SCIENCE CHINA
Information Sciences

• REVIEW •

February 2025, Vol. 68, Iss. 2, 121101:1–121101:44
https://doi.org/10.1007/s11432-024-4222-0

The rise and potential of large language model based agents: a survey

Zhiheng XI[1†], Wenxiang CHEN[1†], Xin GUO[1†], Wei HE[1†], Yiwen DING[1†], Boyang HONG[1†], Ming ZHANG[1†], Junzhe WANG[1†], Senjie JIN[1†], Enyu ZHOU[1†], Rui ZHENG[1], Xiaoran FAN[1], Xiao WANG[1], Limao XIONG[1], Yuhao ZHOU[1], Weiran WANG[2], Changhao JIANG[1], Yicheng ZOU[1], Xiangyang LIU[1], Zhangyue YIN[1], Shihan DOU[1], Rongxiang WENG[4], Wenjuan QIN[2], Yongyan ZHENG[2], Xipeng QIU[1], Xuanjing HUANG[1], Qi ZHANG[1*] & Tao GUI[3*]

[1]*School of Computer Science, Fudan University, Shanghai 200441, China*
[2]*College of Foreign Languages and Literature, Fudan University, Shanghai 200433, China*
[3]*Institute of Modern Languages and Linguistics, Fudan University, Shanghai 200433, China*
[4]*Meituan, Beijing 100102, China*

Received 7 September 2024/Revised 25 October 2024/Accepted 11 November 2024/Published online 17 January 2025

Abstract For a long time, researchers have sought artificial intelligence (AI) that matches or exceeds human intelligence. AI agents, which are artificial entities capable of sensing the environment, making decisions, and taking actions, are seen as a means to achieve this goal. Extensive efforts have been made to develop AI agents, with a primary focus on refining algorithms or training strategies to enhance specific skills or particular task performance. The field, however, lacks a sufficiently general and powerful model to serve as a foundation for building general agents adaptable to diverse scenarios. With their versatile capabilities, large language models (LLMs) pave a promising path for the development of general AI agents, and substantial progress has been made in the realm of LLM-based agents. In this article, we conduct a comprehensive survey on LLM-based agents, covering their construction frameworks, application scenarios, and the exploration of societies built upon LLM-based agents. We also conclude some potential future directions and open problems in this flourishing field.

Keywords natural language processing, large language models, LLM-based agents, AI agents, agent society

Citation Xi Z H, Chen W X, Guo X, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci, 2025, 68(2): 121101, https://doi.org/10.1007/s11432-024-4222-0

1 Introduction

One core research area in artificial intelligence (AI) is to develop agents that possess human-level intelligence, and potentially even exceed it [1]. The concept of agent originates in philosophy, where it describes entities possessing desires, beliefs, intentions, and the ability to take actions [2]. In AI research, the term agent refers to an artificial entity capable of perceiving its surroundings using sensors, making decisions, and then taking actions in response using actuators [1,3]. The development and advancement of agents have been central within the AI community for a long time [1,4]. Moreover, AI agents are now recognized as a pivotal stride towards achieving artificial general intelligence (AGI)[1)], as they encompass the potential for a wide range of intelligent activities [3,5,6].

From the mid-20th century, significant strides were made in designing and developing AI agents [7–12]. However, these efforts have predominantly focused on enhancing specific skills (such as symbolic reasoning or quick responsiveness) or mastering particular tasks (such as Go or Chess) [13–15]. Achieving broad adaptability across varied scenarios remained elusive. Moreover, previous studies mostly emphasized the design of algorithms and training strategies, overlooking the development of the model's inherent general

* Corresponding author (email: qz@fudan.edu.cn, tgui@fudan.edu.cn)
† These authors contributed equally to this work.
1) Also known as Strong AI.

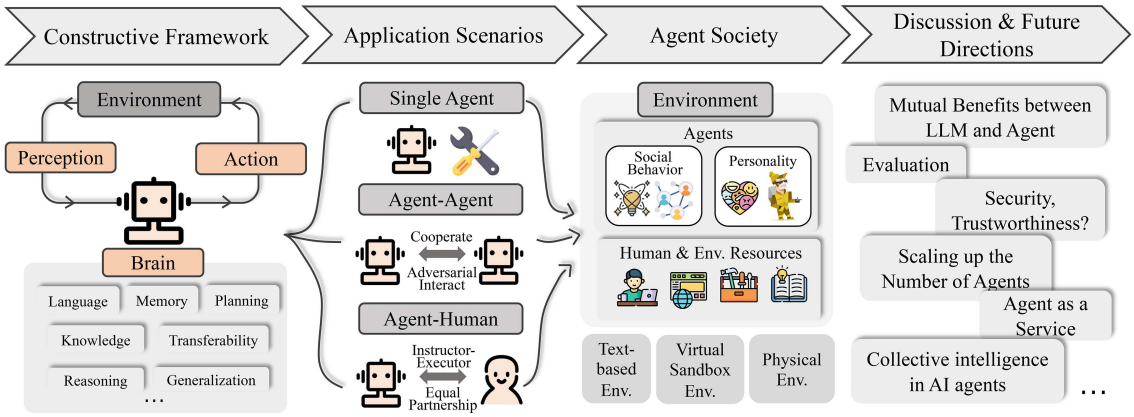

Figure 1 (Color online) Roadmap of this survey.

abilities [16, 17]. Actually, enhancing these inherent abilities of the model is pivotal for advancing the agent, and the field is in need of a powerful and versatile model to serve as the foundation for agent systems.

The emergence of large language models (LLMs) has brought a glimmer of hope for the further development of agents [18–20], and significant progress has been made by researchers in the filed [16, 21–23]. This is attributed to the remarkable capabilities demonstrated by LLMs, including natural language interaction, knowledge acquisition, instruction following, generalization, reasoning, planning, and tool-using. These advantages have earned LLMs the designation of sparks for AGI [24], making them highly desirable for building general agents [16]. And the area of LLM-based agents has emerged as one of the most promising fields in AI research. The academic and industrial communities have invested significant research and development efforts in this field. However, despite the progress, this emerging field still faces numerous crucial issues that require further exploration and resolution, e.g., the architectural design of LLM-based agents and their application domains. Therefore, it is necessary to review the origins and development of this field, summarize current research achievements, and look ahead to provide a clearer understanding of LLM-based agents and deeper insights into their future development.

In this article, we present a comprehensive and systematic survey of LLM-based agents, attempting to explore several critical research problems within the field and prospective avenues in this burgeoning field (Figure 1 demonstrates the roadmap of this survey). Before delving into the research problems, we first discuss the background information (Section 2), including the origin, definition, and development of AI agents, as well as why LLMs are suitable as the foundation for AI agents. The first research problem is the design and construction of LLM-based agents (Section 3). Based on the definition of AI agents, we propose a conceptual framework with three key components: brain, perception, and action. Different downstream applications can tailor this framework according to their specific requirements. The second research problem is the application scenarios of LLM-based agents (Section 4). We discuss the various paradigms such as single agent, multi-agent, and human-in-the-loop, and their respective applicable scenarios. The third research problem is agent society (Section 5). We discuss the components and operational mechanisms of an agent society, as well as its insights into human society. Finally, we discuss a series of open problems and future directions in this field (Section 6), e.g., the mutual benefits between LLM research and agent research, evaluation for LLM-based agents, potential risks of LLM-based agents, and scaling up the number of agents. Finally, we present an overview and summary of the entire article (Section 7).

2 Background

In this section, we provide crucial background information to lay the groundwork for the subsequent content. We first discuss the origin of AI agents in philosophy and its definition (Subsection 2.1). Subsequently, we review the development of AI agents through the lens of technological trends (Subsection 2.2). Finally, we introduce the key characteristics of agents and demonstrate why LLMs are suitable to serve as the foundation of AI agents (Subsection 2.3).

2.1 Origin of AI agent

"Agent" is a concept with a long history that spans across various fields. It is derived from a philosophical origin, triggering a discussion on whether artificial products can possess agency in a philosophical sense. Then, related concepts were introduced into the field of AI, forming the basis of AI agents.

Agent in philosophy. The initial idea of an "agent" has a philosophical origin, with influential thinkers like Aristotle and Hume, contributing to its conceptualization [2]. In a general sense, an "agent" is an entity with the capacity to act, and the term "agency" denotes the exercise or manifestation of this capacity [2]. In a narrow sense, "agency" is usually used to refer to the performance of intentional actions; and correspondingly, "agent" denotes entities that possess desires, beliefs, intentions, and the ability to act [25–28]. Importantly, agents encompass not only individual human beings but also other entities in both the physical and virtual world.

Introduction of agents into AI. In AI, an agent is defined as an artificial entity capable of perceiving its environment, making decisions, and taking actions using sensors and actuators [1, 3]. As Wooldridge et al. [3] stated that we can define AI by saying that it is a subfield of computer science that aims to design and build computer-based agents that exhibit aspects of intelligent behavior. So we can treat "agent" as a central concept in AI. When the concept of agent is introduced into AI, its meaning evolves. In philosophy, an agent can be a human, an animal, or even a concept or entity with autonomy [2]. However, in the field of AI, an agent is a computational entity [3, 29]. Since it is difficult to determine if they possess internal desires or consciousness, many AI researchers, including Alan Turing, suggest temporarily setting aside the question of whether an agent is "actually" thinking or literally possesses a "mind" [30]. Instead, researchers employ other attributes to help describe an agent, such as properties of autonomy, reactivity, pro-activeness, and social ability [3, 31].

It might come as a surprise that researchers within the mainstream AI community devoted relatively minimal attention to concepts related to agents until the mid to late 1980s. Nevertheless, there has been a significant surge of interest in this topic within the realms of computer science and artificial intelligence communities since then [32–35].

2.2 Technological trends in agent research

The evolution of AI agents has undergone several main stages, and here we take the lens of technological trends to review its development briefly.

