SPMF	Sparse Attention Matrix Factorization
LSTM	Long Short-term Memory

References

[1] K. Dautenhahn, Socially intelligent robots: dimensions of human–robot interaction, Phil. Trans. R. Soc. B 362 (1480) (2007) 679–704.

[2] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. 103 (2023) 102274.

[3] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al., Do as i can, not as i say: Grounding language in robotic affordances, 2022, arXiv preprint arXiv:2204.01691.

[4] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al., Inner monologue: Embodied reasoning through planning with language models, 2022, arXiv preprint arXiv:2207.05608.

[5] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, et al., Socratic models: Composing zero-shot multimodal reasoning with language, 2022, arXiv preprint arXiv:2204.00598.

[6] T.B. Sheridan, Human–robot interaction: status and challenges, Hum. Factors 58 (4) (2016) 525–532.

[7] P. Tsarouchi, S. Makris, G. Chryssolouris, Human–robot interaction review and challenges on task planning and programming, Int. J. Comput. Integr. Manuf. 29 (8) (2016) 916–931.

[8] Y.S. Kim, J. Lee, S. Lee, M. Kim, A force reflected exoskeleton-type masterarm for human-robot interaction, IEEE Trans. Syst. Man Cybern. A 35 (2) (2005) 198–212.

[9] A.M. Okamura, Haptic dimensions of human-robot interaction, 2018.

[10] S. Scheggi, F. Chinello, D. Prattichizzo, Vibrotactile haptic feedback for human-robot interaction in leader-follower tasks, in: Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, 2012, pp. 1–4.

[11] A. Gutierrez, V.K. Guda, S. Mugisha, C. Chevallereau, D. Chablat, Trajectory planning in dynamics environment: application for haptic perception in safe human-robot interaction, in: International Conference on Human-Computer Interaction, Springer, 2022, pp. 313–328.

[12] Z. Wang, E. Giannopoulos, M. Slater, A. Peer, Handshake: Realistic human-robot interaction in haptic enhanced virtual reality, Presence 20 (4) (2011) 371–392.

[13] D.P. Losey, C.G. McDonald, E. Battaglia, M.K. O'Malley, A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction, Appl. Mech. Rev. 70 (1) (2018) 010804.

[14] S. Mugisha, V.K. Guda, C. Chevallereau, M. Zoppi, R. Molfino, D. Chablat, Improving haptic response for contextual human robot interaction, Sensors 22 (5) (2022) 2040.

[15] Y. Aydin, N. Arghavani, C. Basdogan, A new control architecture for physical human-robot interaction based on haptic communication, in: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, 2014, pp. 122–123.

[16] H. Wang, K. Kosuge, Control of a robot dancer for enhancing haptic human-robot interaction in waltz, IEEE Trans. Haptics 5 (3) (2012) 264–273.

[17] M. Bianchi, G. Valenza, A. Greco, M. Nardelli, E. Battaglia, A. Bicchi, E.P. Scilingo, Towards a novel generation of haptic and robotic interfaces: integrating affective physiology in human-robot interaction, in: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, 2016, pp. 125–131.

[18] Z.-b. Mao, Y. Asai, A. Wiranata, D.-q. Kong, J. Man, Eccentric actuator driven by stacked electrohydrodynamic pumps, J. Zhejiang Univ. Sci. A 23 (4) (2022) 329–334.

[19] S. Yohanan, K.E. MacLean, The haptic creature project: Social human-robot interaction through affective touch, in: Proceedings of the AISB 2008 Symposium on the Reign of Catz & Dogs: The Second AISB Symposium on the Role of Virtual Creatures in a Computerised Society, Vol. 1, Citeseer, 2008, pp. 7–11.

[20] S. Yohanan, K.E. MacLean, The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature, Int. J. Soc. Robot. 4 (2012) 163–180.

[21] S. Saunderson, G. Nejat, How robots influence humans: A survey of nonverbal communication in social human–robot interaction, Int. J. Soc. Robot. 11 (2019) 575–608.

[22] D. Tsetserukou, R. Tadakuma, H. Kajimoto, N. Kawakami, S. Tachi, Towards safe human-robot interaction: Joint impedance control of a new teleoperated robot arm, in: RO-MAN 2007-the 16th IEEE International Symposium on Robot and Human Interactive Communication, IEEE, 2007, pp. 860–865.

[23] L. Peppoloni, F. Brizzi, C.A. Avizzano, E. Ruffaldi, Immersive ROS-integrated framework for robot teleoperation, in: 2015 IEEE Symposium on 3D User Interfaces (3DUI), IEEE, 2015, pp. 177–178.

[24] R. Chellali, Tele-operation and human robots interactions, Remote Telerobot. 9 (2010).

[25] X. Zhao, X. Chen, Y. He, H. Cao, T. Chen, Varying speed rate controller for human–robot teleoperation based on muscle electrical signals, IEEE Access 7 (2019) 143563–143572.

