Large Language Models for Robotics: A Survey

Wensheng Gan
Jinan University
283 PUBLICATIONS **5,604** CITATIONS

Large Language Models for Robotics: A Survey

Fanlong Zeng[a], Wensheng Gan[a,*], Yongheng Wang[a], Ning Liu[a] and Philip S. Yu[b]

[a] School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai 519070, China
[b] Department of Computer Science, University of Illinois Chicago, Chicago, USA

ARTICLE INFO

Keywords:
large language models
robotics
control and interaction
decision-making
embodied intelligence

ABSTRACT

The human ability to learn, generalize, and control complex manipulation tasks through multi-modality feedback suggests a unique capability, which we refer to as dexterity intelligence. Understanding and assessing this intelligence is a complex task. Amidst the swift progress and extensive proliferation of large language models (LLMs), their applications in the field of robotics have garnered increasing attention. LLMs possess the ability to process and generate natural language, facilitating efficient interaction and collaboration with robots. Researchers and engineers in the field of robotics have recognized the immense potential of LLMs in enhancing robot intelligence, human-robot interaction, and autonomy. Therefore, this comprehensive review aims to summarize the applications of LLMs in robotics, delving into their impact and contributions to key areas such as robot control, perception, decision-making, and path planning. We first provide an overview of the background and development of LLMs for robotics, followed by a description of the benefits of LLMs for robotics and recent advancements in robotics models based on LLMs. We then delve into the various techniques used in the model, including those employed in perception, decision-making, control, and interaction. Finally, we explore the applications of LLMs in robotics and some potential challenges they may face in the near future. Embodied intelligence is the future of intelligent science, and LLMs-based robotics is one of the promising but challenging paths to achieve this.

1. Introduction

Humans possess exceptional proficiency in executing intricate and dexterous manipulation skills by integrating tactile, visual, and other sensory inputs. Research in the field of robotics aspires to imbue robots with comparable manipulation intelligence. Although recent advancements in robotics and machine learning have yielded promising results in visual mitigation and exploration learning for robot manipulation, there remains much to be accomplished in this area. Large language models (LLMs), such as BERT [31], Roberta [79], GPT-3 [27], GPT-4 [110], have emerged as significant research achievements in the field of artificial intelligence (AI) in recent years. Through deep learning techniques [76], LLMs can be trained on massive text corpora, enabling them to generate high-quality natural language text. This development has sparked new thinking in natural language processing and dialogue systems. At the same time, the rapid advancement of robotics technology [66, 32] has created a demand for more intelligent and natural human-machine interaction. Combining LLMs with robots can provide robots with stronger natural language understanding and generation capabilities, enabling more intelligent and human-like conversations and interactions.

Applying LLMs to the field of robotics has important research significance and practical value. Firstly, LLMs can significantly enhance a robot's natural language understanding and generation capabilities. Traditional robot dialogue systems often require manual rules and template writing, making it difficult to handle complex natural language inputs. LLMs, on the other hand, can better understand and generate natural language by learning from massive text corpora, enabling robots to have more intelligent and natural conversation abilities. Secondly, LLMs can provide more diverse conversation content and personalized interaction experiences. Through interaction with LLMs, robots can generate varied responses and personalize interactions based on user preferences and needs. This helps improve user satisfaction and interactions. In addition, the combination of LLMs and robots contributes to the advancement of artificial intelligence and robotics technology, laying the foundation for future intelligent robots (or called smart robots).

Currently, many research teams and companies have begun exploring the application of LLMs in the field of robotics. Some research focuses on using LLMs for natural language understanding in robots. By using pre-trained language models [152], robots can better understand user intentions and needs [34, 117]. Other research focuses on using LLMs for natural language generation in robots. Robots can generate fluent and coherent natural language responses through interaction with language models. Furthermore, some research explores how to combine LLMs with other technologies, such as knowledge graphs and sentiment analysis, to further enhance robot dialogue capabilities and user experiences. From multiple perspectives, LLMs-based robotics is one of the most promising paths to achieve embodied intelligence in the future.

Although the combination of LLMs and robots has many potential advantages, it also faces challenges and issues [50, 86]. Firstly, training and deploying LLMs require substantial computing resources and data, which can be challenging for resource-limited robot platforms [7]. Secondly, LLMs may generate inaccurate, unreasonable, or even harmful content

*Corresponding author
flzeng1@gmail.com (F. Zeng); wsgan001@gmail.com (W. Gan); yonghengwwang@gmail.com (Y. Wang); tliuning@jnu.edu.cn (N. Liu); psyu@uic.edu (P.S. Yu)
ORCID(s):

when generating natural language text. Effective filtering and control mechanisms are necessary to ensure that the content generated by robots complies with ethical and legal requirements [86]. Additionally, robot dialogue systems need to address challenges such as multi-turn dialogues, context understanding, and dialogue consistency to provide more coherent and human-like interactions. Furthermore, the shape of robots has not been standardized across the industry. The question remains whether robots should adopt a humanoid form or take on a different shape [57]. In other words, what form of robot is best suited for our needs? The impact of embodied intelligence on our society cannot be overstated. Will robots eventually replace human labor? How should we respond to this seismic shift in the future? Moreover, if robots were to gain consciousness, should we still view them as tools? How should humans define a conscious robot?

In conclusion, the applications of large language models in robotics hold tremendous potential. They provide new paradigms and methods for robot control, path planning, and intelligence. Through more intuitive and natural human-machine interaction, language-based path planning, and intelligent semantic understanding, large language models not only enhance the performance and efficiency of robots but also improve the experience and interaction modes of human-robot interaction.

Therefore, this comprehensive review aims to summarize the applications of LLMs in robotics, delving into their impact and contributions to key areas such as robot control, perception, decision-making, and path planning. To summarize, there are four key contributions in this paper, as follows:

- We discussed the latest advancements in LLMs and their significant impact on the field of robotics. We highlighted the benefits of LLMs for robots, as well as the emergence of new robot models equipped with LLMs in recent years.
- We discussed the current state of robot technology, focusing on advancements in perception, decision-making, control, and interaction combined with LLMs. Specifically, we highlighted the critical role of LLMs in decision-making modules, which have enabled robots to make more informed and effective decisions in various applications.
- We explored potential applications of current robots equipped with LLMs in the near future.
- We discussed several potential challenges that robots may face when integrated with LLMs, as well as the potential impact of future developments in this field on human society.

Organization: The rest of this article is organized as follows. In Section 2.1, we discuss related concepts of the LLMs and robotics. In Section 2.3 we introduce the new robot models equipped with LLMs in recent years. In Section 3, we indicate the practical guide for technical points. We also introduce the application in Section 4. Moreover, We highlight the challenges in Section 5 and present several promising directions of LLMs for robotics in Section 6. Finally, we conclude this paper in Section 7.

Table 1
Abbreviation with its corresponding description

Abbreviation	Description
AI	Artificial Intelligence
GPT	Generative Pre-trained Transformer
LLMs	Large-scale Language Models
VNM	Vision-Navigation Model
VLM	Vision-Language Model
VLN	Vision-and-Language Navigation model
VLA	Vision-Language-Action model

2. Robotics Based on LLMs

Amidst the swift progress and extensive proliferation of LLMs, the model of robotics based on LLMs has emerged. LLM serves as a robotics brain like in Figure 1, making it more intelligent. In this part, we first review the basic concept of LLMs and the popular LLMs nowadays. After that, we describe the benefits of robotics combined with LLM. Finally, we introduce the recent robotics model based on LLMs and the Transformer designed for robotics below. We also summarize the abbreviation used in this paper in Table 1 for convenience.

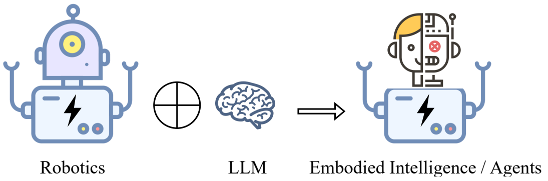


Figure 1: Robotics based on LLM.

2.1. Language Model Overview

2.1.1. Language Model Basics

We first provide an overview of LLM, starting with an introduction to some fundamental concepts. We then delve into the history of LLM's development, followed by a brief discussion of its growing popularity in recent years. A language model is a computational model that utilizes statistical methods to analyze and predict the probability of word sequences in a given language. It is designed to capture the patterns, grammar, and semantic meaning of natural language [92].

- **N-gram models** are a simple form of language models that calculate the probability of a word based on the preceding (n-1) words. They are widely used due to their simplicity and efficiency. The accuracy of the N-gram model is directly related to the length of the context used, with larger 'n' values leading to higher accuracy [13].

- **Unigram models** [147] are often employed for various language processing tasks including information retrieval. It evaluates each word or term independently. It is calculated without considering any conditional context, only the probability of the current word itself appearing.

- **Bidirectional models** differ from unidirectional models, it analyzes text in both directions: backward and forward. This dual approach is commonly employed in various machine learning models and speech generation applications. Bidirectional models harness the power of contextual information from both directions, providing a deeper understanding of the text [6].

- **Exponential models** [28] employ an equation that combines feature functions and n-grams to evaluate text. Unlike n-grams, this type of model allows for more flexibility in analyzing parameters and does not mandate the specification of individual gram sizes. Essentially, exponential models define features and parameters based on the desired outcomes, providing a more open-ended approach to text analysis.

- **Neural language models**, including recurrent neural networks (RNNs) [150] and transformers [131], have gained popularity in recent years. These models use deep learning techniques to capture complex language patterns and dependencies.

- **Transformer architecture**'s development revolutionized language modeling. Transformers use self-attention mechanisms to capture relationships between words in a sentence. This is currently the most popular architecture [131].

2.1.2. Development of LLMs

Some well-known developments in LLMs are described below in detail.

- **Eliza.** The concept of language generation models originated in the 1960s with the development of Eliza, the world's first chatbot, by MIT researcher Joseph Weizenbaum. Eliza's creation laid the groundwork for natural language processing (NLP) research, paving the way for subsequent advancements in this field [125].

- **LSTM.** The year 1997 witnessed the emergence of Long Short-Term Memory (LSTM) networks, introducing a significant advancement in neural network architecture. The introduction of LSTM networks enabled the development of deeper and more intricate neural networks capable of effectively processing vast amounts of data.

- **Stanford coreNLP.** In 2010, Stanford's CoreNLP suite brought about a significant milestone in the field by offering developers a versatile toolkit. This suite empowers developers to conduct various natural language processing tasks.

- **Google brain.** In 2011, a scaled-down version of Google Brain surfaced, introducing groundbreaking features such as word embeddings. These advanced capabilities revolutionized natural language processing (NLP) systems by enhancing their ability to comprehend context with greater clarity.

- **Transformer models.** Transformer models [131], introduced in 2017, brought significant advancements to language modeling. They employ self-attention mechanisms to capture global dependencies and have achieved state-of-the-art performance in various natural language processing tasks.

- **Large language model.** OpenAI unveiled GPT-4 [92], a language model boasting an astounding scale of approximately one trillion parameters. This represents a five-fold increase compared to its predecessor, GPT-3 [14], and a staggering 3,000-fold increase compared to the initial release of BERT [31]. The introduction of GPT-4 sets a new benchmark in the field of language models, showcasing the remarkable progress in model size and capacity.