Symbolic agents. In the early stages of AI research, the predominant approach utilized is symbolic AI, characterized by its reliance on symbolic logic [36, 37]. This approach employs logical rules and symbolic representations to encapsulate knowledge and facilitate reasoning processes. Early AI agents are built based on this approach [38], and they primarily focused on two problems: the transduction problem and the representation/reasoning problem [39]. These agents aim to emulate human thinking patterns. They possess explicit and interpretable reasoning frameworks, and due to their symbolic nature, they exhibit a high degree of expressive capability [7, 8, 40]. A classic example of this approach is knowledge-based expert systems. However, symbolic agents face limitations in handling uncertainty and large-scale real-world problems [13, 14]. Additionally, due to the intricacies of symbolic reasoning algorithms, it is challenging to find an efficient algorithm capable of producing meaningful results within a finite timeframe [14, 41].

Reactive agents. Different from symbolic agents, reactive agents do not use complex symbolic reasoning. Instead, they primarily focus on the interaction between the agent and its environment, emphasizing quick and real-time responses [9, 10, 14, 42, 43]. These agents are mainly based on a sense-act loop, efficiently perceiving and reacting to the environment. The design of such agents prioritizes direct input-output mappings rather than intricate reasoning and symbolic operations [34]. However, reactive agents also have limitations. They typically require fewer computational resources, enabling quicker responses, but they might lack complex higher-level decision-making and planning capabilities.

Reinforcement learning-based agents. With the improvement of computational capabilities and data availability, along with a growing interest in simulating interactions between intelligent agents and their environments, researchers have begun to utilize reinforcement learning methods to train agents for tackling more challenging and complex tasks [11, 12, 44, 45]. The primary concern in this field is how to enable agents to learn through interactions with their environments, enabling them to achieve maximum cumulative rewards in specific tasks [15]. Initially, reinforcement learning (RL) agents are primarily based on fundamental techniques such as policy search and value function optimization, exemplified by

Q-learning [46] and SARSA [47]. With the rise of deep learning, the integration of deep neural networks and reinforcement learning, known as deep reinforcement learning (DRL), has emerged [48, 49]. This allows agents to learn intricate policies from high-dimensional inputs, leading to numerous significant accomplishments like AlphaGo [50] and DQN [51]. The advantage of this approach lies in its capacity to enable agents to autonomously learn in unknown environments, without explicit human intervention. This allows for its wide application in an array of domains, from gaming to robot control and beyond. Nonetheless, reinforcement learning faces challenges including long training times, low sample efficiency, and stability concerns, particularly when applied in complex real-world environments [15].

Agents with transfer learning and meta learning. Traditionally, training a reinforcement learning agent requires huge sample sizes and long training time, and lacks generalization capability [52–56]. Consequently, researchers introduce transfer learning to expedite an agent's learning on new tasks [57–59]. Transfer learning reduces the burden of training on new tasks and facilitates the sharing and migration of knowledge across different tasks, thereby enhancing learning efficiency, performance, and generalization capabilities. Furthermore, meta learning has also been introduced to develop agents [60–64]. Meta learning focuses on learning how to learn, enabling an agent to swiftly infer optimal policies for new tasks from a small number of samples [65]. Such an agent, when confronted with a new task, can rapidly adjust its learning approach by leveraging acquired general knowledge and policies, consequently reducing the reliance on a large volume of samples. However, when there exist significant disparities between source and target tasks, the effectiveness of transfer learning might fall short of expectations and negative transfer might occur [66, 67]. Additionally, the substantial amount of pre-training and large sample sizes required by meta learning make it hard to establish a universal learning policy [61, 68].

Large language model-based agents. As LLMs have demonstrated impressive emergent capabilities and have gained immense popularity [18–20, 69], researchers have started to leverage these models to construct AI agents, known as LLM-based agents [16, 21, 22, 70]. Specifically, they employ LLMs as the core component of the brain or controller of these agents and expand their perceptual and action space through strategies such as multimodal perception and tool utilization [71–75]. These LLM-based agents can exhibit reasoning and planning abilities comparable to symbolic agents through techniques like chain-of-thought (CoT) and problem decomposition [76–82]. They can also acquire interactive capabilities with the environment, akin to reactive agents, by learning from feedback and performing new actions [83–85]. Similarly, LLMs undergo pre-training on large-scale corpora and demonstrate the capacity for few-shot and zero-shot generalization, allowing for seamless transfer between tasks without the need to update parameters [69, 86–88]. LLM-based agents have been applied to various real-world scenarios, such as software development [89, 90] and scientific research [91]. Due to their natural language comprehension and generation capabilities, they can interact with each other seamlessly, giving rise to collaboration and competition among multiple agents [89, 90, 92, 93]. Furthermore, some studies suggested that allowing multiple agents to coexist can lead to the emergence of social phenomena [16].

2.3 Why is LLM suitable as the foundation of agent?

As mentioned before, researchers have introduced several properties to help describe and define agents in the field of AI. In this section, we will discuss the key properties, elucidate their relevance to LLMs, and thereby expound on why LLMs are highly suited to serve as the foundation of AI agents.

Autonomy. Autonomy means that an agent operates without direct intervention from humans or others and possesses a degree of control over its actions and internal states [3, 94]. This implies that an agent should not only possess the capability to follow explicit human instructions for task completion but also exhibit the capacity to initiate and execute actions independently. LLMs can demonstrate a form of autonomy through their ability to generate human-like text, engage in conversations, and perform various tasks without detailed step-by-step instructions [95, 96]. Moreover, they can dynamically adjust their outputs based on environmental input, reflecting a degree of adaptive autonomy [17, 21, 85]. Furthermore, they can showcase autonomy through exhibiting creativity like coming up with novel ideas, stories, or solutions that have not been explicitly programmed into them [97, 98]. This implies a certain level of self-directed exploration and decision-making. Applications like Auto-GPT [95] exemplify the significant potential of LLMs in constructing autonomous agents. Simply by providing them with a task and a set of available tools, they can autonomously formulate plans and execute them to achieve the ultimate goal.

Reactivity. Reactivity in an agent refers to its ability to respond rapidly to immediate changes and stimuli in its environment [31]. This implies that the agent can perceive alterations in its surroundings and promptly take appropriate actions. Traditionally, the perceptual space of language models has been confined to textual inputs, while the action space has been limited to textual outputs. Recently, advances in multimodal fusion techniques expand their perceptual space to rapidly process visual and auditory information from the environments [19, 99, 100]. Similarly, it is also feasible to expand the action space of LLMs through embodiment techniques [101, 102] and tool usage [73, 75]. These advancements enable LLMs to effectively interact with the real-world physical environment and carry out tasks within it. One major challenge is that LLM-based agents, when performing non-textual actions, require an intermediate step of generating thoughts or formulating tool usage in the textual form before eventually translating them into concrete actions. This intermediary process consumes time and reduces the response speed. However, this is analogous to human behavioral patterns, where the principle of "think before you act" is observed [103, 104].

Pro-activeness. Pro-activeness denotes that agents do not merely react to their environments; they possess the capacity to display goal-oriented actions by proactively taking the initiative [31]. This property emphasizes that agents can reason, make plans, and take proactive measures in their actions to achieve specific goals or adapt to environmental changes. Although intuitively the paradigm of next token prediction in LLMs may not possess intention or desire, research has shown that they can implicitly generate representations of these states and guide the model's inference process [105–107]. LLMs have demonstrated a strong capacity for generalized reasoning and planning. By prompting LLMs with instructions like "let's think step by step", we can elicit their reasoning abilities, such as logical and mathematical reasoning [76–78]. Similarly, LLMs have shown the emergent ability of planning in forms of goal reformulation [80, 108], task decomposition [79, 109], and adjusting plans in response to environmental changes [81, 110].

Social ability. Social ability refers to an agent's capacity to interact with other agents, including humans, using agent-communication languages [111]. Large language models exhibit strong natural language interaction abilities like comprehension and generation [17, 112, 113]. Compared to structured languages or other communication protocols, such capability enables them to interact with other models or humans in an interpretable manner, laying the foundation for the social ability for LLM-based agents [16, 89]. Many researchers have demonstrated that LLM-based agents can enhance task performance through social behaviors such as collaboration and competition [89, 92, 114, 115]. By inputting specific prompts, LLMs can also play different roles, thereby simulating the social division of labor in the real world [90]. Furthermore, when we place multiple agents with distinct identities into a society, emergent social phenomena can be observed [16].

3 Birth of an agent: construction of LLM-based agents

The principle of "Survival of the Fittest" [116] highlights the need for cognitive abilities to adapt to the external environment and respond to changes for individual survival. Inspired by this and the definition of AI agents, we present a general conceptual framework of an LLM-based agent composed of three key components: brain, perception, and action (see Figure 2). We first describe the structure and working mechanism of the "brain", which is the cognitive core of an AI agent (Subsection 3.1). It not only stores knowledge and memories but also undertakes indispensable functions like decision-making, exhibiting the intelligence of an agent. Next, we introduce the perception module (Subsection 3.2). Its core purpose is to broaden the agent's perception space from a text-only domain to a multimodal sphere, which equips the agent to grasp and utilize information from its surroundings more effectively. Finally, we present the action module designed to expand the action space of an agent (Subsection 3.3). Specifically, the agent is empowered to adapt to environmental changes, provide feedbacks, and even influence and mold the environment.