[26] Y. Peng, Y. Sakai, K. Nakagawa, Y. Funabora, T. Aoyama, K. Yokoe, S. Doki, Funabot-suit: A bio-inspired and McKibben muscle-actuated suit for natural kinesthetic perception, Biomimetic Intell. Robot. (2023) 100127.

[27] M. Yu, Y. Lin, D. Schmidt, X. Wang, Y. Wang, Human-robot interaction based on gaze gestures for the drone teleoperation, J. Eye Mov. Res. 7 (4) (2014) 1–14.

[28] Y. Gatsoulis, G.S. Virk, A.A. Dehghani-Sanij, On the measurement of situation awareness for effective human-robot interaction in teleoperated systems, J. Cogn. Eng. Decis. Mak. 4 (1) (2010) 69–98.

[29] H. Hedayati, M. Walker, D. Szafir, Improving collocated robot teleoperation with augmented reality, in: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 2018, pp. 78–86.

[30] A. Dünser, M. Lochner, U. Engelke, D.R. Fernández, Visual and manual control for human-robot teleoperation, IEEE Comput. Graph. Appl. 35 (3) (2015) 22–32.

[31] Y.-P. Su, X.-Q. Chen, T. Zhou, C. Pretty, G. Chase, Mixed-reality-enhanced human–robot interaction with an imitation-based mapping approach for intuitive teleoperation of a robotic arm-hand system, Appl. Sci. 12 (9) (2022) 4740.
[32] H. Su, C. Yang, G. Ferrigno, E. De Momi, Improved human–robot collaborative control of redundant robot for teleoperated minimally invasive surgery, IEEE Robot. Autom. Lett. 4 (2) (2019) 1447–1453.
[33] C.L. Sidner, C. Lee, N. Lesh, The role of dialog in human robot interaction, in: International Workshop on Language Understanding and Agents for Real World Interaction, 2003.
[34] C. Shi, M. Shiomi, T. Kanda, H. Ishiguro, N. Hagita, Measuring communication participation to initiate conversation in human–robot interaction, Int. J. Soc. Robot. 7 (2015) 889–910.
[35] I. Papaioannou, C. Dondrup, O. Lemon, Human-robot interaction requires more than slot filling-multi-threaded dialogue for collaborative tasks and social conversation, in: FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction, 2018, pp. 61–64.
[36] J. Han, N. Campbell, K. Jokinen, G. Wilcock, Investigating the use of non-verbal cues in human-robot interaction with a nao robot, in: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), IEEE, 2012, pp. 679–683.
[37] F.I. Doğan, I. Torre, I. Leite, Asking follow-up clarifications to resolve ambiguities in human-robot conversation, in: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2022, pp. 461–469.
[38] C.L. Sidner, C. Lee, L.-P. Morency, C. Forlines, The effect of head-nod recognition in human-robot conversation, in: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, 2006, pp. 290–296.
[39] A. Vega, L.J. Manso, R. Cintas, P. Núñez, Planning human-robot interaction for social navigation in crowded environments, in: Advances in Physical Agents: Proceedings of the 19th International Workshop of Physical Agents (WAF 2018), November 22-23, 2018, Madrid, Spain, Springer, 2019, pp. 195–208.
[40] Y. Li, C.T. Ishi, K. Inoue, S. Nakamura, T. Kawahara, Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction, Adv. Robot. 33 (20) (2019) 1030–1041.
[41] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, N. Hagita, Footing in human-robot conversations: how robots might shape participant roles using gaze cues, in: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, 2009, pp. 61–68.
[42] C.L. Sidner, M. Dzikovska, Human-robot interaction: Engagement between humans and robots for hosting activities, in: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, IEEE, 2002, pp. 123–128.
[43] Z. Mao, Y. Peng, C. Hu, R. Ding, Y. Yamada, S. Maeda, Soft computing-based predictive modeling of flexible electrohydrodynamic pumps, Biomimetic Intell. Robot. 3 (3) (2023) 100114.
[44] Y. Peng, H. Yamaguchi, Y. Funabora, S. Doki, Modeling fabric-type actuator using point clouds by deep learning, IEEE Access 10 (2022) 94363–94375.
[45] M. Sundermeyer, I. Oparin, J.-L. Gauvain, B. Freiberg, R. Schlüter, H. Ney, Comparison of feedforward and recurrent neural network language models, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8430–8434.
[46] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, S. Khudanpur, Recurrent neural network based language model, in: Interspeech, Vol. 2, Makuhari, 2010, pp. 1045–1048.