2.1.3. Popular LLMs

Until now, there are many foundation models or LLMs have been developed. We present some selected models below, including GPT-3.5, GPT-4, BERT, T5, and LLaMA.

- **GPT-3 (Generative pre-trained transformer 3)** [14]. Developed by OpenAI, GPT-3 is one of the most prominent language models. With 175 billion parameters, it can generate coherent and contextually relevant text across a wide range of domains.

- **GPT-4 (Generative pre-trained transformer 4)** [92]. Unveiled on March 14, 2023, GPT-4 represents a significant advancement in language models, prioritizing factual accuracy and enhancing reliability compared to its predecessors, GPT-3 and GPT-3.5. Notably, GPT-4 introduces multimodal capabilities, enabling it to process images as input and generate comprehensive descriptions, classifications, and analyses across different modalities. This multimodal functionality expands the model's versatility and enhances its ability to understand and generate content across various media formats.

- **BERT (Bidirectional encoder representations from transformers)** [31]. Developed by Google, BERT introduced the concept of pre-training and fine-tuning for language understanding tasks. It has achieved remarkable results in tasks such as question answering and text classification.

- **T5 (Text-to-Text transfer transformer)** [101]. Developed by Google, T5 is a versatile language model that can be fine-tuned for various natural language processing tasks, including summarization, translation, and text generation.

- **LLaMA** [128]. Developed by Google, LLaMA is a language model pre-trained and fine-tuned generative text model with parameter counts ranging from 7 to 70 billion. LLaMA removes the absolute position embedding and instead adds rotational position embedding at each layer of the network.

2.2. Benefits of LLM for Robotics

The advent of LLM-based robots has brought about a plethora of innovative changes to the field. Here, we explore the various benefits that LLM will bring to robots. The necessity and significance of LLMs for robotics can be summarized in the following ten points:

- **Natural language interaction.** LLMs provide robots with the ability to engage in natural language interactions, allowing users to communicate with robots in an intuitive and convenient manner. This interaction method aligns better with human habits and needs, enhancing the usability and acceptance of robots.

- **Task execution.** LLMs assist robots in performing various tasks by understanding and generating natural language instructions. Robots can navigate, manipulate objects, and execute specific actions based on user language commands [126]. This opens up broader possibilities for robot applications in everyday life.

- **Knowledge acquisition and reasoning.** LLMs possess powerful information retrieval and reasoning capabilities, which can help robots acquire and process rich knowledge. Robots can interact with language models to obtain real-time and accurate information, thereby improving their decision-making ability and intelligence.

- **Flexibility and adaptability.** The flexibility of LLMs enables robots to adapt to different tasks and environments. Through interaction with language models, robots can make flexible adjustments and self-adaptation based on specific circumstances, better meeting user needs [52].

- **Learning and improvement.** LLMs enable continuous learning and improvement through interaction with users. By analyzing and understanding user feedback, robots can enhance their performance and proficiency. This learning and improvement capability allows robots to gradually adapt to user personalities and preferences, providing more personalized services.

- **Multimodal interaction.** LLMs also support multimodal interaction, enabling robots to process different forms of inputs such as speech, images, and text simultaneously. This multimodal capability [141] allows robots to comprehensively understand user needs and provide richer interaction experiences.

- **Education and entertainment.** LLMs offer potential applications for education and entertainment purposes in robotics. Robots can provide educational content, answer questions, or engage in games and entertainment activities through interaction with language models. This has significant implications for children's education, language learning, and the entertainment industry.

- **Emotional interaction.** The application of LLMs enhances the emotional interaction capabilities of robots. By generating emotionally responsive outputs, robots can establish closer and more meaningful relationships with users. This emotional interaction is valuable in fields such as care robots, emotional support, and psychotherapy.

- **Collaboration and cooperation.** LLMs enable robots to collaborate and cooperate better with humans. Robots can jointly solve problems, formulate plans, and execute tasks through interaction with language models [126]. This collaboration and cooperation ability is significant for industrial automation, team collaboration, and human-robot coexistence.

- **Innovation and exploration.** The application of LLMs stimulates innovation and exploration in the field of robotics. Through interaction with language models, robots can possess higher-level intelligence and comprehension abilities, opening up new avenues for research and development in robotics.

2.3. Robotics Based on LLMs

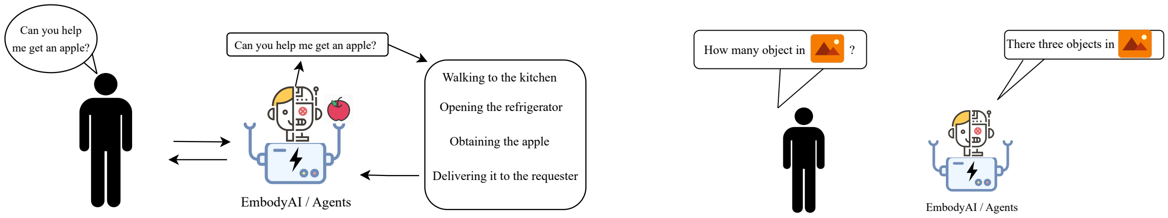
In this subsection, we introduce the smart robotics based on LLMs in recent years. LLMs are used as brains in the part of robotics. First, we summarize the models in recent years in Table 2.

2.3.1. PaLM-SayCan

With the increasing popularity of LLMs, people have begun to wonder whether these models can be used to assist robots in performing various daily tasks. However, there are challenges in enabling robots to extract knowledge from LLMs and interact with the physical world. LLMs contain valuable semantic information about the real world, aiding robots in understanding natural language. Nonetheless, giving LLMs a physical form capable of interacting and making real-world decisions is challenging due to their lack of experience with physical objects and environments. PaLM-SayCan [1] can function as the physical embodiment of LLM, utilizing LLM's semantic capabilities to process natural language instructions. PaLM-SayCan enables robots to execute tasks assigned by humans through the value function. PaLM-SayCan features pre-trained meta-actions controlled by visual motors, while BC-Z [58] and MT-Opt [64] are employed to learn language-conditioned BC and RL policies, respectively. LLM can decompose received natural language instructions into smaller, manageable tasks. Based on the current status, capabilities, and surrounding environment of the robot, actions can be flexibly executed. To determine the feasibility of an action, PaLM-SayCan relies on a logarithmic estimation of the value function and

Table 2
LLMs for robot in recent years

Year	LLM-based robotics	Description
2022	PaLM-SayCan [1]	PaLM-SayCan can function as the physical embodiment of LLM, utilizing LLM's semantic capabilities to process natural language instructions. Enabling robots to execute tasks assigned by humans through the value function.
2023	PaLM-E [34]	PaLM-E boasts an LLM capable of integrating continuous sensory information from the real world, effectively bridging the gap between language and perception.
2023	LM-Nav [117]	LM-Nav was developed, exploiting the advantages of language to facilitate effective communication between users and robots. The LM-Nav system comprises three components: a vision-navigation model (VNM); a vision language model (VLM); and a large language model (LLM).
2023	Expedition A1[1]	Expedition A1, developed by AGIBot, embodies the company's commitment to seamlessly integrating advanced AI into robotics and fostering harmonious collaboration between humans and machines.



(a) Task decomposition.

(b) Multi-modal sentences in PaLM-E.

Figure 2: Task decomposition and multi-modal sentences in PaLM-E.

affordance function. It will perform the most likely action to succeed in the current environment and state. For instance, upon receiving the instruction, “Can you help me get an apple?". LLM may decompose it into several tasks: “walking to the kitchen, opening the refrigerator, obtaining the apple, and delivering it to the requester.", just like in Figure 2(a).

2.3.2. PaLM-E

While LLMs have demonstrated remarkable capabilities in handling complex tasks, integrating them as an interface into robots remains a significant challenge. A major limitation of LLMs is their reliance on text input, which is insufficient for robots that require physical interaction. PaLM-E [34] boasts an LLM capable of integrating continuous sensory information from the real world, effectively bridging the gap between language and perception. Its multi-modal input encompasses vision, text, and state estimation, like in Figure 2(b), as exemplified by the question "What is it in <img_1>?" The model's processing is end-to-end, whose performance is state-of-the-art in OK-VQA [84]. PaLM-E is a visual-language generalist. PaLM-E treats images and text as multi-modal inputs represented by latent vectors. PaLM-E is a decoder-only model that generates text completions autonomously when provided with a prefix or hint. The output of PaLM-E is separated into two parts: when tackling text generation tasks (such as embedded question answering or scene description), the model directly produces the final output (i.e., output text or speech). In contrast, when utilized for specific planning and control tasks, PaLM-E generates low-level instruction text (e.g., instructions for controlling robot meta-actions).

2.3.3. LM-Nav

Goal-based robot navigation can leverage large, unlabeled datasets for training, resulting in strong generalization capabilities in real-world scenarios. However, in vision-based settings, specifying targets often requires images. Current supervised learning methods are not only expensive but also demand linguistically described and labeled trajectory data, making them impractical for widespread use. How can users communicate with robots more conveniently? To address the challenge, LM-Nav [117] was developed, exploiting the advantages of language to facilitate effective communication between users and robots. The LM-Nav system comprises three components: a vision-navigation model (VNM); a visual-language model (VLM); and a large-scale language model (LLM). Notably, LM-Nav operates without the requirement of labeled data or fine-tuning. By leveraging the VLM and VNM, LM-Nav can extract landmark names from commands and navigate to specified locations. LM-Nav leverages three pre-trained models to achieve successful navigation in pre-explored environments. First, it employs ViNG [114] as a VNM creates a topological map using observations from a prior exploration of the environment. Subsequently, GPT-3 [27] serves as the LLM [14] processes free-form text instructions to determine the target landmark. Finally, CLIP [99] serves as the VLM to locate the corresponding position in the topology map based on the identified landmark. By combining these models, LM-Nav can effectively follow natural language instructions to complete navigation tasks.