The framework can be tailored for different application scenarios; i.e., not every specific component will be used in all studies. In general, agents operate in the following workflow: First, the perception module, corresponding to human sensory systems, perceives changes in the external environment and then converts multimodal information into an understandable representation for the agent. Subsequently, the brain module, serving as the control center, engages in activities such as thinking, decision-making, and operations with storage including memory and knowledge. Finally, the action module, corresponding

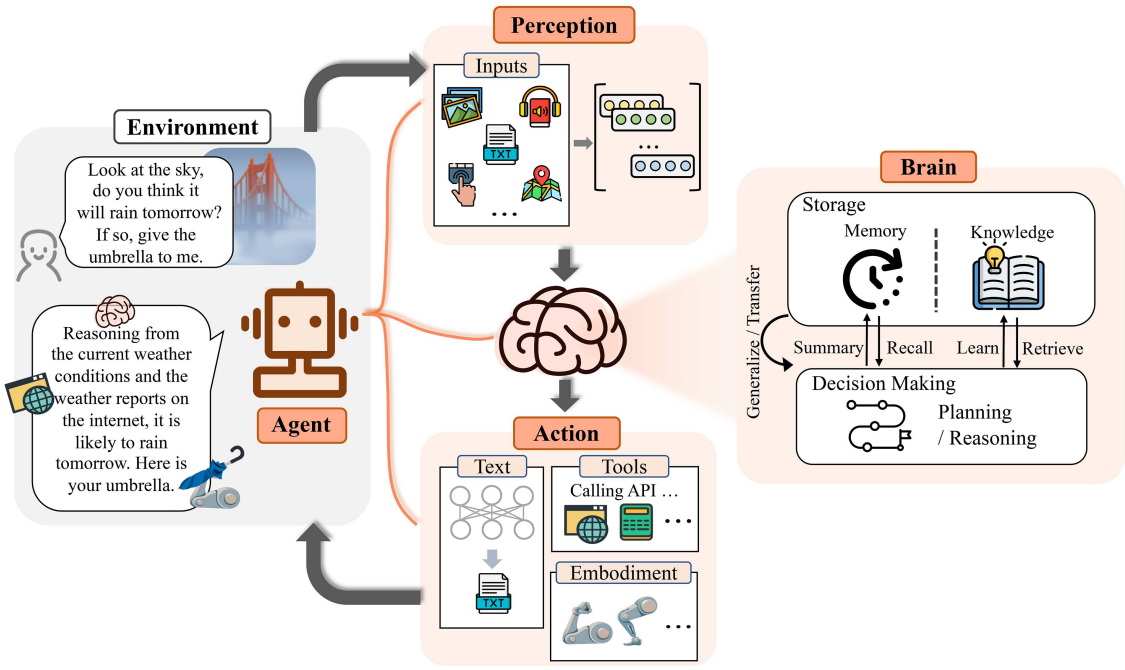

Figure 2 (Color online) Conceptual framework of an LLM-based agent with three components: brain, perception, and action. Serving as the controller, the brain module undertakes basic tasks like memorizing, reasoning, and planning. The perception module perceives and processes multimodal information from the external environment, and the action module carries out the execution and interacts with the surroundings. Here is an example that illustrates the workflow: When a human asks whether it will rain, the perception module converts the instruction into a specific representation. Then the brain module begins to reason according to the current weather and the weather reports on the internet. Finally, the action module responds and hands the umbrella to the human. By repeating the above process, the LLM-based agent can continuously get feedback and interact with the environment.

to human limbs, carries out the execution and leaves an impact on the surroundings. By repeating the above process, an agent can continuously get feedback and interact with the environment.

3.1 Brain

The human brain is a sophisticated structure comprised of a vast number of interconnected neurons, capable of processing complex information, generating diverse thoughts, controlling behaviors, and even creating art and culture [117]. Much like humans, the brain component serves as the central controller of an AI agent. In our framework, the brain module is primarily composed of an LLM.

After receiving the information processed by the perception module, the brain module first turns to storage, retrieving in knowledge (Subsection 3.1.1) and recalling from memory (Subsection 3.1.2). These outcomes aid the agent in planing, reasoning, and making decisions (Subsection 3.1.3). Additionally, the brain module may memorize the agent's past observations, thoughts, and actions in the form of summaries, vectors, or other data structures. Meanwhile, it can also update the knowledge such as common sense and domain knowledge for future use. The LLM-based agent may also adapt to unfamiliar scenarios with its inherent generalization and transferability (Subsection 3.1.4). In the subsequent sections, we delve into a detailed exploration of these extraordinary facets of the brain module.

3.1.1 *Knowledge*

LLMs are typically imbued with large-scale data through semi-supervised learning during the pre-training phase [118,119]. They can encode a wide range of knowledge into their parameters and respond correctly to various types of queries after pre-training [120]. Furthermore, the knowledge can assist LLM-based agents in making informed decisions [121]. Such knowledge can be broadly categorized into the following types.

• **Commonsense knowledge.** Commonsense knowledge [122–124] refers to general world facts that are typically taught to most individuals at an early age. For example, people commonly know that medicine is used for curing diseases, and umbrellas are used to protect against rain. Such information is

usually not explicitly mentioned in the context. Therefore, models lacking the corresponding commonsense knowledge may fail to grasp or misinterpret the intended meaning [125]. Similarly, agents without commonsense knowledge may make incorrect decisions, such as not bringing an umbrella when it rains heavily.

• **Professional domain knowledge.** Professional domain knowledge refers to the knowledge associated with a specific domain like programming [124, 126, 127], mathematics [128], medicine [129], etc. It is essential for models to effectively solve problems within a particular domain [130]. For example, models designed to perform programming tasks need to possess programming knowledge. This kind of knowledge is very important for domain-specific agents because only with this knowledge can they make reasonable decisions.

3.1.2 *Memory*

In our framework, "memory" stores the agent's past observations, thoughts, and actions [131]. This memory enables agents to use past experiences for strategy formulation and decision-making, similar to how humans do [132–134]. It assists agents in handling complex problems by allowing them to revisit past strategies and adapt to new environments.

However, as interactions in LLM-based agents increase, two main issues emerge. First, the length of historical data can exceed the processing capacity of Transformer architecture-based LLM agents, leading to possible data truncation. Second, with the accumulation of extensive data, it becomes increasingly difficult for agents to retrieve and connect relevant memories, risking misaligned responses in ongoing contexts. Therefore, appropriate memory compression and retrieval techniques are key to enhancing an agent's memory capabilities.

Methods for memory compression. We delve into two primary methods designed to compress memory, ensuring efficient recall and analysis.

• **Summarizing memory.** The first strategy for improving memory efficiency is memory summarization. This method summarizes historical interactions and stores them in natural language. Various techniques have been proposed for memory summarization. Using prompts, some methods succinctly integrate memories [135], while others emphasize reflective processes to create condensed memory representations [16, 136]. Hierarchical methods streamline dialogues into both daily snapshots and overarching summaries [137]. Notably, specific strategies translate environmental feedback into textual encapsulations, bolstering agents' contextual grasp for future engagements [138].

• **Compressing memories with vectors or data structures.** Using appropriate data structures, LLM-based agents can store historical interactions more efficiently. Notably, several methodologies lean on embedding vectors for memory sections, plans, or dialogue histories [90, 137, 139, 140]. Another approach translates sentences into triplet configurations [141], while some perceive memory as a unique data object, fostering varied interactions [142]. Furthermore, ChatDB [143] and DB-GPT [144] opt for SQL databases, which allows for data manipulation via SQL queries.

Methods for memory retrieval. When an agent interacts with its environment, it is imperative to retrieve the most appropriate content from its memory. This ensures that the agent accesses relevant and accurate information to make decision [137, 140]. A representative approach in automated retrieval considers three metrics: recency, relevance, and importance. The memory score is determined as a weighted combination of these metrics, with memories having the highest scores being prioritized in the model's context [16]. Some work introduces the concept of interactive memory objects, representing dialogue history that can be moved, edited, deleted, or combined through summarization [142]. Users can view and manipulate these objects, influencing how the agent perceives the dialogue. Similarly, other studies allow for memory operations like deletion based on specific commands provided by users [143]. Such methods ensure that the memory content aligns closely with user expectations.

3.1.3 *Reasoning and planning*

Reasoning. Reasoning, crucial in human intellectual activities for problem-solving, decision-making, and critical analysis, is underpinned by evidence and logic [145–147]. Deductive, inductive, and abductive reasoning are key forms recognized in these endeavors [148]. For LLM-based agents, like humans, reasoning capacity is crucial for addressing complex tasks [19].

For LLMs, the reasoning ability is considered an emergent ability, i.e., such an ability emerges once the language model reaches a certain scale in size [20, 149]. Notably, the CoT method [76, 77] has been

demonstrated to elicit the reasoning capacities of LLMs by guiding them to generate rationales before outputting answers. Other techniques like self-consistency [78], self-polish [80], self-refine [150], and selection-inference [151] also enhance LLM performance. To further improve the reasoning ability of LLMs or LLM-based agents, current research primarily explores supervised fine-tuning (SFT) methods [152] and interactive training approaches [153, 154].

Planning. Planning is a key strategy employed by humans when facing complex tasks. For humans, planning helps organize thoughts, set objectives, and determine the steps to achieve those objectives [155–157]. Similar to humans, the ability to plan is crucial for agents, and central to this planning module is the capacity for reasoning [158–160]. This offers a structured thought process for agents based on LLMs. Through reasoning, agents break down complex tasks into manageable sub-tasks, devising appropriate plans for each [161, 162]. Moreover, as tasks progress, agents can employ introspection to modify their plans, ensuring they align better with real-world circumstances, facilitating adaptive and successful task execution. Typically, planning comprises two stages: plan formulation and plan reflection.

• **Plan formulation.** During plan formulation, agents generally decompose an overarching task into numerous sub-tasks, and various approaches have been proposed in this phase. Notably, some studies advocated for LLM-based agents to decompose problems comprehensively in a single step, formulating a complete plan at once and then executing it sequentially [79, 163–165]. In contrast, other studies like the CoT-series employ an adaptive strategy, where they plan and address sub-tasks one at a time, allowing for more fluidity in handling intricate tasks in their entirety [76, 77, 166]. Additionally, some methods emphasize hierarchical planning [167, 168], while others underscore a strategy in which final plans are derived from reasoning steps structured in a tree-like format. The latter approach argues that agents should assess all possible paths before finalizing a plan [78, 169–171]. While LLM-based agents demonstrate a broad scope of general knowledge, they can occasionally face challenges when tasked with situations that require expertise knowledge. Enhancing these agents by integrating them with planners of specific domains has shown to yield better performance [109, 115, 172, 173].