[47] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, S. Khudanpur, Extensions of recurrent neural network language model, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011, pp. 5528–5531.
[48] T. Mikolov, G. Zweig, Context dependent recurrent neural network language model, in: 2012 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2012, pp. 234–239.
[49] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.
[50] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.
[51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[52] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le, R. Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, 2019, arXiv preprint arXiv:1901.02860.
[53] S. Sukhbaatar, E. Grave, P. Bojanowski, A. Joulin, Adaptive attention span in transformers, 2019, arXiv preprint arXiv:1905.07799.
[54] R. Al-Rfou, D. Choe, N. Constant, M. Guo, L. Jones, Character-level language modeling with deeper self-attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3159–3166.
[55] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, 2019, arXiv preprint arXiv:1904.10509.
[56] N. Kitaev, Ł. Kaiser, A. Levskaya, Reformer: The efficient transformer, 2020, arXiv preprint arXiv:2001.04451.
[57] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, Ł. Kaiser, Universal transformers, 2018, arXiv preprint arXiv:1807.03819.
[58] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, in: International Conference on Machine Learning, PMLR, 2018, pp. 4055–4064.
[59] E. Parisotto, F. Song, J. Rae, R. Pascanu, C. Gulcehre, S. Jayakumar, M. Jaderberg, R.L. Kaufman, A. Clark, S. Noury, et al., Stabilizing transformers for reinforcement learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 7487–7498.
[60] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[61] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv preprint arXiv:1909.11942.
[62] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019, arXiv preprint arXiv:1910.01108.
[63] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[64] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguist. 8 (2020) 64–77.
[65] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.
[66] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[67] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[68] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.
[69] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019, arXiv preprint arXiv:1910.13461.
[70] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, A. Zeng, Code as policies: Language model programs for embodied control, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 9493–9500.
[71] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, Q. Liu, Tinybert: Distilling bert for natural language understanding, 2019, arXiv preprint arXiv:1909.10351.
[72] OpenAI, GPT-4 technical report, 2023, arXiv:2303.08774.
[73] Z. Mao, Y. Asai, A. Yamanoi, Y. Seki, A. Wiranata, A. Minaminosono, Fluidic rolling robot using voltage-driven oscillating liquid, Smart Mater. Struct. 31 (10) (2022) 105006.
[74] Z. Mao, K. Yoshida, J.-w. Kim, Fast packaging by a partially-crosslinked SU-8 adhesive tape for microfluidic sensors and actuators, Sensors Actuators A 289 (2019) 77–86.
[75] T. Brants, A.C. Popat, P. Xu, F.J. Och, J. Dean, Large language models in machine translation, 2007.
[76] K.M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, Adv. Neural Inf. Process. Syst. 28 (2015).
[77] A. Fan, M. Lewis, Y. Dauphin, Hierarchical neural story generation, 2018, arXiv preprint arXiv:1805.04833.
[78] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A.H. Miller, S. Riedel, Language models as knowledge bases? 2019, arXiv preprint arXiv:1909.01066.
[79] N.S. Keskar, B. McCann, L.R. Varshney, C. Xiong, R. Socher, Ctrl: A conditional transformer language model for controllable generation, 2019, arXiv preprint arXiv:1909.05858.
[80] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, R. Liu, Plug and play language models: A simple approach to controlled text generation, 2019, arXiv preprint arXiv:1912.02164.
[81] A. Anand, A. Anand, V. Setty, et al., Query understanding in the age of large language models, 2023, arXiv preprint arXiv:2306.16004.