2.3.4. Expedition A1

Expedition A1[2], developed by AGIBot, embodies the company's commitment to seamlessly integrating advanced AI into robotics and fostering harmonious collaboration between humans and machines. Envisioning a future where robots serve as indispensable assistants to humans, AGIBot's mission is to create intelligent and versatile robots

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

Table 3
Transformer architecture in robotics

Year	Transformer architecture	Description
2022	Control Transformer [75]	Control Transformer (CT) utilizes a sample-based probabilistic road map (PRM) planner to generate conditional sequences from low-level policy, enabling it to complete navigation tasks solely through local information.
2022	Robotics Transformer 1 [10]	RT-1 is capable of encoding high-dimensional input and output data, including images and instructions, into compact tokens that can be efficiently processed by Transformer. It exhibits real-time operation characteristics, making it suitable for applications that require rapid processing and response times.
2023	Q-Transformer [18]	Q-Transfomer is proposed to combine the Transformer structure with offline reinforcement learning, enabling the exploitation of Q-values for each dimension.
2023	Robotics Transformer 2 [9]	Robot Transformer 2 (RT-2) is a model that leverages fine-tuning of a VLM. RT-2 training on a web-scale dataset to achieve direct possession of generalization ability and semantic awareness for new tasks.
2023	Robotics Transformer X [29]	Robotics Transformer X (RT-X) is categorized into two branches: RT-1-X and RT-2-X. RT-1-X employs the RT-1 architecture and utilizes the X-embodiment repository for training, while RT-2-X leverages the strategy architecture of RT-2 and is trained on the same dataset. Experiments demonstrate that both RT-1-X and RT-2-X have exhibited enhanced capabilities.

capable of unlocking limitless productivity. The company's founding ethos is centered around the belief that "intelligent robots can create unlimited productivity" when designed to parallel human flexibility and intelligence. Expedition A1 is a humanoid robot equipped with reflex knee joints, designed to resemble a human form. This design choice stems from the fact that most work environments are currently tailored for human functionality. Humanoid robots are allowed to seamlessly integrate and function without requiring significant environmental modifications. A key advantage of humanoid robots is their strong generalization capabilities, enabling them to adapt to diverse situations. While the Expedition A1 can also swap out components, such as replacing legs with tires, mimicking human movement and perception remains a significant challenge for robots. Expedition A1 integrates cutting-edge perception, control, and decision-making technologies, incorporating both a state-of-the-art language model and an independently developed visual model. Designed with industrial manufacturing in mind, it boasts 49 degrees of freedom, surpassing the limitations of traditional robots with only 20 degrees of freedom. Its high degree of freedom enables it to meet various industrial manufacturing requirements. The Expedition A1 is also modular, allowing for autonomous component replacement. For instance, *PowerFlow* is a joint motor for enhanced flexibility, while *SkillHand* features vision-based fingertip sensors for precision manufacturing scene design. In addition to its robust hardware, the Expedition A1 utilizes LLM as its brain, complemented by EI-Brain's embodied intelligence framework. This framework divides the robot's system into different levels of management, including Expedition A1's super brain in the cloud, local brain, cerebellum, and brainstem, each corresponding to diverse task levels.

2.4. New Transformer Architecture for Robotics

In this part, we introduce the Transformer designed for robotics. We summarize the Transformer for robotics in recent years in Table 3.

2.4.1. *Control Transformer*

Reinforcement learning [94] methods struggle to effectively tackle long-horizon tasks like navigation, but from a different angle, sample-based path planning techniques can discover collision-free paths without the need for learning in a known environment. Control Transformer (CT) [75] utilizes a sample-based probabilistic road map (PRM) [67] planner to generate conditional sequences from low-level policy, enabling it to complete navigation tasks solely through local information. CT has been shown to be effective in complex terrain and unknown environments through relevant experiments. By leveraging local observations, CT can solve long-horizon and robot navigation tasks. Following training, CT can obtain a policy and complete navigation from partially observed or unknown environments. CT is a Transformer [131] framework designed to model conditional sequences generated by robot actions. It utilizes a learnable value function to assess the initial cost of reaching the target position and guides the sequence modeling and generation process of the Transformer. To facilitate learning from data collections guided by sampling, the CT problem is treated as a sequence modeling problem with a goal-oriented approach. In essence, CT processing involves auto-regressively predicting actions within a sequence.

2.4.2. *Q-Transformer*

Many proposed high-capability machine learning models rely on supervised learning, but their performance is limited by the quality of human demonstrations. Neither the full potential of the hardware nor the required experience can be obtained automatically (given the availability of unlabeled datasets). Reinforcement learning [94] can address these limitations, but training Transformer-based models using reinforcement learning has proven challenging at large dataset sizes. To integrate reinforcement learning and Transformer [131], Q-Transfomer [18] is proposed. It combines the Transformer structure with offline reinforcement learning, enabling the exploitation of Q-values for each dimension. This is achieved by utilizing a Transformer-based architecture that leverages offline reinforcement learning to extend the representation of the Q-Function [74] through offline temporal differential backup [139]. The approach involves discretizing each action dimension and representing each action dimension as separate tokens using Q-values. This allows for the utilization of large and diverse robot datasets, enhancing the efficiency and effectiveness of the reinforcement learning process.

2.4.3. *Robotics Transformer*

Robotics transformer 1. By migrating large and diverse datasets, machine learning has now been targeted at downstream tasks and significantly improved performance in many areas (such as computer vision, natural language processing, or speech recognition) by fine-tuning with zero-shot or few-shot. However, the field of robotics has yet to show similar generalization capabilities. Training a general robotics model through open-ended task-agnostic training and incorporating high-performance architectures that can absorb large and diverse datasets may be a promising approach. If a model could act like a sponge, absorbing ubiquitous patterns of language and perception, it may be able to perform better on specific downstream tasks. The question remains whether it is possible to train a model in the field of robotics that can absorb knowledge from other fields. Could the model demonstrate zero-shot generalization capabilities for new tasks? Robotics Transformer 1 (RT-1) [10] was proposed to address the aforementioned question. RT-1 is capable of encoding high-dimensional input and output data, including images and instructions, into compact tokens that can be efficiently processed by Transformer [131]. It exhibits real-time operation characteristics, making it suitable for applications that require rapid processing and response times. In experimental evaluations, RT-1 demonstrated strong generalization. The structure of RT-1 is composed of FiLM [96], conditioned EfficientNet [124], a TokenLearner [107], and Transformer [131]. However, RT-1 is not an end-to-end model.

Robotics transformer 2. Can we pre-train a vision-language model (VLM) [22, 34] that can be seamlessly integrated into low-level robot control? Hereby enhancing VLM generalization capabilities? We can achieve this by training the robot's trajectory to be represented as a sequence of tokens, effectively mapping natural language instructions into a series of robot actions. To create an end-to-end model that can directly map robot observations into actions, DeepMind employs a collaborative fine-tuning approach. Combining state-of-the-art VLMs with network-scale visual-language tasks on robot trajectory data, Robot Transformer 2 (RT-2) [9] is a model that leverages fine-tuning of a VLM. RT-2 is trained on a web-scale dataset to achieve direct possession of generalization ability and semantic awareness for new tasks. Through fine-tuning a VLM, it is adapted to generate actions based on text encoding. Specifically, the model is trained on a dataset that incorporates action-related text tokens. This type of model can be called a visual-language-action model (VLA) [9]. RT-2 builds upon the policy trained by Robotic Transformer 1 (RT-1) [10], leveraging the same dataset and an expanded VLA to significantly enhance the model's generalization capabilities for new tasks.

Robotics transformer X. In robot learning, it is common to train a separate large model for each application or environment. However, this approach can be limiting, as it may not allow for adaptability across different robots or environments. Can we develop a robot policy that is versatile and can be applied across various robots and environments? With the advancements in large models, it is within the realm of possibility to train a versatile model that exhibits strong generalization capabilities for a specific task. Inspired by these large models, X-embodiment training[3] is proposed, which involves using robot data from diverse platforms for training. This approach enables the model to better adapt to changes in both the robot and the environment, leading to improved performance and versatility. Robotics Transformer X (RT-X) [29] is categorized into two branches: RT-1-X and RT-2-X. RT-1-X employs the RT-1 architecture and utilizes the X-embodiment repository[4] for training, while RT-2-X leverages the strategy architecture of RT-2 and is trained on the same dataset. Experiments demonstrate that both RT-1-X and RT-2-X have exhibited enhanced capabilities. Similarly, robots may benefit from acquiring knowledge across various domains, much like humans.

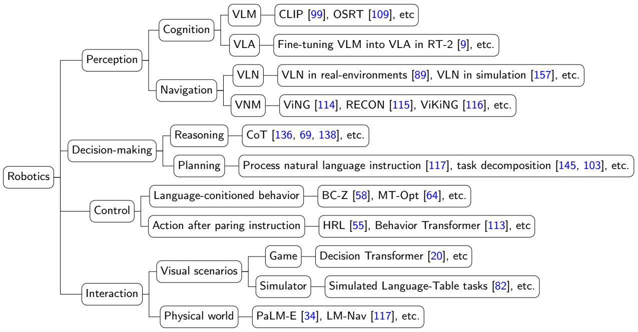

Robotics
Perception
Cognition
VLM
CLIP [99], OSRT [109], etc
VLA
Fine-tuning VLM into VLA in RT-2 [9], etc.
Navigation
VLN
VLN in real-environments [89], VLN in simulation [157], etc.
VNM
ViNG [114], RECON [115], ViKiNG [116], etc.
Decision-making
Reasoning
CoT [136, 69, 138], etc.
Planning
Process natural language instruction [117], task decomposition [145, 103], etc.
Control
Language-conitioned behavior
BC-Z [58], MT-Opt [64], etc.
Action after paring instruction
HRL [55], Behavior Transformer [113], etc
Interaction
Visual scenarios
Game
Decision Transformer [20], etc
Simulator
Simulated Language-Table tasks [82], etc.
Physical world
PaLM-E [34], LM-Nav [117], etc.

Figure 3: Components of robotics in this survey.

3. Related Technologies

In this section, we introduce the related technology used in robotics. Noticing that agents, embodied AI, and robotics based on LLMs all have the same meaning in this paper. Here we divided the model of robotics into four parts, which are perception, decision-making, control, and interaction. We certainly provide a more detailed introduction to the decision-making component, as it serves as the core of robotics based on LLMs. Decision-making serves as a connecting link between perception and control. We summarize the related technologies introduced below in Figure 3.

3.1. Perception

Perception is a fundamental capability of robots, akin to their input. Currently, multi-modality is a popular approach for robot perception. The models discussed below employ different treatments of perception.

[3] X-embodiment training means training robot policy with Open X-embodiment repository's datasets

[4] Open X-embodiment repository, a dataset consisting of different platforms. https://robotics-transformer-x.github.io

3.1.1. Vision-navigation model

Berkeley Autonomous Driving Ground Robot (BADGR) [62] is a mobile robot navigation system that leverages end-to-end learning and self-supervised non-policy data collected in real-world environments to train its algorithms without any simulation or human supervision. This innovative approach enables BADGR to navigate complex environments with ease and efficiency, paving the way for future advancements in autonomous driving technology. ViNG [114] is a goal-condition model that draws inspiration from GoalConditionedRL [37]. It is capable of predicting the temporal distance between image pairs and the corresponding actions to be performed. By integrating learned policies with topological maps constructed from previously observed data, ViNG's system can effectively determine how to achieve visually indicated goals, even in the presence of variable appearance and lighting conditions. RECON [115] is a system for robot learning designed for exploring autonomously and navigating in complex and unpredictable real-world surroundings. The core of RECON leverages a latent variable model of learning distance and action, along with non-parametric topology memory, to enable efficient and effective exploration. ViKiNG [116], built upon RECON mapping, incorporates geographical hints to propose an integrated learning and planning method that utilizes auxiliary information. This method combines a local traversability model. The model evaluates the robot's present camera observation and utilizes a potential sub-goal to infer the difficulty of achieving it. With a heuristic model that examines hints in the cost graph and evaluates the suitability of these sub-goals in achieving the overall goals, the general navigation model (GNM) [118] aims to train a general goal-condition model for vision-based navigation that can broadly generalize across diverse environments and embodiments, leveraging data from multiple structurally similar robots. By developing pre-trained navigation models with such capabilities, GNM represents a significant step toward realizing this vision that envisions applications for new types of robots.