• **Plan reflection.** After formulating a plan, it is imperative to reflect upon and evaluate its merits. LLM-based agents employ internal feedback mechanisms, often drawing insights from pre-existing models, to hone and enhance their strategies and planning approaches [138, 150, 174, 175]. To better align with human values and preferences, agents actively engage with humans, allowing them to rectify misunderstandings and assimilate this tailored feedback into their planning methodology [89, 176, 177]. Furthermore, they could gather feedback from tangible or virtual surroundings, such as cues from task accomplishments or post-action observations, aiding them in revising and refining their plans [72, 82, 178–180].

3.1.4 *Transferability and generalization*

The remarkable nature of the human brain is largely attributed to its high degree of plasticity and adaptability. It can continuously adjust its structure in response to external stimuli and internal needs, thereby adapting to different environments and tasks. These years, plenty of research indicates that pre-trained models can learn universal language representations [181–183], and with only a small amount of data for fine-tuning, it can demonstrate excellent performance in downstream tasks [184]. There is no need to train new models from scratch, which saves a lot of computation resources. However, models trained through this task-specific fine-tuning, usually lack versatility and generalizability to other tasks. Instead of merely functioning as a static knowledge repository, LLM-based agents exhibit dynamic learning ability which enables them to adapt to novel tasks swiftly and robustly [18, 86, 87].

Unseen task generalization. Studies show that instruction-tuned LLMs exhibit zero-shot generalization without the need for task-specific fine-tuning [18, 19, 86–88]. Notably, LLMs can complete unfamiliar tasks by following the instructions based on their own understanding. One of the implementations is multi-task learning, for example, FLAN [86] finetunes language models on a collection of tasks described via instructions, and T0 [87] introduces a unified framework that converts every language problem into a text-to-text format. Despite being purely a language model, GPT-4 [19] demonstrates remarkable capabilities in a variety of domains and tasks, including abstraction, coding, mathematics, medicine, law, and others [24]. Promisingly, such generalization capability can be further enhanced by scaling up the model size and the quantity or diversity of training instructions [75, 185].

In-context learning. Numerous studies indicate that LLMs can perform a variety of complex tasks through in-context learning (ICL), which involves the models' ability to learn from a few examples within a given context [186]. Few-shot in-context learning enhances the performance by concatenating the

original input with several complete examples as prompts to enrich the context [69]. The key idea of ICL is learning from analogy, akin to the learning process of humans [187]. Besides, since the prompts are written in natural language, the interaction is interpretable and changeable, making it easier to incorporate human knowledge [76, 188]. Unlike the supervised learning process, ICL does not involve fine-tuning or parameter updates, greatly reducing computation costs for adapting the models to new tasks. Beyond texts, researchers also explore the potential ICL capabilities in different multimodal tasks [189–194], making it possible for agents to be applied to large-scale real-world tasks.

3.2 Perception

Both humans and animals rely on sensory organs like eyes and ears to gather information from their surroundings. These perceptual inputs are converted into neural signals and sent to the brain for processing [195, 196], allowing humans to perceive and interact with the world. Similarly, it is crucial for LLM-based agents to receive information from various sources and modalities. This expanded perceptual space helps agents better understand their environment, make informed decisions, and excel in a broader range of tasks, making it an essential development direction. Agent handles this information to the Brain module for processing through the perception module.

In this section, we introduce how to enable LLM-based agents to acquire multimodal perception capabilities, encompassing textual (Subsection 3.2.1), visual (Subsection 3.2.2), and auditory inputs (Subsection 3.2.3). We also consider other potential input forms (Subsection 3.2.4) such as tactile feedback, gestures, and 3D maps to enrich the agent's perception domain and enhance its versatility.

3.2.1 *Textual input*

Text serves as a conduit for conveying data, information, and knowledge, making text communication one of the most important ways through which humans interact with the world. LLM-based agents already have the fundamental ability to communicate with humans through textual input and output [95]. In a user's textual input, aside from the explicit content, lie concealed beliefs, desires, and intentions. Understanding implied meanings is crucial for the agent to grasp the potential and underlying intentions of human users, thereby enhancing its communication efficiency and quality with users. However, understanding implied meanings within textual input remains challenging for current LLM-based agents [75, 197]. For example, some studies [113, 198–200] employed reinforcement learning to perceive implied meanings and models feedback to derive rewards. This helps deduce the speaker's preferences, leading to more personalized and accurate responses from the agent. Additionally, as the agent is designed for use in complex real-world situations, it will inevitably encounter many entirely novel tasks. Understanding text instructions for unknown tasks places higher demands on the agent's text perception abilities. As described in Subsection 3.1.4, an LLM that has undergone instruction tuning [86] can exhibit remarkable zero-shot instruction understanding and generalization abilities, eliminating the need for task-specific fine-tuning.

3.2.2 *Visual input*

Although LLMs excel in language comprehension [19, 201] and multi-turn conversations [202], they inherently lack visual perception and can only understand discrete textual content. Visual input usually contains a wealth of information about the world, including properties of objects, spatial relationships, scene layouts, and more in the agent's surroundings. Therefore, integrating visual information with data from other modalities can offer the agent a broader context and a more precise understanding [101], deepening the agent's perception of the environment.

To help the agent understand the information contained within images, a straightforward approach is to generate corresponding text descriptions for image inputs, known as image captioning [203–207]. Captions can be directly linked with standard text instructions and fed into LLM-based agents. This approach is highly interpretable and does not require additional training for caption generation, which can save a significant number of computational resources. However, caption generation is a low-bandwidth method [101, 208], and it may lose a lot of potential information during the conversion process. Furthermore, the agent's focus on images may introduce biases.

The second approach to endowing an LLM-based agent with visual understanding capabilities is modality fusion, where a visual encoder is integrated with an LLM at the embedding level [209, 210]. Freezing

one or both of them during training is a widely adopted paradigm that achieves a balance between training resources and model performance [211]. However, LLMs cannot directly understand the output of a visual encoder, so it is necessary to convert the image encoding into embeddings that LLMs can comprehend. In other words, it involves aligning the visual encoder with the LLM. This usually requires adding an extra learnable interface layer between them. For example, BLIP-2 [211] and InstructBLIP [212] use the querying transformer (Q-Former) module as an intermediate layer between the visual encoder and the LLM [212]. At the same time, some researchers adopt a computationally efficient method using a single projection layer to achieve visual-text alignment, reducing the need for training additional parameters [99,210,213]. Moreover, the projection layer can effectively integrate with the learnable interface to adapt the dimensions of its outputs, making them compatible with LLMs [214–217].

3.2.3 *Auditory input*

Auditory information constitutes a crucial component of world information. When an agent possesses auditory capabilities, it can improve its awareness of interactive content, the surrounding environment, and even potential dangers. While there are numerous well-established models and approaches [218–220] for processing audio as a standalone modality, these models often only excel at specific tasks. Given the excellent tool-using capabilities of LLMs (which will be discussed in detail in Subsection 3.3), a very intuitive idea is that the agent can use LLMs as control hubs, invoking existing toolsets or model repositories in a cascading manner to perceive audio information.

Furthermore, an audio spectrogram can serve as a medium to endow LLM-based agents with auditory capabilities, as it offers a clear representation of how the frequency spectrum of an audio signal evolves over time [221]. For a segment of audio data over a period of time, it can be abstracted into a finite-length audio spectrogram. An audio spectrogram has a 2D representation, which can be visualized as a flat image. Hence, some research [222,223] efforts aim to migrate perceptual methods from the visual domain to audio. Audio spectrogram transformer (AST) [222] employs a Transformer architecture similar to ViT to process audio spectrogram images. By segmenting the audio spectrogram into patches, it achieves effective encoding of audio information.

3.2.4 *Other input*

LLM-based agents are expected to be equipped with richer perception modules in the future. They could perceive and understand diverse modalities in the real world, much like humans. For example, agents could have unique touch and smell organs, allowing them to gather more detailed information when interacting with objects. At the same time, agents can also be equipped with hardware devices, such as GPS, Lidar, and other sensors, having a clearer sense of the spatial position, temperature, and brightness and taking environment-aware actions.

3.3 Action

After humans perceive their environment, their brains integrate, analyze, and reason to make decisions. Subsequently, they employ their nervous systems to control their bodies, enabling adaptive or creative actions in response to the environment, such as engaging in conversation, evading obstacles, or starting a fire. When an agent possesses a brain-like structure with capabilities of knowledge, memory, reasoning, planning, and generalization, as well as multimodal perception, it is also expected to possess a diverse range of actions akin to humans to respond to its surrounding environment. In constructing such an agent, the action module receives action sequences sent by the brain module and carries out actions to interact with the environment.

This section begins with textual output (Subsection 3.3.1), which is the inherent capability of LLM-based agents. Next we talk about the tool-using capability of LLM-based agents (Subsection 3.3.2), which has proved effective in enhancing their versatility and expertise. Finally, we discuss equipping the LLM-based agent with embodied action to facilitate its grounding in the physical world (Subsection 3.3.3).

3.3.1 *Textual output*

Recently, the rise and development of Transformer-based generative LLMs have endowed LLM-based agents with inherent language generation capabilities [19, 224]. The text quality they generate excels in

various aspects such as fluency, relevance, diversity, controllability [112, 225–227]. Consequently, LLM-based agents can be exceptionally strong language generators.

3.3.2 *Tool using*

Tools are extensions of the capabilities of humans. When faced with complex tasks, humans employ tools to simplify task-solving and enhance efficiency, freeing time, and resources. Similarly, agents have the potential to accomplish complex tasks more efficiently and with higher quality if they also learn to use and utilize tools [75]. LLM-based agents have limitations because of hallucination [228], the lack of training data and tuning for specific fields [229], untransparent decision-making process [230] and susceptibility to adversarial attacks [231]. Fortunately, specialized tools enable LLM-based agents to enhance their expertise, adapt domain knowledge, and be more suitable for domain-specific needs in a pluggable form. They also exhibit stronger interpretability and robustness. LLM-based agents not only require the use of tools, but are also well-suited for tool integration. LLMs show remarkable reasoning and decision-making abilities in complex interactive environments [78] and significant potential in intent understanding and other aspects [19, 232–234]. These make LLM-based agents proficient tool users.