[82] K. Heafield, Kenlm: Faster and smaller language model queries, in: Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011, pp. 187–197.
[83] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[84] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.
[85] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, C. Feichtenhofer, Videoclip: Contrastive pre-training for zero-shot video-text understanding, 2021, arXiv preprint arXiv:2109.14084.
[86] A. Guzhov, F. Raue, J. Hees, A. Dengel, Audioclip: Extending clip to image, text and audio, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 976–980.
[87] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K.V. Alwala, A. Joulin, I. Misra, Imagebind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190.
[88] M. Chen, J. Tworek, H. Jun, Q. Yuan, H.P.d.O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, 2021, arXiv preprint arXiv:2107.03374.
[89] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, et al., Palm: Scaling language modeling with pathways, 2022, arXiv preprint arXiv:2204.02311.
[90] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, T. Chen, MotionGPT: Human motion as a foreign language, 2023, arXiv preprint arXiv:2306.14795.
[91] Z. Zhou, B. Wang, Ude: A unified driving engine for human motion generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5632–5641.
[92] H. Yi, H. Liang, Y. Liu, Q. Cao, Y. Wen, T. Bolkart, D. Tao, M.J. Black, Generating holistic 3d human motion from speech, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 469–480.
[93] T.H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, PLoS Digit. Health 2 (2) (2023) e0000198.
[94] E.M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big??? in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623.
[95] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: A review, Sci. Robot. 3 (21) (2018) eaat5954.
[96] R. Crutzen, G.-J.Y. Peters, S.D. Portugal, E.M. Fisser, J.J. Grolleman, An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study, J. Adoles. Health 48 (5) (2011) 514–519.
[97] G. Gordon, S. Spaulding, J.K. Westlund, J.J. Lee, L. Plummer, M. Martinez, M. Das, C. Breazeal, Affective personalization of a social robot tutor for children's second language skills, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016.
[98] O. Topsakal, E. Topsakal, Framework for a foreign language teaching software for children utilizing AR, voicebots and ChatGPT (large language models), J. Cognit. Syst. 7 (2) (2022) 33–38.
[99] E. Nichols, L. Gao, R. Gomez, Collaborative storytelling with large-scale neural language models, in: Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games, 2020, pp. 1–10.
[100] E. Schwitzgebel, D. Schwitzgebel, A. Strasser, Creating a large language model of a philosopher, 2023, arXiv preprint arXiv:2302.01339.
[101] X. Li, H. Zhong, B. Zhang, J. Zhang, A general Chinese chatbot based on deep learning and its—application for children with ASD, Int. J. Mach. Learn. Comput. 10 (4) (2020) 519–526.
[102] R. Romero-García, R. Martínez-Tomás, P. Pozo, F. de la Paz, E. Sarriá, Q-CHAT-NAO: A robotic approach to autism screening in toddlers, J. Biomed. Inform. 118 (2021) 103797.
[103] M.-Y. Day, S.-R. Shaw, AI customer service system with pre-trained language and response ranking models for university admissions, in: 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), IEEE, 2021, pp. 395–401.
[104] A.K. Kushwaha, A.K. Kar, Language model-driven chatbot for business to address marketing and selection of products, in: Re-Imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, Proceedings, Part I, Springer, 2020, pp. 16–28.
[105] K.T. Pham, A. Nabizadeh, S. Selek, Artificial intelligence and chatbots in psychiatry, Psychiatric Q. 93 (1) (2022) 249–253.
[106] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al., Lamda: Language models for dialog applications, 2022, arXiv preprint arXiv:2201.08239.
[107] G. Li, H.A.A.K. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, Camel: Communicative agents for" mind" exploration of large scale language model society, 2023, arXiv preprint arXiv:2303.17760.
[108] D. Nyga, S. Roy, R. Paul, D. Park, M. Pomarlan, M. Beetz, N. Roy, Grounding robot plans from natural language instructions with incomplete world knowledge, in: Conference on Robot Learning, PMLR, 2018, pp. 714–723.
[109] C.H. Song, J. Wu, C. Washington, B.M. Sadler, W.-L. Chao, Y. Su, Llm-planner: Few-shot grounded planning for embodied agents with large language models, 2022, arXiv preprint arXiv:2212.04088.
[110] Y. Mu, Q. Zhang, M. Hu, W. Wang, M. Ding, J. Jin, B. Wang, J. Dai, Y. Qiao, P. Luo, Embodiedgpt: Vision-language pre-training via embodied chain of thought, 2023, arXiv preprint arXiv:2305.15021.
[111] C. Huang, O. Mees, A. Zeng, W. Burgard, Visual language maps for robot navigation, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 10608–10615.
[112] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al., Rt-1: Robotics transformer for real-world control at scale, 2022, arXiv preprint arXiv:2212.06817.
[113] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al., Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023, arXiv preprint arXiv:2307.15818.
[114] D. Driess, F. Xia, M.S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., Palm-e: An embodied multimodal language model, 2023, arXiv preprint arXiv:2303.03378.
[115] B.Y. Lin, C. Huang, Q. Liu, W. Gu, S. Sommerer, X. Ren, On grounded planning for embodied tasks with language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 13192–13200.
[116] X. Zhao, M. Li, C. Weber, M.B. Hafez, S. Wermter, Chat with the environment: Interactive multimodal perception using large language models, 2023, arXiv preprint arXiv:2303.08268.
[117] A.Z. Ren, A. Dixit, A. Bodrova, S. Singh, S. Tu, N. Brown, P. Xu, L. Takayama, F. Xia, J. Varley, et al., Robots that ask for help: Uncertainty alignment for large language model planners, 2023, arXiv preprint arXiv:2307.01928.
[118] Z. Yang, C.R. Garrett, D. Fox, Sequence-based plan feasibility prediction for efficient task and motion planning, 2022, arXiv preprint arXiv:2211.01576.
[119] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, A. Garg, Progprompt: Generating situated robot task plans using large language models, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 11523–11530.
[120] Z. Mandi, S. Jain, S. Song, RoCo: Dialectic multi-robot collaboration with large language models, 2023, arXiv preprint arXiv:2307.04738.
[121] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models (2021), 2021, arXiv preprint arXiv:2112.04359.
[122] S. Shahriar, K. Hayawi, Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations, 2023, arXiv preprint arXiv:2302.13817.