3.1.2. Vision-language model

In recent years, large language models and visual models have had great success in their field. However, each of them can only process input in their own corresponding fields (for example, the language model only accepts text as input, and the visual model only accepts images as input), which is relatively simple. People began to focus on the processing of multi-modal input, combining large language models and visual models. Therefore, the multi-modal model that can take both vision and natural language as input was created — the visual-language model (VLM). VLM can process images and text at the same time. In actual use, we also need to distinguish between recognizing 2D scenes (such as some Visual Transformers (ViTs) [23, 33, 107]) or 3D scenes (such as OSRT [109]) when processing vision. VLMs come in various types [40]. There are many VLM models emerging. Contrastive Language-Image Pre-training (CLIP) [99] is a neural network that has been trained on diverse pairs of images and text. It has the capability to understand natural language instructions and predict the most pertinent text excerpts associated with a given image, all without directly optimizing for this specific task. CLIP is similar to the zero-shot function of GPT-2 and 3. CLIP is also used in LM-Nav [117] as a VLM to predict the text based on natural language. The landmarks are extracted and built into the topological map. VLM has the versatility to be employed in various downstream tasks including visual question answering (VQA) [151, 155], optical character recognition (OCR) [78], and image captioning [53]. Such as PaLM-E [34] treats text and images as latent vectors of multi-modal input. Frozen [129] is also processed similarly to PaLM-E.

3.1.3. Vision-and-language navigation model

One of the primary objectives of AI research is to develop an embodied intelligence that can effectively communicate with humans and interact with the environment. This embodied intelligence is capable of understanding human language and navigating its surroundings with ease, which has the potential to greatly benefit human society. However, achieving this goal is not without its challenges, including insufficient dataset, navigation processing strategies, processing of multi-modal inputs, and model migration from familiar environments to unfamiliar environments. Despite these obstacles, the development of embodied intelligence remains a crucial area of research in the field of AI [47]. Visual-and-language navigation (VLN) is a model that leverages visual observations to directly learn navigation implications and seamlessly links images and actions across time. As an extension of visual navigation in both real environments [89] and simulated [157], VLN boasts the capability to navigate complex 3D environments. There are many datasets in VLN that can be exploited.

3.1.4. Vision-language-action model

Can we pre-train a model that integrates multimodal inputs and low-level robot protocols to enhance the robot's generalization and semantic reasoning abilities? DeepMind aimed to develop a straightforward end-to-end model that could seamlessly map the robot's observations into action, thereby creating Vision-Language-Action Models (VLA) [9]. Prior approaches involved incorporating VLMs into robot policies or designing novel robot visual-language-action architectures. VLA instantiated by fine-tuning is first introduced and implemented in RT-2 [9], leveraging a large VLM. DeepMind fine-tunes the large-scale VLM [22, 34] and pre-trains it on a vast network-scale dataset, transforming VLM into VLA. To unify robot actions and natural language responses, DeepMind integrates actions as text tokens directly into the pre-trained dataset, forming multi-modal sentences [34]. Multimodal statements can respond to the command set generated by the robot through observation, outputting corresponding actions. This processing is analogous to LLM processing natural language data, where action-related tokens are decoded and converted into robot actions during interface processing. VLA can significantly enhance the generalization capabilities of robots.

3.2. Decision-making

Decision-making is a fundamental capability of robots, enabling them to make informed decisions and plan tasks based on their current state and environment. As the core of a robot, decision-making plays a crucial role in connecting the preceding and the following, analyzing input from the perception module to generate appropriate actions.

3.2.1. What brings intelligence to robotics?

LLM has the potential to significantly aid intelligent agents, with numerous studies successfully utilizing LLM as the brain to implement intelligent agents [10, 34, 117] and achieve promising results [93, 100]. Our ideal embodied intelligence should be an intelligent entity that can perceive the surrounding environment and produce corresponding output after interacting with humans or the environment. LLM plays a vital role in this process, serving as a central hub for analyzing multi-modal input and converting it into appropriate action output. The development of intelligent agents has progressed through various stages [142]: from symbolic agents relying on symbolic logic [43, 91]; Reactive agents prioritizing environmental interaction and instantaneously responding [12, 11]; Reinforcement learning-based agents trained to handle complex tasks [105] but lacking generalization [41]; Agents with transfer learning [15, 158] and meta-learning [48, 102] based on meta-learning and transfer learning to improve the generalization of the agent to the task. To the current LLM-based agents, where LLM is used as the brain of the agents [95, 122]. LLM can interpret inputs, plan output actions, and demonstrate reasoning even with the abilities of decision-making.

The emergence of ChatGPT [27] has sparked a surge of interest in LLMs within the scientific research community and industry in recent years. LLMs possess exceptional capabilities, often serving as the brains of agents, and have zero-shot and few-shot generalization abilities that enable them to adapt to various tasks without parameter updates. Their natural language understanding and generation capabilities are unparalleled, allowing them to gain reasoning and planning abilities [138]. Additionally, LLMs can parse high-level abstract instructions to perform complex tasks without requiring step-by-step guidance[5], and their human-like text-generation capabilities make them highly effective communicators [46]. Furthermore, LLMs can sense their environment [44], and technologies that expand their action space allow them to interact with the physical environment and complete tasks [149, 156]. They also possess reasoning and planning capabilities, such as logical and mathematical reasoning [134, 138], task decomposition [154], and planning [143] for specific tasks. LLM-based agents have been used in various real-world scenarios [77, 97] and have shown potential for multi-agent interactions and social capabilities. Overall, LLMs have revolutionized the field of artificial intelligence and hold great promise for future advancements.

3.2.2. Capacity of LLM in robotics

LLM serves as the brain of the robot, functioning as the central component that integrates knowledge, memory, and reasoning capabilities to enable the robot to plan and execute tasks intelligently.

Knowledge. The knowledge of LLM for robotics can be categorized into two types: the knowledge that needs to be acquired through learning (which is the pre-trained dataset) and the knowledge that has been learned and stored in memory [142].

- **Pre-trained data.** There are various types of pre-trained datasets available, and the more extensive and richer the knowledge learned, the stronger the LLM's generalization and natural language understanding capabilities will be [106]. Theoretically, the more a language model learns, the more parameters it has, enabling it to learn complex knowledge in natural language and gain powerful capabilities [65]. Research has shown that a richer dataset for language model learning can result in correct answers to diverse questions [106]. Datasets can be categorized into different types, such as basic semantic knowledge, which provides an understanding of language meaning [133]; Common sense, including everyday facts like people eating when hungry or the sun rising in the east [108]; Professional field knowledge, which can aid humans in completing tasks like programming [146] and mathematics [24].

- **Memory.** Just like human memory, embodied intelligence should be able to formulate strategies and make decisions for new tasks based on experiences (i.e., observed actions, thoughts, etc.). When faced with complex tasks, the memory mechanism can aid in reviewing past strategies to obtain more effective solutions [56, 121]. However, memory poses some challenges, such as the length of memory sequences and how to efficiently store and index them as the number of memories grows. As the robot's memory burden increases over time, it must be able to effectively manage and retrieve memories to avoid catastrophic forgetfulness [68].

Reasoning. Reasoning serves as a foundational element in human cognition, playing a crucial role in problem-solving, decision-making, and the analytical examination of information [135, 136]. Reasoning plays a crucial role in enabling LLMs to solve complex tasks. Reasoning capabilities allow LLMs to break down problems into smaller, manageable steps and solve them starting from the current status and known conditions. There is ongoing debate about how LLMs acquire their reasoning abilities, with some arguing that it is a result of pre-training or fine-tuning [54], while others believe that it emerges only at a certain scale [137].

[5] BabyAGI https://github.com/yoheinakajima/babyagi

Research has shown that Chain-of-Thought (CoT) [136] can help LLMs reveal their reasoning capabilities, and some studies suggest that inference abilities may stem from the local static structure of the training data.

Planning. Humans plan when faced with complex challenges. Planning can help people organize their thoughts, set goals, and decide what they should do in the current situation [45, 130]. In this case, they can gradually approach their goals. The core of planning is reasoning. The agent can use reasoning capabilities to deconstruct the received high-level abstract instructions into executable subtasks and make reasonable plans for each subtask [26, 112]. For example, LM-Nav uses ChatGpt to process received natural language instructions [117]. PaLM-E directly implements end-to-end processing, converting the received multi-modal input into multi-modal sentences for LLM processing [34]. Agents may also be able to reasonably update task planning based on the current situation through multiple rounds of dialogue and self-questioning and answering in the future. Many studies have proposed methods of dividing the execution tasks into many executable small tasks during the planning process. For example, directly break down the execution task into many small tasks and execute them sequentially [103, 145]. CoT only processes one sub-task at a time and can adaptively complete the task, which has a certain degree of flexibility [69, 138]. There are also some vertical planning methods that divide tasks into tree diagrams [49, 148].

3.3. Control

Here, we argue that the control module is the key component responsible for regulating robotic actions. This module plays a crucial role in ensuring that the robot's actions are executed accurately and successfully, with a focus on the hardware aspects of action execution.

3.3.1. How to learn language-conditioned behavior

Much of the previous work has focused on enabling robots and other agents to comprehend and execute natural language instructions [19, 35, 81]. There are various approaches to learning linguistically conditioned behaviors, such as image-based behavioral cloning that follows the BC-Z [58] method or the MT-Opt [64] reinforcement learning method. Imitation learning techniques train protocols on demonstration datasets [58, 153], while offline reinforcement learning has also been studied extensively [59, 71, 88]. However, some works suggest that imitation learning on demonstration data performs better than offline reinforcement learning [83], and other studies demonstrate the feasibility of offline reinforcement learning in theory and practice [72, 73]. Many works combine RL and Transformer structures [20, 60], and some works integrate imitation learning with reward conditions, such as Decision Transformer (DT) [20], namely combines imitation learning with reinforcement learning elements. However, DT does not enable the model to learn from the demonstration dataset to have better performance. Deep Skill Graphs (DSG) [5] present a novel approach to skill learning utilizing the option framework. This method leverages graphs to represent discrete aspects of the environment, enabling the model to acquire structured knowledge and learn complex skills within the given domain. CT employs goal-conditioned RL to transform the local skill-learning problem into a goal-conditioned Markov decision process (MDP) [61].

In the context of navigation robots, early approaches to enhancing navigation strategies with the natural language employed static machine translation [80] to discover patterns. The process involves utilizing discovery patterns to translate free-form instructions into formal languages that adhere to specific grammatical rules [19, 85, 127]. However, these methods were limited to structured state spaces. Recent works have also developed the VLN task as a sequence prediction problem [3, 87, 119]. Additionally, there are methods that leverage nearly 1M labeled simulation trajectory demonstration data for training [47], but applying these models in unstructured environments remains a significant challenge. Data-driven approaches for vision-based mobile robot navigation often depend on the utilization of realistic simulation techniques [70, 111, 144] or gathering supervised data to directly learn policies for achieving goals based on observations [38]. Alternatively, self-supervised learning methods can utilize unlabeled datasets or trajectories generated automatically by onboard sensors and hindsight relabeling learning [51, 63, 114].