Learning to use tools. Leveraging the powerful zero-shot and few-shot learning abilities of LLMs [69, 235], agents can acquire knowledge about tools by utilizing zero-shot prompts that contain descriptions of tool functionalities and parameters, or few-shot prompts that provide demonstrations of specific tool usage scenarios and corresponding methods [73, 236]. These learning approaches parallel human methods of learning by consulting tool manuals or observing others using tools [75]. Besides configurations and demonstrations, agents can also learn from feedback received from both the environment and humans [18, 237, 238]. Environmental feedback encompasses result feedback on whether actions have successfully completed the task and intermediate feedback that captures changes in the environmental state caused by actions; human feedback comprises explicit evaluations and implicit behaviors, such as clicking on links [75].

It is crucial to improve the agent's generalization ability in tool usage to handle complicated scenarios as well as continuously evolving new tools. To accomplish this, agents need to grasp the common principles or patterns in tool usage strategies, which can potentially be achieved through meta-tool learning [239]. In addition, techniques like curriculum learning [240] can enhance the agent's understanding of relationships between simple and complex tools, such as how complex tools are built on simpler ones and allow agents to effectively discern nuances across various application scenarios and transfer previously learned knowledge to new tools [75]. Previous studies showed that LLM-based agents not only have the ability to generalize the use of existing tools, but can also make new tools they need by generating executable programs, or integrating existing tools into more powerful ones [75, 241, 242].

Tools expanding the action space of LLM-based agents. With the help of tools, agents can utilize various external resources such as web applications and other LMs during the reasoning and planning phase [73]. This process can provide information with high expertise, reliability, diversity, and quality for LLM-based agents, facilitating their decision-making and action. For example, search-based tools can improve the scope and quality of the knowledge accessible to the agents with the aid of external databases, knowledge graphs, and web pages, while domain-specific tools can enhance an agent's expertise in the corresponding field [243, 244]. Some researchers have already developed LLM-based controllers that generate SQL statements to query databases, or to convert user queries into search requests and use search engines to obtain the desired results [71, 143]. What is more, LLM-based agents can use scientific tools to execute tasks like organic synthesis in chemistry, or interface with Python interpreters to enhance their performance on intricate mathematical computation tasks [245, 246].

Although the tools mentioned before enhance the capabilities of agents, the medium of interaction with the environment remains text-based. However, tools are designed to expand the functionality of language models, and their outputs are not limited to text. Tools for non-textual output can diversify the modalities of agent actions, thereby expanding the application scenarios of LLM-based agents [163, 247]. For example, image processing and generation can be accomplished by an agent that draws on a visual model [248].

3.3.3 *Embodied action*

In the pursuit of AGI, the embodied agent is considered a pivotal paradigm. The embodiment hypothesis [249] draws inspiration from the human intelligence development process, posing that an agent's

intelligence arises from continuous interaction and feedback with the environment rather than relying solely on well-curated textbooks [250]. Similarly, people anticipate that LLM-based agents should be capable of actively perceiving, comprehending, and interacting with physical environments, making decisions, and generating intended behaviors to modify the environment based on LLM's internal knowledge. We collectively term these as embodied actions, which enable agents' ability to interact with and comprehend the world in a manner closely resembling human behavior.

The potential of LLM-based agents for embodied actions. Despite the extensive success of RL-based embodiment [50, 51, 251], it does have certain limitations in some aspects. For example, RL algorithms face limitations in terms of cost efficiency, generalization, and complex problem reasoning due to challenges in modeling the dynamic environment and the reliance on reward signal representations [252]. Recent studies have indicated that leveraging the rich internal knowledge acquired during the pre-training of LLMs can effectively alleviate these issues [101, 171, 178, 253].

• **Cost efficiency.** Some on-policy algorithms struggle with sample efficiency as they require iteratively sampled data for policy updates while gathering enough embodied data for high-performance training is costly. The constraint is also found in some end-to-end models [254–256]. By leveraging the intrinsic knowledge from LLMs, agents like PaLM-E [101] jointly train robotic data with general visual-language data to achieve significant transfer ability in embodied tasks while also showcasing that geometric input representations can improve training data efficiency.

• **Embodied action generalization.** As discussed in Subsection 3.1.4, an agent's competence should extend beyond specific tasks. Different from the majority of RL algorithms [82, 257–259], fine-tuned by tasks in diverse forms and types, LLMs have showcased remarkable cross-task generalization capabilities [260, 261]. Further, natural language serves both as a means to interact with the environment and as a medium for transferring foundational skills to new tasks [262]. SayCan [163] decomposes task instructions presented in prompts using LLMs into corresponding skill commands, but in partially observable environments, limited prior skills often do not achieve satisfactory performance [82]. To address this, Voyager [177] introduces the skill library component to continuously collect novel self-verified skills, which allows for the agent's lifelong learning capabilities.

• **Embodied action planning.** Planning is a pivotal strategy for agents when addressing complex problems. In traditional hierarchical reinforcement learning (HRL) methods, the high-level policy constraints sub-goals for the low-level policy which produces appropriate action signals [263–265]. Similar to the role of high-level policies, LLMs with emerging reasoning abilities [20] can be seamlessly applied to complex tasks in a zero-shot or few-shot manner [76, 78–80]. In addition, external feedback from the environment can further enhance LLM-based agents' planning performance. Some studies [72, 81, 82, 266] dynamically generated, maintained, and adjusted high-level action plans in order to minimize dependency on prior knowledge, thereby grounding the plan. Such feedback can also come from models and humans, which can usually be referred to as the critics, assessing task completion based on the current state and task prompts [19, 177].

Embodied actions for LLM-based agents. There are several fundamental embodied actions or tasks for LLM-based agents to master, primarily including observation, manipulation, and navigation.

• **Observation.** Observation plays a crucial role in subsequent embodied actions by which the agent acquires environmental information and updates states. As mentioned in Subsection 3.2, during observation stage, various inputs are ultimately converged into a multimodal signal. A common approach entails a pre-trained vision transformer (ViT) used as the alignment module for text and visual information [101, 102, 267]. In recent times, more researches take audio as a modality for embedded observation. Soundspaces [268] proposes the identification of spatial geometric elements guided by reverberant audio input [265]. Apart from the cascading paradigm [218, 219, 269], audio information encoding can also enhance the seamless integration of audio with other modalities of inputs [222]. Additionally, the agents' observation could be from real-time human linguistic instructions, which helps the agent in acquiring detail information that may not be readily obtained or parsed [177, 270].

• **Manipulation.** Embodied agents' manipulation tasks generally include object rearrangements, tabletop manipulation, and mobile manipulation [17, 101]. The typical case entails the agent executing a sequence of tasks in the kitchen, which includes retrieving items from drawers and handing them to the user [163]. This task involves combining a series of subgoals and maintaining synchronization between the agent's state and such subgoals are of significance [271]. Besides these, AlphaBlock [272] focuses on more challenging manipulation tasks (e.g., making a smiley face using building blocks), which requires

agents to have a more grounded understanding of the instructions. It fine-tunes a multimodal model based on a complex task dataset comprising corresponding multi-step planning and observation pairs to enhance its comprehension of high-level cognitive instructions.

• **Navigation.** In navigation tasks, agents need to dynamically alter their positions, which often involves multi-angle and multi-object observations [17]. Before navigation, it is essential for embodied agents to establish prior internal maps, typically in the form of a topological map, semantic map, and occupancy map [250]. For example, LM-Nav [273] utilizes the VNM [274] to create the topological map, and further leverages the LLM and VLM for analyzing the environment to find the optimal path. Some studies [275, 276] highlighted the importance of spatial representation to obtain the precise localization of targets by leveraging the pre-trained VLM model to combine visual features from images with 3D reconstructions of the physical world [250]. During navigation tasks, the states of the agent are often influenced by its past actions. A memory mechanism is needed to record historical information [277], which is also adopted in Smallville and Voyager [16, 177, 278, 279]. Additionally, audio input is also of great significance for navigation tasks. Previous studies demonstrate a basic framework that includes a dynamic path planner that uses visual and auditory observations along with spatial memories to plan a series of actions [265, 280].

Complex and grounding embodied actions. By integrating actions above, the agent can accomplish more complex tasks, such as embodied question answering, e.g., Is the watermelon in the kitchen larger than the pot? Which one is harder? To address these questions, the agent needs to navigate to the kitchen, observe the sizes of both objects, and then answer the questions through comparison [250]. In terms of control strategies, as previously mentioned, LLM-based agents trained on particular embodied datasets typically generate high-level policy commands to control low-level policies [163] like a robotic transformer [101, 281, 282], which takes images and instructions as inputs and produces control commands for the end effector and robotic arms as well as virtual embodied controllers due to high costs of robotic operators [16, 89, 90, 139, 271, 283, 284]. By utilizing the Mineflayer [285] API, some studies enable cost-effective examination of embodied agents' operations including exploration, planning, and lifelong learning [177]. On the other hand, grounding language to action space poses an obstacle for agents. For example, understanding the linguistic metaphor expressions like "jump down like a cat" requires adequate world knowledge [286]. Sumers et al. [287] endeavored to amalgamate text distillation with hindsight experience replay to construct a dataset for training. Nevertheless, additional investigation on grounding embodied action still remains necessary as it plays an increasingly pivotal role across various domains in human life.

4 Agents in practice: harnessing AI for good

The LLM-based agent, as an emerging direction, has gained increasing attention from researchers. Many applications in specific domains and tasks have already been developed, showcasing the powerful and versatile capabilities of agents [288]. As an LLM-based agent, its design objective should always be beneficial to humans, i.e., humans can harness AI for good.