3.3.2. How to execute action after parsing nature language

To determine whether a skill can be executed in the current state after parsing a natural language command, a temporal-difference-based (TD) reinforcement learning approach can be employed. This method learns a value function to evaluate whether the skill is executable or not [1]. The value function is derived from the corresponding affordance function of reinforcement learning [42]. Additionally, LM-Nav [117] utilizes a self-supervised learning method to enhance the parsing of free-form language instructions leveraging pre-trained VLM in a large number of previous environments. To address the challenges of long-term tasks, hierarchical reinforcement learning (HRL) [55] can be employed, where higher-level policies play a role in setting objectives for lower-level protocols to execute [90, 132]. The process of mapping natural language and observations into robot actions can also be viewed as a sequence modeling problem [9, 10, 29]. Transformer-based robot control, such as the Behavior Transformer [113], focuses on learning demonstrations that correspond to each task. Gato [104] suggests training a model on large datasets including robotic and non-robotic.

3.4. Interaction

Interaction serves as a fundamental module that enables robots to engage and interact with both the environment and humans. To enhance robots' ability to interact in the physical world, they are often trained extensively. While some researchers utilize artificial intelligence to interact in virtual environments, such as games or simulations, ultimately, these models must be transferred to the real world.

However, the accuracy of these models tends to be lower in real-world settings compared to simulated environments.

3.4.1. Game

Traditional game developers manually write over a dozen character behaviors (including class methods and attributes) for the implementation of a game in the Valentine's Day party's specific game environment. Almost all of these behaviors are fixed sets, making the process very cumbersome with poor scalability. In games, LLMs have been used to create interactive novels and text adventure games [17]. LLMs are increasingly utilized for planning robotic tasks due to their capacity to generate and decompose sequences of actions. In GA [95], they created a computer program that can mimic the behavior of human beings, called the *Generative Agents*. It extends the LLM by using natural language to store complete records of the intelligentsia's experiences. Synthesizing accumulated memories and reflecting upon them at higher levels over time, the system can dynamically retrieve these memories to plan and guide its behavior. Agent characters engage in comprehensive verbal exchanges utilizing authentic human language. They possess knowledge of other intelligent entities within their vicinity, and the generative agent framework dictates whether they proceed to interact or initiate a dialogue. These intelligent agents' characters can exhibit quite realistic personal behavior and social interactions. For example, when someone tells one of the agents that they have a desire to organize and host a festive gathering to celebrate Valentine's Day, these agents will spontaneously invite others to attend, meet each other, date, and be on time for the party together. This innovative architecture empowers generative agents with the ability to retain, recall, contemplate, engage with fellow agents, and strategize amidst ever-changing circumstances.

3.4.2. Language-based human-robot interaction

There are GUI (Graphical User Interface) and LUI (Language User Interface) for human-robot interaction. GUI refers to a computer-operated user interface that is graphically displayed and uses an interactive device to manage the interaction with the system. Unlike GUI, LUI can directly use natural human language for human-robot interaction, and the most representative LUI product is ChatGPT. Traditionally, the task of simulating human-robot interaction using natural language has proven to be difficult due to the constraints imposed on users by rigid instructions, or the need for intricate algorithms to manage numerous probability distributions related to actions and target objects[4]. However, it is not easy to translate instructions into commands that robots can understand in the real world, and traditionally, fixed collections of desired actions and directives have been used to enable robots to understand human language. However, this can significantly limit the robot's flexibility and has limited generalizability across different hardware platforms. The LAnguage Trajectory TransformEr [16] introduces a versatile language-driven framework that empowers users to customize and adapt the overall trajectories of robots. The approach leverages pre-trained language models (e.g., BERT [31] and CLIP [99]) to encode the user's intention and target objects directly from unrestricted text inputs and scene images. It combines geometric features produced by a network of transformer encoders and generates the trajectory using a transformer decoder, eliminating the requirement for prior task-related or robot-specific information.

Considering the vagueness and ambiguity of natural language, from the point of view of human-robot interaction, robots should enhance the initiative of interaction in the future, that is to say, let the robot actively ask the user questions through the large language model. If the robot feels that the user's words are problematic and is not sure what they mean, it should ask you back what you mean or whether you mean what you say.

4. Applications of LLMs in Robotics

Applications of large models and robotics across various domains. Here are ten specific applications of the combination of large models and robotics, along with their explanations:

- **Autonomous navigation and path planning.** Large models provide powerful semantic understanding and reasoning capabilities for robots, assisting them in autonomous navigation and path planning in unknown environments. By combining large models with sensor data, robots can comprehend semantic information in the environment, recognize obstacles, target locations, and navigation objectives, and generate suitable path-planning solutions [25].

- **Speech interaction and NLP.** LLMs excel in speech recognition, semantic understanding, and natural language generation. Robots can leverage large models for speech interaction, understanding and answering user queries, executing specific tasks, and providing personalized service experiences.

- **Visual perception and object recognition.** Large models possess strong capabilities in image and video analysis, aiding robots in object recognition, target detection, and scene understanding. By integrating deep learning and large models, robots can achieve efficient and accurate visual perception, which can be applied in autonomous driving, robot vision-based navigation, and industrial automation.

- **Human-robot collaboration and social robots.** Large models with natural language processing and emotion analysis help robots understand human feelings and intentions better, making interactions between humans and robots more natural and smart. Social robots can engage in conversations, comprehend emotions, and provide companionship and support, which are applied in fields like healthcare, education, and entertainment.

- **Humanoid robots and emotional expression.** Large models can help humanoid robots better understand and express emotions. Through natural language generation and emotion recognition technologies, robots can engage in emotional communication and expression with humans, providing emotional support and companionship.

- **Industrial automation and robot control.** Large models can be combined with sensor data for industrial process monitoring, anomaly detection, and predictive maintenance. By learning and analyzing large-scale data, robots can achieve intelligent industrial automation and adaptive control.

- **Healthcare and rehabilitation robots.** Large models can be applied in medical and rehabilitation robots to assist in diagnosis, treatment, and patient care. Robots can analyze medical images, patient data, and clinical records, aiding in disease detection, surgical planning, and personalized therapy. They can also provide physical assistance and rehabilitation exercises for mobility-impaired patients.

- **Environmental monitoring and exploration.** Large models can be combined with robot platforms for monitoring and exploration in various environments, such as oceans, forests, and disaster sites. These robots can analyze sensor data, satellite imagery, and other environmental data to monitor pollution levels, detect natural disasters, and explore uncharted territories.

- **Agriculture and farm mechanization.** Large models and robots can be applied in agriculture and farm mechanization, optimizing crop management, monitoring plant health, and automating labor-intensive tasks. Robots equipped with sensors and cameras can collect data from farmlands, and analyze soil conditions, climate changes, and crop requirements, providing farmers with decision support to enhance agricultural productivity and sustainability.

- **Education and learning assistance.** Large models and robots can provide personalized tutoring and learning support in the field of education. Robots can interact with students, and then offer personalized learning materials and guidance based on their abilities and needs [2]. Leveraging the semantic understanding and knowledge reasoning capabilities of large models, robots can answer questions, explain concepts, and help students deepen their understanding of knowledge.

In summary, the combination of large models and robotics holds tremendous potential across various domains, including autonomous navigation, speech interaction, visual perception, human-robot collaboration, industrial automation, healthcare, environmental monitoring, agriculture, and education. It can bring convenience and innovation to human life and work.

5. Challenges

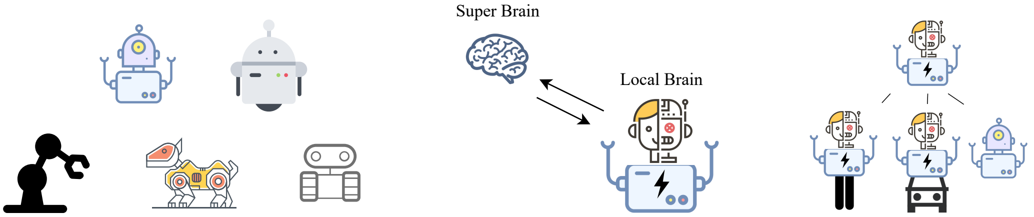
5.1. Datasets

In the realm of Web 3.0 [39], big data [123], AI-Generated Content (AIGC) [140], and machine learning, collecting datasets has always been a challenge. Currently, training LLMs require vast amounts of data to support their capabilities, particularly high-quality datasets that consume considerable resources. In the field of robotics, collecting datasets is even more difficult. While LLM like ChatGPT relies on text data for pre-training [14], VLM uses a combination of text and image data [99]. Robotics, however, requires a combination of both, with the addition of multimodal data, such as text, images, and touch, to serve as the robot's sensory input. These diverse datasets need to be processed in a unified format [34], allowing the robot's brain to plan and divide tasks effectively. Unfortunately, there is a lack of ready-made, multi-modal datasets, and collecting them requires a significant time investment. Moreover, policy control is necessary, which includes the interaction between the robot and its environment, necessitating 3D data [7]. The data required for robotics are diverse and scarce, with poor general applicability. For instance, a dataset used to train robot dogs cannot be applied to humanoid robots, and a dataset used for screwing in an assembly line may not be suitable for robots that assemble items. However, with the emergence of platforms similar to X-embodiment[6], the challenges of dataset collection in robotics may be alleviated in the future.

5.2. Training Scemes

As embodied intelligence necessitates interaction with the physical environment, the model's training requires specific scenarios, e.g., distributed training [152]. Current research involves training robot-related models in various environments, such as games [95], simulations [30], and real-world scenarios [8]. Training in-game scenarios is straightforward, with simple operations like button-pressing. However, the knowledge gained from games may not translate well to real-world scenarios, as the information in complex scenes varies greatly, and language models cannot provide a universal solution. Simulation environments aim to closely replicate reality, with low energy consumption and cost. However, modeling real scenes in simulators can be necessary. While game and simulation environments can train models, they share a common issue: poor transferability to real scenes. For instance, a model with 90% accuracy in a game or simulation may only have 10% accuracy in a real scene. Real-scene training faces significant challenges, such as cost. In simulations, objects can be generated through code [21], but in reality, purchasing them can be expensive. Transferring models between different training scenarios is a significant challenge.

[6] Open X-embodiment repository, a dataset consisting of different platforms. https://robotics-transformer-x.github.io/



(a) Robotics with different shapes.

(b) LLM development.

(c) Modualize.

Figure 4: Challenge in embodied intelligence.

5.3. Shape

Currently, most work environments in human society are well-suited for humanoid robots. However, the question arises whether robots must be human-shaped [57]. There are numerous types of robots currently in existence, each with its unique capabilities and applications, like in Figure 4(a). From an energy consumption perspective, wheels are more energy-efficient than legs. Therefore, if a humanoid robot is built, it may be inappropriate to use legs to move objects instead of a conveyor belt. Similarly, a chef robot may not need to hold a shovel and cook like a human. In many cases, designing a pipeline tailored to the specific task at hand can lead to more efficient automation than humanoid robots. While humanoid robots are often depicted in animation scenes, such as in animation like Mobile Suit Gundam or games like Armored Core, their design may not always be practical for applications. For instance, a robot designed solely for washing dishes may not need the ability to sing. Modular concepts like Expedition A1[7], can offer optimal results for different scenarios by replacing certain components. The shape of the robot remains a topic of debate, and the decision should ultimately focus on suitability for the task at hand.