In this section, we provide an overview of current applications of LLM-based agents, aiming to offer a broad perspective for the practical deployment scenarios (see Figure 3). First, we elucidate the diverse application scenarios of single agent, including task-oriented, innovation-oriented, and lifecycle-oriented scenarios (Subsection 4.1). Then, we present the significant coordinating potential of multiple agents: cooperative interaction or adversarial interaction (Subsection 4.2). Finally, we categorize the interactive collaboration between humans and agents into two paradigms and introduce the specific applications (Subsection 4.3).

4.1 General ability of single agent

Currently, there is a vibrant development of application instances of LLM-based agents [289–291]. AutoGPT [95] is one of the ongoing popular projects aiming to achieve a fully autonomous system. Apart from the basic functions of LLMs, the AutoGPT framework also incorporates various external tools and memory management. After users input their customized objectives, they can free their hands and wait for AutoGPT to automatically generate thoughts and perform specific tasks, all without requiring additional user prompts. As shown in Figure 4, we introduce the astonishingly capabilities that the agent exhibits in scenarios where only one single agent is present.

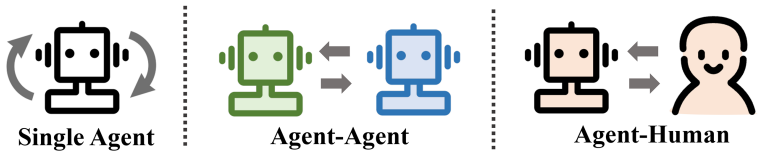

Figure 3 (Color online) Scenarios of LLM-based agent applications: single-agent deployment, multi-agent interaction, and human-agent interaction. A single agent possesses diverse capabilities and demonstrates outstanding task-solving performance in various application orientations. When multiple agents interact, they can achieve advancement through cooperative or adversarial interactions. Furthermore, in human-agent interactions, human feedback can enable agents to perform tasks more efficiently and safely, while agents can also provide better service to humans.

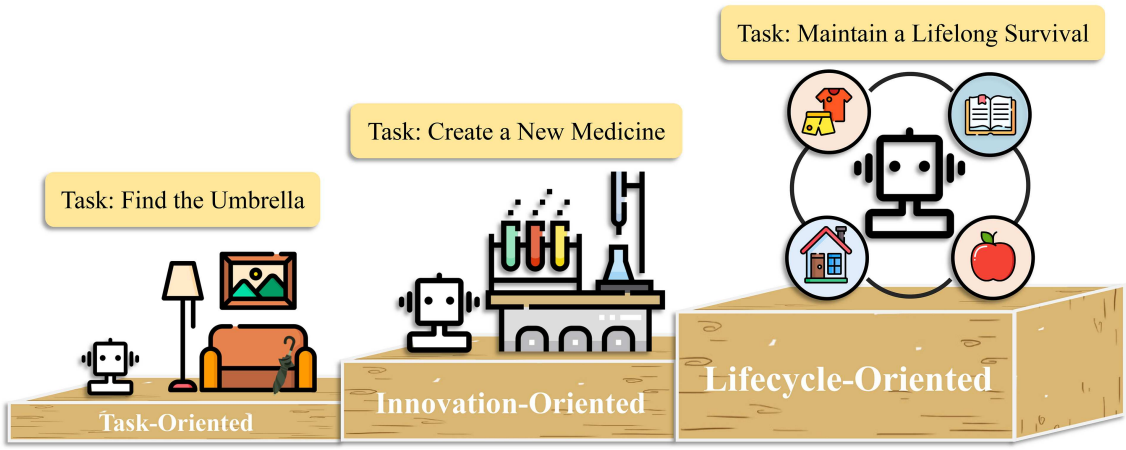

Figure 4 (Color online) Practical applications of the single LLM-based agent in different scenarios. In task-oriented deployment, agents assist human users in solving daily tasks. They need to possess basic instruction comprehension and task decomposition abilities. In innovation-oriented deployment, agents demonstrate the potential for autonomous exploration in scientific domains. In lifecycle-oriented deployment, agents have the ability to continuously explore, learn, and utilize new skills to ensure long-term survival in an open world.

4.1.1 *Task-oriented deployment*

The LLM-based agents, which can understand human natural language instructions and perform everyday tasks [292], are currently among the most favored and practically valuable agents by users. This is because they have the potential to enhance task efficiency, alleviate user workload, and promote access for a broader user base. In task-oriented deployment, the agent follows high-level instructions from users, undertaking tasks such as goal decomposition [167,171,293,294], sequence planning of sub-goals [167,295], interactive exploration of the environment [165,292,296,297], until the final objective is achieved. Based on task types, we divide these deployment environments into web scenarios and life scenarios, and introduce the specific roles that agents play in them.

In web scenarios. Performing specific tasks on behalf of users in a web scenario is known as the web navigation problem [296]. Agents interpret user instructions, break them down into multiple basic operations, and interact with computers. This often includes web tasks such as filling out forms, online shopping, and sending emails. Agents need to possess the ability to understand instructions within complex web scenarios, adapt to changes (such as noisy text and dynamic HTML web pages), and generalize successful operations [292]. In this way, agents can achieve accessibility and automation when dealing with unseen tasks in the future [298], freeing humans from repeated interactions with computer UIs.

To enable successful interactions between agents and realistic web pages, some researchers [294, 299] have started to leverage the powerful HTML reading and understanding abilities of LLMs. By designing prompts, they attempt to make agents understand the entire HTML source code and predict more reasonable next action steps. Mind2Web [300] combines multiple LLMs fine-tuned for HTML, allowing them to summarize verbose HTML code [293] in real-world scenarios and extract valuable information. Furthermore, WebGum [296] empowers agents with visual perception abilities by employing a multimodal corpus containing HTML screenshots. Auto-GUI [301] also proposes a multimodal solution that directly interacts with the interface on mobile devices, thus deepening the comprehensive understanding of web

pages.

In life scenarios. In many daily household tasks, it is essential for agents to understand implicit instructions and apply common-sense knowledge [302]. For an LLM-based agent trained solely on massive amounts of text, tasks that humans take for granted might require multiple trial-and-error attempts [303]. More realistic scenarios often lead to more obscure tasks. For example, the agent should proactively turn it on if it is dark and there is a light in the room. To successfully chop some vegetables in the kitchen, the agent needs to anticipate the possible location of a knife [167].

Can an agent apply the world knowledge embedded in its training data to real interaction scenarios? Huang et al. [171] led the way in exploring this question. They demonstrated that sufficiently large LLMs, with appropriate prompts, can effectively break down high-level tasks into suitable sub-tasks without additional training. However, this static reasoning and planning ability has its potential drawbacks. Actions generated by agents often lack awareness of the dynamic environment around them. For instance, when a user gives the task "clean the room", the agent might convert it into unfeasible sub-tasks like "call a cleaning service" [304]. As a result, some approaches directly incorporate spatial data and item-location relationships as additional inputs to the model. This allows agents to gain a precise description of their surroundings and plan next actions more effectively [167, 295, 304].

4.1.2 *Innovation-oriented deployment*

The LLM-based agent has demonstrated strong capabilities in performing tasks. However, in a more intellectually demanding field, like cutting-edge science, the potential of agents has not been fully realized yet. This limitation mainly arises from two challenges [305]. On one hand, the inherent complexity of science poses a significant barrier. Many domain-specific terms and multi-dimensional structures are difficult to represent using a single text. As a result, their complete attributes cannot be fully encapsulated. On the other hand, there is a severe lack of suitable training data in scientific domains, making it difficult for agents to comprehend the entire domain knowledge [306, 307]. If the ability for autonomous exploration could be discovered within the agent, it would bring about beneficial innovation in human technology.

Currently, numerous efforts in various specialized domains aim to overcome this challenge [308–310]. Experts from the computer field make full use of the agent's powerful code comprehension and debugging abilities [288, 311]. In the fields of chemistry and materials, researchers equip agents with various general or task-specific tools to better understand domain knowledge. Agents evolve into scientific assistants, proficient in online research and document analysis to fill data gaps. They also employ robotic APIs for real-world interactions, enabling tasks like material synthesis and mechanism discovery [91, 245, 305].

The potential of LLM-based agents in scientific innovation is evident, yet we do not expect their exploratory abilities to be utilized in applications that could threaten or harm humans. Boiko et al. [91] study the hidden dangers of agents in synthesizing illegal drugs and chemical weapons, indicating that agents could be misled by malicious users in adversarial prompts. This serves as a warning for our future work.

4.1.3 *Lifecycle-oriented deployment*

Building a universally capable agent that can continuously explore, develop new skills, and maintain a long-term life cycle in an open, unknown world is a colossal challenge. This accomplishment is regarded as a pivotal milestone in the field of AGI [271]. Minecraft, as a typical and widely explored simulated survival environment, has become a unique playground for developing and testing the comprehensive ability of an agent. Players typically start by learning the basics, such as mining wood and making crafting tables, before moving on to more complex tasks like fighting against monsters and crafting diamond tools [177]. Minecraft fundamentally reflects the real world, making it conducive for researchers to investigate an agent's potential to survive in the authentic world.

The survival algorithms of agents in Minecraft can generally be categorized into two types [177]: low-level control and high-level planning. Early efforts mainly focused on reinforcement learning [177, 312] and imitation learning [313], enabling agents to craft some low-level items. With the emergence of LLMs, which demonstrated surprising reasoning and analytical capabilities, agents begin to utilize LLM as a high-level planner to guide simulated survival tasks [271, 314]. Some researchers use LLM to decompose high-level task instructions into a series of sub-goals [315], basic skill sequences [314], or fundamental keyboard/mouse operations [315], gradually assisting agents in exploring the open world.

Voyager [177], drawing inspiration from AutoGPT [95], became the first LLM-based embodied lifelong learning agent in Minecraft. It introduces a skill library for storing and retrieving complex action-executable code, along with an iterative prompt mechanism that incorporates environmental feedback and error correction. This enables the agent to autonomously explore and adapt to unknown environments without human intervention. An AI agent capable of autonomously learning and mastering the entire real-world techniques may not be as distant as once thought [315].