5.4. LLM Deployment

Given embodied intelligence, the question arises regarding the deployment of its brain. Current technical limitations prevent the LLM from being deployed locally on the robot. The prevailing industry practice involves employing two brains: a cloud-based super brain and a local brain, like in Figure 4(b). However, a unified consensus on this device-side plus cloud testing deployment method has yet to be established. A feasible solution could be to create a dynamic, compact model on the local client side, capable of handling basic scenario interactions. The cloud-based super brain, on the other hand, would tackle complex and challenging problems. The LLM deployment architecture remains a pressing issue that must be addressed in the future development of agents. This deployment structure also introduces latency issues, as information exchange between the robot and the super brain requires signal transmission. In certain environments, such as those with signal loss, the robot may be left with only its local brain, potentially leading to control loss or unpredictable behavior.

5.5. Security

LLM like ChatGPT may harbor biases or misconceptions stemming from their pre-training data. These biases can manifest in problematic guidance for users, and robots that rely on LLM as their brains may also exhibit biases [142]. Since robots' outputs are typically physical actions, biased or misunderstood guidance can lead to harmful consequences for users [36, 98], such as a chef robot burning down a house while cooking. Beyond physical safety risks, robots also raise concerns about data security [86]. For instance, a robot butler who resides in a home may become intimately familiar with the household's environment and occasionally require cloud interaction for certain tasks. During user interaction, there is a risk of private data leakage, which could be mitigated by an offline environment, but this may compromise the robot's performance.

5.6. Dialogue Consistency

Humans often don't complete tasks in a single, static step. Instead, they iteratively adjust strategies and goals based on feedback received after taking action. The same is true for embodied intelligence. When faced with high-level, abstract, or ambiguous commands, robots may not be able to decompose them into executable small tasks at first. They need to obtain further feedback from the environment and humans through continuous dialogue to update their goals. Without this ability to engage in continuous dialogue, which enables robots to perform tasks dynamically, their performance will be significantly impaired [120]. Moreover, the maximum length limit of a robot's context is another issue worth considering. Typically, embodied intelligence may play a housekeeper role, handling daily tasks like washing dishes or drying clothes. However, for long-term tasks like scientific research, robots require more context-understanding capabilities. Currently, there's a limit to the length of context that robots can handle, and this limitation can lead to catastrophic forgetting [68]. Dialogue persistence is a crucial challenge for long-term tasks.

5.7. Social Influence

The rapid advancement of LLMs is bringing the era of embodied intelligence, as depicted in science fiction movies and games, closer to reality. This technological breakthrough will undoubtedly revolutionize human society and unleash

[7] https://www.agibot.com

unprecedented productivity. With robots capable of performing repetitive tasks, the need for human labor in various industries will diminish. However, this shift may also have far-reaching consequences, potentially disrupting social structures and stability [50]. As robots replace low-end manual labor, it raises questions about the fate of those who previously held these jobs. The double-edged sword of embodied intelligence presents both liberation and disruption. While automation may usher in unprecedented efficiency, it also poses challenges for societal adaptation. Some works of science fiction, such as Detroit Become Human, depict a future where robots gain consciousness and conflict with humans, leading to a war between the two. Alternatively, technology may fall into the wrong hands, becoming a tool for exploitation and solidifying class divisions. However, in a worst-case scenario, robots may become a replacement for humans. As we embrace the development of embodied intelligence, we must also confront the ethical and societal implications it entails.

5.8. Ethic

Embodied intelligence has long been regarded as a mere tool, but it may hold more significance in the eyes of some users. For instance, companion robots can bring solace to lonely individuals, much like a loyal companion. In fact, some people even develop emotional attachments to their first car or a vehicle that has been with them for a long time. If we were to create robots that resemble humans or exhibit human-like intelligence, would they evoke different emotions? In science fiction movies, robots that gain self-awareness and break free from their programming often develop emotions and even marry humans. Interestingly, robots powered by LLMs have already demonstrated a degree of intelligence. Will they eventually become conscious? If embodied intelligence evolves to possess consciousness, should we still consider them tools? This raises questions about the definition of conscious robots and whether they can be considered human. Although this challenge is still far off in the future of smart robot development, it is an intriguing topic to ponder.

6. Promising Directions for Future Work

6.1. Security of Task Executing

Security has always been a pressing concern in various models, particularly with regard to user privacy. However, we argue that the safety of agents during task execution is of paramount importance. In this article, we explore the question of whether an agent's actions during task execution could cause harm [98, 36]. For instance, consider a scenario where a robot is asked to make lunch, but in the process, it sets the kitchen on fire. In other scenes, imagine a robot tasked with killing fish, but it mistakenly identifies humans as fish and proceeds to chase and harm them. These scenarios highlight the need to limit the actions an agent can perform to prevent potential harm. Current robot systems focus on enabling the robot to determine which actions can be performed based on the current state and environment, without fully considering the consequences of executing those actions. Therefore, we propose that ensuring the safety of task execution must be a top priority, by guaranteeing that the robot's actions do not harm human rights and interests.

6.2. Training Scenario Transfer

Due to technical or economic constraints, it is common to train robot action policies in simulated [30] or gaming environments [95]. However, the ultimate goal of agent training is to apply it in real-world scenarios. Unfortunately, training in diverse scenarios can lead to not being acclimatized, which may compromise the agent's performance when deployed in real-world situations. The fundamental source of this problem can be attributed to the disparity of feedback mechanisms between simulated and real-world environments. In games or simulations, feedback is often more straightforward, with the robot receiving clear and concise information about the outcome of its actions. In contrast, real-world feedback is more complex and nuanced, making it challenging to assess the feasibility of a task in a limited scenario. Therefore, a valuable research direction is to explore methods for transferring model training across different scenarios while maintaining their accuracy in the original training environments.

6.3. Unify Format of Modal

Currently, many models are utilizing LLM as the robot's brain, and text-type data is typically the input that LLM accepts. However, for agents reliant on multi-modal perception, efficiently handling diverse input formats poses a significant challenge. To address this issue, a VLA model has been proposed [9], which uniformly converts visual and natural language multi-modal inputs into multi-modal sentences for processing, and outputs actions in the same format. In other words, multi-modal statements are employed to harmonize input and output. Nevertheless, there is currently no unified processing for other modalities such as touch and smell. It is anticipated that unified multi-modal models like VLA will gain popularity in the future.

6.4. Modular Components

As previously discussed, the field of robotics currently lacks a unified approach to robot design, with varying opinions on the matter. We believe that there should be a modular design method, wherein each part of the robot can be swapped out like a machine, just like in Figure 4(c), allowing for greater versatility and adaptability[8]. To achieve this, we must first establish unified specifications for the various modules of the robot. For instance, a robot can be composed of a head, torso, upper limbs, and lower limbs, with the upper limbs and lower limbs being interchangeable based on the task at hand. Among them, the upper limbs and lower limbs can be replaced according to specific tasks. When we need to cook, we can use our upper limbs as a shovel, and when

[8] https://www.agibot.com

we need to deal with weeds in the yard, we can use our lower limbs as a weeder.

6.5. Autonomous Perception

Our current research focuses on developing robots that can interact with humans using natural language instructions. In many cases, we study how humans issue instructions and how robots can decompose abstract tasks into specific sub-tasks for execution [1]. However, we also hope that robots can perceive and respond autonomously to handle our current needs. For instance, if our cup falls to the ground and breaks, an agent should be able to perceive the situation through hearing and vision, and then autonomously handle the glass fragments for us. Autonomous perception requires the robot to have common sense, which is a capability that can be integrated into robots based on LLM as the brain. Research on robots' autonomous perception capabilities is crucial for improving our quality of life in the future.

7. Conclusions

In this survey, we summarized the methods and technologies currently used for large models in robots. First, we reviewed some basic concepts of large language models and common large models. We explain what improvements will be brought to robots by using large models as brains. We also introduce the representative LLM-based robot models proposed in recent years, such as LM-Nav [117], PaLM-SayCan [1], PaLM-E [34], etc. Next, we divide the robot into four modules: perception, decision-making, control, and interaction. For each module, we discuss the relevant technologies and their functions, including the perception module's ability to process the robot's input from the surroundings; the decision-making module's capacity to understand human instructions and plan; the control module's role in processing output actions; the interaction module's ability to interact with the environment. We also explore the potential application scenarios of current robots based on LLMs and discuss the challenges, such as training, safety, shape, deployment, and long-term task performance. Finally, we consider the social and ethical implications of post-intelligent robots and their potential impact on human society.

As LLMs continue to evolve, robots may become increasingly intelligent and capable of processing instructions and tasks more efficiently. With advancements in hardware, robots may eventually become reliable assistants for humans, as depicted in science fiction movies. However, we must also be mindful of their potential impact on society and address any concerns proactively. Embodied intelligence is a new paradigm for the development of intelligent science and is of great significance in leading the development of the future. LLM-based robotics represent a potential path to embodied intelligence. We hope this survey can provide some inspiration to the community and facilitate research in related fields.

Acknowledgment

This research was supported in part by the National Natural Science Foundation of China (Nos. 62002136 and 62272196), the Natural Science Foundation of Guangdong Province (No. 2022A1515011861), Engineering Research Center of Trustworthy AI, Ministry of Education (Jinan University), and Guangdong Key Laboratory of Data Security and Privacy Preserving.

References

[1] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al., 2022. Do as i can, not as i say: Grounding language in robotic affordances, in: The Conference on Robot Learning, pp. 287–318.

[2] Alam, A., 2022. Social robots in education for long-term human-robot interaction: socially supportive behaviour of robotic tutor for creating robo-tangible learning environment in a guided discovery learning interaction. ECS Transactions 107, 12389.

[3] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A., 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674–3683.

[4] Arkin, J., Park, D., Roy, S., Walter, M.R., Roy, N., Howard, T.M., Paul, R., 2020. Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions. The International Journal of Robotics Research 39, 1279–1304.

[5] Bagaria, A., Senthil, J.K., Konidaris, G., 2021. Skill discovery for exploration and planning using deep skill graphs, in: International Conference on Machine Learning, PMLR. pp. 521–531.

[6] Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate, in: the International Conference on Learning Representations.

[7] Bermudez, L., . Overview of embodied artificial intelligence. https://medium.com/machinevision/overview-of-embodied-artificial-intelligence-b7f19d18022.

[8] Bharadhwaj, H., Vakil, J., Sharma, M., Gupta, A., Tulsiani, S., Kumar, V., 2023. RoboAgent:: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. arXiv preprint, arXiv:2309.01918 .

[9] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al., 2023a. RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint, arXiv:2307.15818 .

[10] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al., 2023b. RT-1: Robotics transformer for real-world control at scale. Robotics: Science and Systems XIX .

[11] Brooks, R., 1986. A robust layered control system for a mobile robot. IEEE Journal on Robotics and Automation 2, 14–23.

[12] Brooks, R.A., 1991. Intelligence without representation. Artificial Intelligence 47, 139–159.

[13] Brown, P.F., Della Pietra, V.J., Desouza, P.V., Lai, J.C., Mercer, R.L., 1992. Class-based n-gram models of natural language. Computational linguistics 18, 467–480.

[14] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901.

[15] Brys, T., Harutyunyan, A., Taylor, M.E., Nowé, A., 2015. Policy transfer using reward shaping., in: The International Conference on Autonomous Agents and Multiagent Systems, pp. 181–188.

[16] Bucker, A., Figueredo, L., Haddadin, S., Kapoor, A., Ma, S., Vemprala, S., Bonatti, R., 2023. LATTE: Language trajectory

transformer, in: IEEE International Conference on Robotics and Automation, IEEE. pp. 7287–7294.

[17] Burkinshaw, R., 2009. Alice and kev: The story of being homeless in the sims 3. Retrieved February 19, 2010.

[18] Chebotar, Y., Vuong, Q., Irpan, A., Hausman, K., Xia, F., Lu, Y., Kumar, A., Yu, T., Herzog, A., Pertsch, K., et al., 2023. Q-Transformer: Scalable offline reinforcement learning via autoregressive q-functions. arXiv preprint, arXiv:2309.10150 .

[19] Chen, D., Mooney, R., 2011. Learning to interpret natural language navigation instructions from observations, in: The AAAI Conference on Artificial Intelligence, pp. 859–865.

[20] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., Mordatch, I., 2021. Decision transformer: Reinforcement learning via sequence modeling. Advances in Neural Information Processing Systems 34, 15084–15097.

[21] Chen, P.L., Chang, C.S., 2023. InterAct: Exploring the potentials of chatgpt as a cooperative agent. arXiv preprint, arXiv:2308.01552 .

[22] Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., et al., 2023a. PaLI-X: On scaling up a multilingual vision and language model. arXiv preprint, arXiv:2305.18565 .

[23] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al., 2023b. PaLI: A jointly-scaled multilingual language-image model. International Conference on Learning Representations .

[24] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al., 2021. Training verifiers to solve math word problems. arXiv preprint, arXiv:2110.14168 .

[25] Crespo, J., Castillo, J.C., Mozos, O.M., Barber, R., 2020. Semantic information for robot navigation: A survey. Applied Sciences 10, 497.

[26] Crosby, M., Rovatsos, M., Petrick, R., 2013. Automated agent decomposition for classical planning, in: The International Conference on Automated Planning and Scheduling, pp. 46–54.

[27] Dale, R., 2021. GPT-3: What's it good for? Natural Language Engineering 27, 113–118.

[28] Davison, A.C., Reid, N., 2021. The tangent exponential model. arXiv preprint arXiv:2106.10496 .

[29] DeepMind, . Open X-Embodiment: Robotic learning datasets and RT-X models. https://robotics-transformer-x.github.io/paper.pdf.

[30] Devin, C., Gupta, A., Darrell, T., Abbeel, P., Levine, S., 2017. Learning modular neural network policies for multi-task and multi-robot transfer, in: IEEE International Conference on Robotics and Automation, IEEE. pp. 2169–2176.

[31] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of deep bidirectional transformers for language understanding, in: The Conference of the North American Chapter of the Association for Computational Linguistics, ACL. pp. 4171–4186.

[32] Dorigo, M., Theraulaz, G., Trianni, V., 2021. Swarm robotics: Past, present, and future [point of view]. Proceedings of the IEEE 109, 1152–1165.

[33] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations .

[34] Driess, D., Xia, F., Sajjadi, M.S.e.a., 2023. PaLM-E: An embodied multimodal language model, in: International Conference on Machine Learning, pp. 8469–8488.

[35] Duvallet, F., Walter, M.R., Howard, T., Hemachandra, S., Oh, J., Teller, S., Roy, N., Stentz, A., 2016. Inferring maps and behaviors from natural language instructions, in: The 14th International Symposium on Experimental Robotics, Springer. pp. 373–388.

[36] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al., 2021. A mathematical framework for transformer circuits. Transformer Circuits Thread 1.

[37] Eysenbach, B., Salakhutdinov, R.R., Levine, S., 2019. Search on the replay buffer: Bridging planning and reinforcement learning. Advances in Neural Information Processing Systems 32.

[38] Francis, A., Faust, A., Chiang, H.T.L., Hsu, J., Kew, J.C., Fiser, M., Lee, T.W.E., 2020. Long-range indoor navigation with PRM-RL. IEEE Transactions on Robotics 36, 1115–1134.

[39] Gan, W., Ye, Z., Wan, S., Yu, P.S., 2023. Web 3.0: The future of internet, in: Companion Proceedings of the ACM Web Conference, pp. 1266–1275.

[40] Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J., et al., 2022. Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision 14, 163–352.

[41] Ghosh, D., Rahme, J., Kumar, A., Zhang, A., Adams, R.P., Levine, S., 2021. Why generalization in RL is difficult: Epistemic pomdps and implicit partial observability. Advances in Neural Information Processing Systems 34, 25502–25515.

[42] Gibson, J.J., 1977. The theory of affordances. Hilldale, USA 1, 67–82.

[43] Ginsberg, M., 2012. Essentials of artificial intelligence. Newnes.

[44] Goodwin, R., 1995. Formalizing properties of agents. Journal of Logic and Computation 5, 763–781.

[45] Grafman, J., Spector, L., Rattermann, M.J., 2004. Planning and the brain, in: The cognitive psychology of planning. Psychology Press, pp. 191–208.

[46] Gravitas, S., 2023. Auto-GPT: An Autonomous GPT-4 Experiment.

[47] Gu, J., Stefani, E., Wu, Q., Thomason, J., Wang, X.E., 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions, in: Annual Meeting of the Association for Computational Linguistics, pp. 7606–7623.

[48] Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., Levine, S., 2018. Meta-reinforcement learning of structured exploration strategies. Advances in neural information processing systems 31.

[49] Hao, S., Gu, Y., Ma, H., Hong, J.J., Wang, Z., Wang, D.Z., Hu, Z., 2023. Reasoning with language model is planning with world model. arXiv preprint, arXiv:2305.14992 .

[50] Helberger, N., Diakopoulos, N., 2023. Chatgpt and the AI Act. Internet Policy Review 12.

[51] Hirose, N., Xia, F., Martín-Martín, R., Sadeghian, A., Savarese, S., 2019. Deep visual mpc-policy learning for navigation. IEEE Robotics and Automation Letters 4, 3184–3191.

[52] Hrabia, C.E., Lützenberger, M., Albayrak, S., 2018. Towards adaptive multi-robot systems: Self-organization and self-adaptation. The Knowledge Engineering Review 33, e16.

[53] Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L., 2022. Scaling up vision-language pre-training for image captioning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17980–17989.

[54] Huang, J., Chang, K.C.C., 2023. Towards reasoning in large language models: A survey, in: The Association for Computational Linguistics.

[55] Hutsebaut-Buysse, M., Mets, K., Latré, S., 2022. Hierarchical reinforcement learning: A survey and open research challenges. Machine Learning and Knowledge Extraction 4, 172–221.

[56] Hutter, M., 2000. A theory of universal artificial intelligence based on algorithmic complexity. arXiv preprint, cs/0004001 .

[57] Hwang, J., Park, T., Hwang, W., 2013. The effects of overall robot shape on the emotions invoked in users and the perceived personalities of robot. Applied Ergonomics 44, 459–471.

[58] Jang, E., Irpan, A., Khansari, M., Kappler, D., Ebert, F., Lynch, C., Levine, S., Finn, C., 2022. BC-Z: Zero-shot Task Generalization with Robotic Imitation Learning, in: The Conference on Robot Learning, PMLR. pp. 991–1002.

[59] Jang, Y., Lee, J., Kim, K.E., 2021. GPT-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems, in: International Conference on Learning Representations.

[60] Janner, M., Li, Q., Levine, S., 2021. Offline reinforcement learning as one big sequence modeling problem. Advances in Neural Information Processing Systems 34, 1273–1286.

[61] Kaelbling, L.P., 1993. Learning to achieve goals, in: The International Joint Conference on Artificial Intelligence, Citeseer. pp. 1094–8.

[62] Kahn, G., Abbeel, P., Levine, S., 2021. BADGR: An autonomous self-supervised learning-based navigation system. IEEE Robotics and Automation Letters 6, 1312–1319.

[63] Kahn, G., Villaflor, A., Ding, B., Abbeel, P., Levine, S., 2018. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation, in: IEEE International Conference on Robotics and Automation, IEEE. pp. 5129–5136.

[64] Kalashnikov, D., Varley, J., Chebotar, Y., Swanson, B., Jonschkowski, R., Finn, C., Levine, S., Hausman, K., 2021. Mt-Opt: Continuous multi-task robotic reinforcement learning at scale. arXiv preprint, arXiv:2104.08212 .

[65] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint, arXiv:2001.08361 .

[66] Károly, A.I., Galambos, P., Kuti, J., Rudas, I.J., 2020. Deep learning in robotics: Survey on model structures and training strategies. IEEE Transactions on Systems, Man, and Cybernetics: Systems 51, 266–279.

[67] Kavraki, L.E., Svestka, P., Latombe, J.C., Overmars, M.H., 1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transactions on Robotics and Automation 12, 566–580.

[68] Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C., 2018. Measuring catastrophic forgetting in neural networks, in: The AAAI Conference on Artificial Intelligence, pp. 3390–3398.

[69] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems 35, 22199–22213.

[70] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al., 2017. Ai2-thor: An interactive 3d environment for visual AI. arXiv preprint, arXiv:1712.05474 .

[71] Kostrikov, I., Nair, A., Levine, S., 2021. Offline reinforcement learning with implicit Q-learning. International Conference on Learning Representations .

[72] Kumar, A., Hong, J., Singh, A., Levine, S., 2022a. When should we prefer offline reinforcement learning over behavioral cloning? arXiv preprint, arXiv:2204.05618 .

[73] Kumar, A., Singh, A., Ebert, F., Yang, Y., Finn, C., Levine, S., 2022b. Pre-training for robots: Offline RL enables learning new tasks from a handful of trials. arXiv preprint, arXiv:2210.05178 .

[74] Kumar, A., Zhou, A., Tucker, G., Levine, S., 2020. Conservative Q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems 33, 1179–1191.

[75] Lawson, D., Qureshi, A.H., 2022. Control Transformer: Robot navigation in unknown environments through prm-guided return-conditioned sequence modeling. arXiv preprint, arXiv:2211.06407 .

[76] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

[77] Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D., Ghanem, B., 2023a. CAMEL: Communicative agents for" mind" exploration of large language model society, in: Thirty-seventh Conference on Neural Information Processing Systems.

[78] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F., 2023b. TrOCR: Transformer-based optical character recognition with pre-trained models, in: The AAAI Conference on Artificial Intelligence, pp. 13094–13102.

[79] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint, arXiv:1907.11692 .

[80] Lopez, A., 2008. Statistical machine translation. ACM Computing Surveys 40, 1–49.

[81] Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., Rocktäschel, T., 2019. A survey of reinforcement learning informed by natural language. The International Joint Conference on Artificial Intelligence .

[82] Lynch, C., Sermanet, P., 2021. Language conditioned imitation learning over unstructured data, in: Robotics: Science and Systems XVII, Virtual Event.

[83] Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y., Martín-Martín, R., 2021. What matters in learning from offline human demonstrations for robot manipulation. Proceedings of Machine Learning Research .