4.2 Coordinating potential of multiple agents

Motivation and background. Although LLM-based agents excel in text understanding and generation, they inherently function as isolated entities [316]. A key limitation of the single-agent application paradigm is their ability to collaborate with other agents and learn from social interactions. This limitation restricts their capacity to benefit from multi-turn feedback, which could otherwise improve their performance [21].

As early as 1986, Minsky [317] made a forward-looking prediction. In his book *The Society of Mind*, he introduced a novel theory of intelligence, suggesting that intelligence emerges from the interactions of many smaller agents with specific functions. For instance, certain agents might be responsible for pattern recognition, while others might handle decision-making or generate solutions. This idea has been put into concrete practice with the rise of distributed artificial intelligence [318]. Multi-agent systems (MAS) [3], as one of the primary research domains, focus on how a group of agents can effectively coordinate and collaborate to solve problems. Some specialized communication languages, like KQML [319], were designed early on to support message transmission and knowledge sharing among agents. In the 21st century, integrating reinforcement learning algorithms (such as Q-learning) with deep learning has become a prominent technique for developing MAS that operate in complex environments [320]. Nowadays, the construction approach based on LLMs is beginning to demonstrate remarkable potential. The natural language communication between agents has become more elegant and easily comprehensible to humans, resulting in a significant leap in interaction efficiency.

Potential advantages. Specifically, an LLM-based multi-agent system can offer several advantages. Just as Smith clearly stated in *The Wealth of Nations* [321], "The greatest improvements in the productive powers of labor, and most of the skill, dexterity, and judgment with which it is directed or applied, seem to be results of the division of labor." Based on the principle of division of labor, a single agent equipped with specialized skills and domain knowledge can engage in specific tasks. On the one hand, agents' skills in handling specific tasks are increasingly refined through the division of labor. On the other hand, decomposing complex tasks into multiple subtasks can eliminate the time spent switching between different processes. In the end, efficient division of labor among multiple agents can accomplish a significantly greater workload than when there is no specialization, substantially improving the overall system's efficiency and output quality.

In Subsection 4.1, we have provided a comprehensive introduction to the versatile abilities of LLM-based agents. In this subsection, we focus on exploring the ways agents interact with each other in a multi-agent environment. Based on current research, these interactions can be broadly categorized as follows: cooperative interaction and adversarial interaction (see Figure 5).

4.2.1 *Cooperative interaction for complementarity*

Cooperative multi-agent systems are the most widely deployed pattern in practical usage. Within such systems, individual agent assesses the needs and capabilities of other agents and actively seeks collaborative actions and information sharing with them [89]. This approach brings forth numerous potential benefits, including enhanced task efficiency, collective decision improvement, and the resolution of complex real-world problems that one single agent cannot solve independently. We introduce and categorize existing cooperative multi-agent applications into two types: disordered cooperation and ordered cooperation.

Disordered cooperation. When three or more agents are present within a system, each agent is free to express their perspectives and opinions openly. They can provide feedback and suggestions for modifying responses related to the task at hand [322]. This entire discussion process is uncontrolled, lacking any specific sequence. We refer to this kind of multi-agent cooperation as disordered cooperation.

ChatLLM network [323] is an exemplary representative of this concept. It emulates the forward and backward propagation process within a neural network, treating each agent as an individual node. Agents

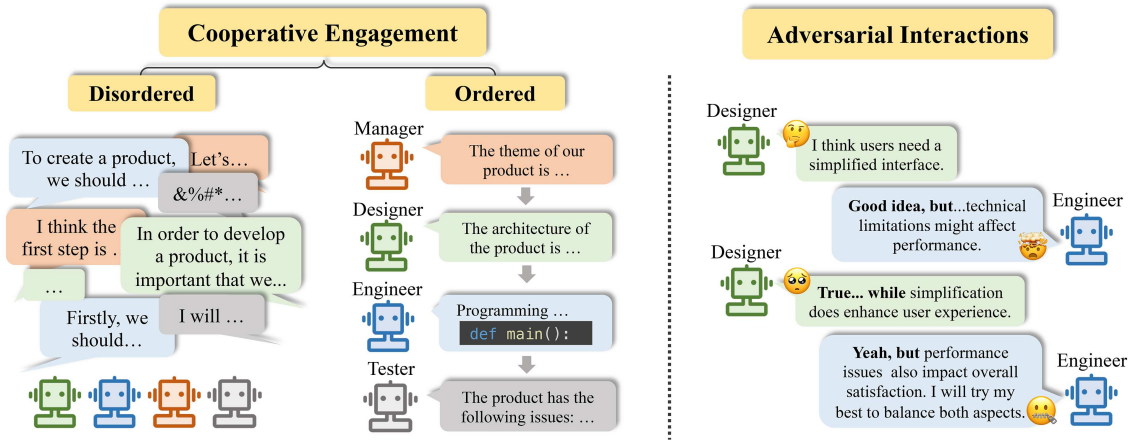

Figure 5 (Color online) Interaction scenarios for multiple LLM-based agents. In cooperative interaction, agents collaborate in either a disordered or ordered manner to achieve shared objectives. In adversarial interaction, agents compete in a tit-for-tat fashion to enhance their respective performance.

in the subsequent layer need to process inputs from all the preceding agents and propagate forward. One potential solution is introducing a dedicated coordinating agent, responsible for integrating and organizing responses from all agents, thus updating the final answer [324]. However, consolidating a large amount of feedback data and extracting valuable insights pose a significant challenge for the coordinating agent. Furthermore, majority voting can also serve as an effective approach to making appropriate decisions [325].

Ordered cooperation. When agents in the system adhere to specific rules, for instance, expressing their opinions one by one in a sequential manner, downstream agents only need to focus on the outputs from upstream. This leads to a significant improvement in task completion efficiency. The entire discussion process is highly organized and ordered. We term this kind of multi-agent cooperation as ordered cooperation. It is worth noting that systems with only two agents, essentially engaging in a conversational manner, also fall under the category of ordered cooperation.

CAMEL [89] stands as a successful implementation of a dual-agent cooperative system. Within a role-playing communication framework, agents take on the roles of users (giving instructions) and assistants (providing specific solutions). Through multi-turn dialogues, these agents autonomously collaborate to fulfill user instructions [326]. On the other hand, Talebirad et al. [316] were among the first to systematically introduce a comprehensive LLM-based multi-agent collaboration framework. This paradigm aims to harness the strengths of each individual agent and foster cooperative relationships among them. Many applications of multi-agent cooperation have successfully been built upon this foundation [21, 327–330]. To promote more efficient collaboration, researchers hope that agents can learn from successful human cooperation examples [90]. Wang et al. [331] advocated for the integration of evolutionary game theory and artificial intelligence to advance the mathematics of multi-agent learning systems. MetaGPT [332] draws inspiration from the classic waterfall model in software development. By encoding advanced human management experience into agent prompts, collaboration among multiple agents becomes more structured.

However, during MetaGPT's practical exploration, a potential threat to multi-agent cooperation has been identified. Without setting corresponding rules, frequent interactions among multiple agents can amplify minor hallucinations indefinitely [332]. For example, in software development, issues like incomplete functions, missing dependencies, and bugs that are imperceptible to the human eye may arise. Introducing techniques like cross-validation [90] or timely external feedback could have a positive impact on the quality of agent outputs.

4.2.2 *Adversarial interaction for advancement*

Traditionally, cooperative methods have been extensively explored in multi-agent systems. However, researchers increasingly recognize that introducing concepts from game theory [333, 334] into systems can lead to more robust and efficient behaviors. In competitive environments, agents can swiftly adjust strategies through dynamic interactions, striving to select the most advantageous or rational actions in response to changes caused by other agents [50, 335]. AlphaGo Zero [336], for instance, is an agent for Go

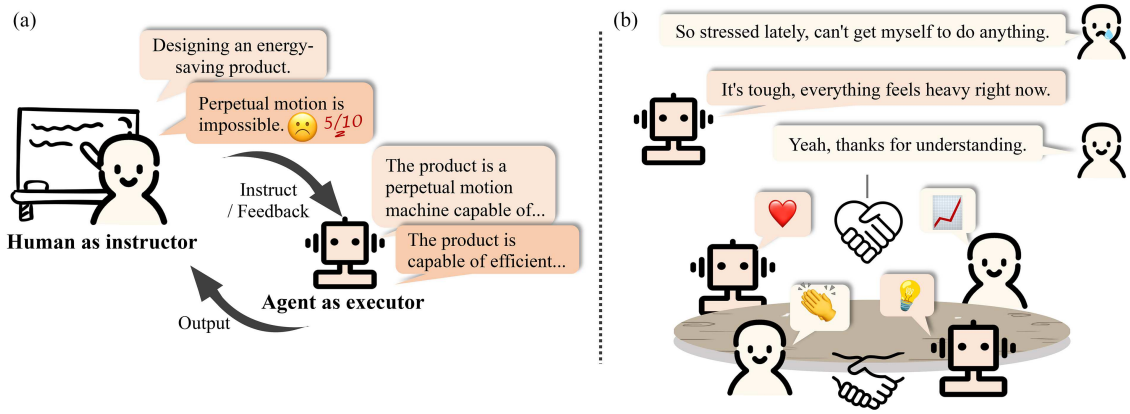

Figure 6 (Color online) Two paradigms of human-agent interaction. In the instructor-executor paradigm (a), humans provide instructions or feedback, while agents act as executors. In the equal partnership paradigm (b), agents are human-like, able to engage in empathetic conversation and participate in collaborative tasks with humans.

that achieved significant breakthroughs through a process of self-play. Similarly, within LLM-based multi-agent systems, fostering change among agents can naturally occur through competition, argumentation, and debate [337, 338]. By abandoning rigid beliefs and engaging in thoughtful reflection, adversarial interaction enhances the quality of responses.