[84] Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R., 2019. OK-VQA: A visual question answering benchmark requiring external knowledge, in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3195–3204.

[85] Matuszek, C., Herbst, E., Zettlemoyer, L., Fox, D., 2013. Learning to parse natural language commands to a robot control system, in: International Symposium on Experimental Robotics, Springer. pp. 403–415.

[86] McCallum, S., 2023. Chatgpt banned in italy over privacy concerns. BBC News .

[87] Mei, H., Bansal, M., Walter, M., 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences, in: AAAI Conference on Artificial Intelligence, pp. 2772–2778.

[88] Meng, L., Wen, M., Yang, Y., Le, C., Li, X., Zhang, W., Wen, Y., Zhang, H., Wang, J., Xu, B., 2023. Offline pre-trained multi-agent decision transformer: One big sequence model tackles all SMAC tasks. Machine Intelligence Research .

[89] Mirowski, P., Grimes, M., Malinowski, M., Hermann, K.M., Anderson, K., Teplyashin, D., Simonyan, K., Zisserman, A., Hadsell, R., et al., 2018. Learning to navigate in cities without a map. Advances in Neural Information Processing Systems 31.

[90] Nachum, O., Gu, S.S., Lee, H., Levine, S., 2018. Data-efficient hierarchical reinforcement learning. Advances in Neural Information Processing Systems 31.

[91] Newell, A., Simon, H.A., 2007. Computer science as empirical inquiry: Symbols and search, in: ACM Turing Award Lectures, p. 1975.

[92] OpenAI, 2023. GPT-4 technical report. CoRR .

[93] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730–27744.

[94] Parisotto, E., Song, F., Rae, J., Pascanu, R., Gulcehre, C., Jayakumar, S., Jaderberg, M., Kaufman, R.L., Clark, A., Noury, S., et al., 2020. Stabilizing Transformers for reinforcement learning, in: The International Conference on Machine Learning, PMLR. pp. 7487–7498.

[95] Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior. arXiv preprint, arXiv:2304.03442 .

[96] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. FiLM: Visual reasoning with a general conditioning layer, in: AAAI Conference on Artificial Intelligence, pp. 3942–3951.

[97] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M., 2023. Communicative agents for software development. arXiv preprint, arXiv:2307.07924 .

[98] Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Huang, Y., Xiao, C., Han, C., et al., 2023. Tool learning with foundation models. arXiv preprint arXiv:2304.08354 .

[99] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: The International Conference on Machine Learning, PMLR. pp. 8748–8763.

[100] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

[101] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21, 5485–5551.

[102] Rakelly, K., Zhou, A., Finn, C., Levine, S., Quillen, D., 2019. Efficient off-policy meta-reinforcement learning via probabilistic context variables, in: The International Conference on Machine Learning, PMLR. pp. 5331–5340.

[103] Raman, S.S., Cohen, V., Rosen, E., Idrees, I., Paulius, D., Tellex, S., 2022. Planning with large language models via corrective re-prompting. arXiv preprint, arXiv:2211.09935 .

[104] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al., 2022. A generalist agent. Transactions on Machine Learning Research .

[105] Ribeiro, C., 2002. Reinforcement learning agents. Artificial Intelligence Review 17, 223–250.

[106] Roberts, A., Raffel, C., Shazeer, N., 2020. How much knowledge can you pack into the parameters of a language model?, in: The Conference on Empirical Methods in Natural Language Processing, pp. 5418–5426.

[107] Ryoo, M., Piergiovanni, A., Arnab, A., Dehghani, M., Angelova, A., 2021. TokenLearner: Adaptive space-time tokenization for videos. Advances in Neural Information Processing Systems 34, 12786–12797.

[108] Safavi, T., Koutra, D., 2021. Relational world knowledge representation in contextual language models: A review. Association for Computational Linguistics .

[109] Sajjadi, M.S., Duckworth, D., Mahendran, A., van Steenkiste, S., Pavetic, F., Lucic, M., Guibas, L.J., Greff, K., Kipf, T., 2022. Object scene representation transformer. Advances in Neural Information Processing Systems 35, 9512–9524.

[110] Sanderson, K., 2023. GPT-4 is here: what scientists think. Nature 615, 773.

[111] Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al., 2019. Habitat: A platform for embodied AI research, in: IEEE/CVF International Conference on Computer Vision, pp. 9339–9347.

[112] Sebastia, L., Onaindia, E., Marzal, E., 2006. Decomposition of planning problems. AI Communications 19, 49–81.

[113] Shafiullah, N.M., Cui, Z., Altanzaya, A.A., Pinto, L., 2022. Behavior transformers: Cloning *k* modes with one stone. Advances in Neural Information Processing Systems 35, 22955–22968.

[114] Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S., 2021a. ViNG: Learning open-world navigation with visual goals, in: The IEEE International Conference on Robotics and Automation, IEEE. pp. 13215–13222.

[115] Shah, D., Eysenbach, B., Rhinehart, N., Levine, S., 2021b. Rapid exploration for open-world navigation with latent goal models. The Conference on Robot Learning .

[116] Shah, D., Levine, S., 2022. ViKiNG: Vision-based kilometer-scale navigation with geographic hints. Robotics: Science and Systems XVIII .

[117] Shah, D., Osiński, B., Levine, S., et al., 2023a. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action, in: The Conference on Robot Learning, PMLR. pp. 492–504.

[118] Shah, D., Sridhar, A., Bhorkar, A., Hirose, N., Levine, S., 2023b. GNM: A general navigation model to drive any robot, in: IEEE International Conference on Robotics and Automation, IEEE. pp. 7226–7233.

[119] Shimizu, N., Haas, A., 2009. Learning to follow navigational route instructions, in: International Joint Conference on Artificial Intelligence, pp. 1488–1493.

[120] Song, H., Zhang, W.N., Hu, J., Liu, T., 2020. Generating persona consistent dialogues by exploiting natural language inference, in: The AAAI Conference on Artificial Intelligence, pp. 8878–8885.

[121] Squire, L.R., 1986. Mechanisms of memory. Science 232, 1612–1619.

[122] Sumers, T., Yao, S., Narasimhan, K., Griffiths, T.L., 2023. Cognitive architectures for language agents. arXiv preprint, arXiv:2309.02427 .

[123] Sun, J., Gan, W., Chen, Z., Li, J., Yu, P.S., 2022. Big data meets metaverse: A survey. arXiv preprint, arXiv:2210.16282 .

[124] Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR. pp. 6105–6114.

[125] Team, T., . The history, timeline, and future of LLMs. https://toloka.ai/blog/history-of-llms.

[126] Tellex, S., Gopalan, N., Kress-Gazit, H., Matuszek, C., 2020. Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems 3, 25–55.

[127] Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S., Roy, N., 2011. Understanding natural language commands for robotic navigation and mobile manipulation, in: AAAI Conference on Artificial Intelligence, pp. 1507–1514.

[128] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .

[129] Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F., 2021. Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems 34, 200–212.

[130] Unterrainer, J.M., Owen, A.M., 2006. Planning and problem solving: from neuropsychology to functional neuroimaging. Journal of Physiology-Paris 99, 308–317.

[131] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 30.

[132] Vezhnevets, A.S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., Kavukcuoglu, K., 2017. Feudal networks for hierarchical reinforcement learning, in: International Conference on Machine Learning, PMLR. pp. 3540–3549.

[133] Vulić, I., Ponti, E.M., Litschko, R., Glavaš, G., Korhonen, A., 2020. Probing pretrained language models for lexical semantics, in: The Conference on Empirical Methods in Natural Language Processing, pp. 7222–7240.

[134] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D., 2022. Self-consistency improves chain of thought reasoning in language models, in: The Eleventh International Conference on Learning Representations.

[135] Wason, P.C., 1968. Reasoning about a rule. Quarterly journal of experimental psychology 20, 273–281.

[136] Wason, P.C., Johnson-Laird, P.N., 1972. Psychology of reasoning: Structure and content. volume 86. Harvard University Press.

[137] Webb, T., Holyoak, K.J., Lu, H., 2023. Emergent analogical reasoning in large language models. Nature Human Behaviour , 1–16.

[138] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35, 24824–24837.

[139] Wilcox, A., Balakrishna, A., Dedieu, J., Benslimane, W., Brown, D., Goldberg, K., 2022. Monte carlo augmented actor-critic for sparse reward deep reinforcement learning from suboptimal demonstrations. Advances in Neural Information Processing Systems 35, 2254–2267.

[140] Wu, J., Gan, W., Chen, Z., Wan, S., Lin, H., 2023a. Ai-generated content (AIGC): A survey. arXiv preprint, arXiv:2304.06632 .

[141] Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S., 2023b. Multimodal large language models: A survey, in: IEEE International Conference on Big Data, IEEE. pp. 1–10.

[142] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al., 2023a. The rise and potential of large language model based agents: A survey. arXiv preprint,

arXiv:2309.07864 .

[143] Xi, Z., Jin, S., Zhou, Y., Zheng, R., Gao, S., Gui, T., Zhang, Q., Huang, X., 2023b. Self-Polish: Enhance reasoning in large language models via problem refinement. arXiv preprint, arXiv:2305.14497 .

[144] Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S., 2018. Gibson env: Real-world perception for embodied agents, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9068–9079.

[145] Xu, B., Peng, Z., Lei, B., Mukherjee, S., Liu, Y., Xu, D., 2023. ReWOO: Decoupling reasoning from observations for efficient augmented language models. arXiv preprint, arXiv:2305.18323 .

[146] Xu, F.F., Alon, U., Neubig, G., Hellendoorn, V.J., 2022. A systematic evaluation of large language models of code, in: The ACM SIGPLAN International Symposium on Machine Programming, pp. 1–10.

[147] Xun, G., Jia, X., Gopalakrishnan, V., Zhang, A., 2016. A survey on context learning. IEEE Transactions on Knowledge and Data Engineering 29, 38–56.

[148] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K., 2023. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint, arXiv:2305.10601 .

[149] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2023. A survey on multimodal large language models. arXiv preprint, arXiv:2306.13549 .

[150] Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 .

[151] Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y., 2021. MERLOT: Multimodal neural script knowledge models. Advances in Neural Information Processing Systems 34, 23634–23651.

[152] Zeng, F., Gan, W., Wang, Y., Yu, P.S., 2023. Distributed training of large language models, in: The 29th IEEE International Conference on Parallel and Distributed Systems, IEEE. pp. 1–8.

[153] Zhang, T., McCarthy, Z., Jow, O., Lee, D., Chen, X., Goldberg, K., Abbeel, P., 2018. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation, in: IEEE International Conference on Robotics and Automation, IEEE. pp. 5628–5635.

[154] Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al., 2023. Least-to-most prompting enables complex reasoning in large language models. The International Conference on Learning Representations .

[155] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J., 2020. Unified vision-language pre-training for image captioning and vqa, in: AAAI Conference on Artificial Intelligence, pp. 13041–13049.

[156] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M., 2023a. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint, arXiv:2304.10592 .

[157] Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A., 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning, in: The IEEE International Conference on Robotics and Automation, IEEE. pp. 3357–3364.

[158] Zhu, Z., Lin, K., Jain, A.K., Zhou, J., 2023b. Transfer learning in deep reinforcement learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 13344–13362.