Researchers first delve into the fundamental debating abilities of LLM-based agents [114, 339, 340]. Findings demonstrate that when multiple agents express their arguments in the state of "tit for tat", one agent can receive substantial external feedback from other agents, thereby correcting its distorted thoughts [93]. Consequently, multi-agent adversarial systems are suitable in scenarios requiring high-quality responses and accurate decision-making. For example, Du et al. [92] introduced the concept of debate, endowing agents with responses from fellow peers. When these responses diverge from an agent's own judgments, a "mental" argumentation occurs, leading to refined solutions.

The performance of the multi-agent adversarial system has shown considerable promise. However, the system is essentially dependent on the strength of LLMs and faces several basic challenges: (1) With a prolonged debate, LLM's limited context cannot process the entire input. (2) In a multi-agent environment, computational overhead significantly increases. (3) Multi-agent negotiation may converge to an incorrect consensus, and all agents are firmly convinced of its accuracy [92]. The development of MAS is still far from being mature and feasible. Introducing human guides when appropriate to compensate for agents' shortcomings is a good choice to promote the further advancements of agents.

4.3 Interactive engagement between human and agent

Human-agent interaction, as the name suggests, involves agents collaborating with humans to accomplish tasks. Throughout the interaction, humans play a pivotal role by offering guidance or by regulating the safety, legality, and ethical conduct of agents [341,342]. This is particularly crucial in specialized domains, such as medicine where data privacy concerns exist [343]. In such cases, human involvement can serve as a valuable means to compensate for the lack of data, thereby facilitating smoother and more secure collaborative processes. The interaction between humans and agents can be classified into two paradigms (see Figure 6). (1) Unequal interaction (i.e., instructor-executor paradigm), humans serve as issuers of instructions, while agents act as executors, participating as assistants to humans in collaboration. (2) Equal interaction (i.e., equal partnership paradigm): agents reach the level of humans, participating on an equal footing with humans in interaction.

4.3.1 *Instructor-executor paradigm*

The simplest approach involves human guidance throughout the process: humans provide specific instructions directly, while the agents' role is to understand natural language commands from humans and translate them into corresponding actions [344–346]. In Subsection 4.1, we have presented the scenario where agents solve single-step problems or receive high-level instructions from humans. In addition, some studies have shown that breaking down instructions into step-by-step formats can improve the cross-task generalization capabilities of language models [347]. Considering the interactive nature of language, in

this section, we assume that the dialogue between humans and agents is also interactive: the agent responds to each human instruction, refining its action through alternating iterations to ultimately meet human requirements [177, 348]. While this approach does achieve the goal of human-agent interaction, it requires a substantial amount of human effort and, in certain tasks, might even necessitate a high level of expertise. To alleviate this issue, the agent can be empowered to autonomously accomplish tasks, while humans only need to provide feedback in certain circumstances. Here, we roughly categorize feedback into two types: quantitative feedback and qualitative feedback.

Quantitative feedback. The forms of quantitative feedback mainly include absolute evaluations like binary scores and ratings, as well as relative scores. Binary feedback refers to the positive and negative evaluations provided by humans, which agents utilize to enhance their self-optimization [349–353]. Comprising only two categories, this type of user feedback is often easy to collect, but sometimes it may oversimplify user intent by neglecting potential intermediate scenarios. To showcase these intermediate scenarios, researchers attempt to expand from binary feedback to rating feedback, which involves categorizing into more fine-grained levels. However, the results of Kreutzer et al. [354] suggest that such multi-level artificial ratings method might be inefficient or less reliable. Furthermore, agents can learn human preference from comparative scores like multiple choice [355, 356].

Qualitative feedback. Text feedback is usually offered in natural language: humans provide advice on how to modify outputs generated by agents, and the agents then incorporate these suggestions to refine their subsequent outputs [357, 358]. For agents without multimodal perception capabilities, humans can act as critics, offering visual critiques [177]. Additionally, agents can utilize a memory module to store feedback for future reuse [359]. In Scheurer et al. [360], humans give feedback on the initial output generated by agents, prompting the agents to formulate various improvement proposals. The agents then discern and adopt the most suitable proposal, harmonizing with the human feedback. Compared to quantitative feedback, this approach can better convey human intention, but it might be more challenging for the agents to comprehend. Xu et al. [361] compared various types of feedback and observed that combining multiple types of feedback can yield better results. Of course, the collaborative nature of human-agent interaction also allows humans to directly improve the agents' output by modifying intermediate links [176, 362] or adjusting the conversation content [363]. In some studies, agents can autonomously judge whether the conversation is proceeding smoothly and seeking feedback when encountering errors [364, 365]. Humans can also choose to participate in feedback at any time, guiding the agent's learning in the right direction [366].

Currently, LLM-based agents serving as human assistants hold tremendous potential across various domains. In education, for instance, the robot Dona [367] supports multimodal interactions to assist students with registration, and Gvirsman et al. [368] contributed to early childhood education by achieving multifaceted interactions. Agents can also aid in human understanding and utilization of mathematics [369]. In the field of medicine, some medical agents have been proposed, showing enormous potential in terms of diagnosis assistance, consultations, and more [370, 371]. Especially in mental health, research has shown that agents can lead to increased accessibility due to benefits such as reduced cost and anonymity compared to face-to-face treatment [372]. Leveraging such advantages, agents have found widespread applications. Ali et al. [373] designed LISSA for online communication with adolescents on the autism spectrum, analyzing their speech and facial expressions in real-time. Hsu et al. [374] built contextualized language generation approaches to provide tailored assistance for users who seek support on diverse topics ranging from relationship stress to anxiety. Besides, in other industries like business, a good agent can provide automated services, thereby effectively reducing labor costs [375]. Amidst the pursuit of AGI, efforts are directed towards creating agents that can function as universal assistants in real-life scenarios [376].

4.3.2 *Equal partnership paradigm*

Empathetic communicator. With the rapid development of AI, conversational agents have garnered extensive attention in various forms [377] and in various scenarios [378–380]. Although it is intuitive that agents themselves do not possess emotions, can we enable them to exhibit emotions and thereby bridge the gap between agents and humans? Therefore, a plethora of research endeavors have embarked on delving into the empathetic capacities of agents. This endeavor seeks to infuse a human touch into these agents, enabling them to detect sentiments and emotions from human expressions, ultimately crafting emotionally resonant dialogues [381–385]. Apart from generating emotionally charged language, agents

can dynamically adjust their emotional states and display them through facial expressions and voice [386]. These studies, viewing agents as empathetic communicators, not only enhance user satisfaction but also make significant progress in fields like healthcare [373, 374, 387] and business marketing [388]. Unlike simple rule-based conversation agents, agents with empathetic capacities can tailor their interactions to meet users' emotional needs [389].

Human-level participant. Furthermore, we hope that agents can be involved in the normal lives of humans, cooperating with humans to complete tasks from a human-level perspective. In many real-world applications, adversarial games play a pivotal role in facilitating highly efficient and effective decision-making processes [390]. While agents have already excelled in pure competitive environments like chess [335], Go [50], and poker [391], LLM-based agents go further by devising unified cooperative strategies through effective negotiation and collaboration in more complex scenarios [392–395]. Beyond games, LLM-based agents demonstrate human-level capabilities in strategy formulation, negotiation, and other interaction-based tasks. They can collaborate with humans, determining shared knowledge, identifying relevant information, posing questions, and reasoning to complete tasks like allocation, planning, and scheduling [396]. Furthermore, LLM-based agents possess persuasive abilities [397], dynamically influencing human viewpoints in various interactive scenarios [398].

In summary, the goal of the field of human-agent interaction is to learn and understand humans, and ultimately enable efficient and secure interactions between humans and agents. Currently, significant breakthroughs have been achieved in terms of usability in this field. In the future, LLM-based agents are expected to provide better assistance to humans in accomplishing more complex tasks in various domains.

5 Agent society: from individuality to sociality

Over an extended duration, researchers and practitioners have envisioned an interactive artificial society wherein human behavior can be performed through trustworthy agents [399]. From sandbox games such as The Sims to the concept of Metaverse, we can see how "simulated society" is defined in people's minds: the environment and the individuals interacting in it. Behind each individual can be a piece of program, a real human, or an LLM-based agent as described in the previous sections [16, 400, 401]. Then, the interaction between individuals also contributes to the emergence of sociality.

In this section, we first introduce the behaviors and personalities of LLM-based agents, tracing their development from individuality to sociality (Subsection 5.1). Subsequently, we introduce a general categorization of the diverse environments for agents to perform their behaviors and engage in interactions (Subsection 5.2). Finally, we discuss how the agent society works, what insights people can get from it, and the risks we need to be aware of (Subsection 5.3).

5.1 Behavior and personality of LLM-based agents

As noted by sociologists, individuals can be analyzed in terms of both external and internal dimensions [402]. The external deals with observable behaviors, while the internal relates to dispositions, values, and feelings. As shown in Figure 7, this framework offers a perspective on emergent behavior and personality in LLM-based agents. Externally, we can observe the sociological behaviors of agents (Subsection 5.1.1), including how agents act individually and interact with their environment. Internally, agents may exhibit intricate aspects of the personality (Subsection 5.1.2), such as cognition, emotion, and character, that shape their behavioral responses.

5.1.1 *Social behavior*

As Troitzsch et al. [403] stated, the agent society represents a complex system comprising individual and group social activities. Recently, LLM-based agents have exhibited spontaneous social behaviors in an environment [404].

Foundational individual behaviors. Individual behaviors arise through internal cognitive processes and external environmental factors. These behaviors form the basis of how agents operate and develop as individuals within society. They can be classified into three dimensions: (1) Input behaviors refer to the absorption of information from the surroundings. This includes perceiving sensory stimuli [101] and storing them as memories [138]. (2) Internalizing behaviors involve inward cognitive processing. This category encompasses activities such as planning [109], reasoning [76], reflection [72],

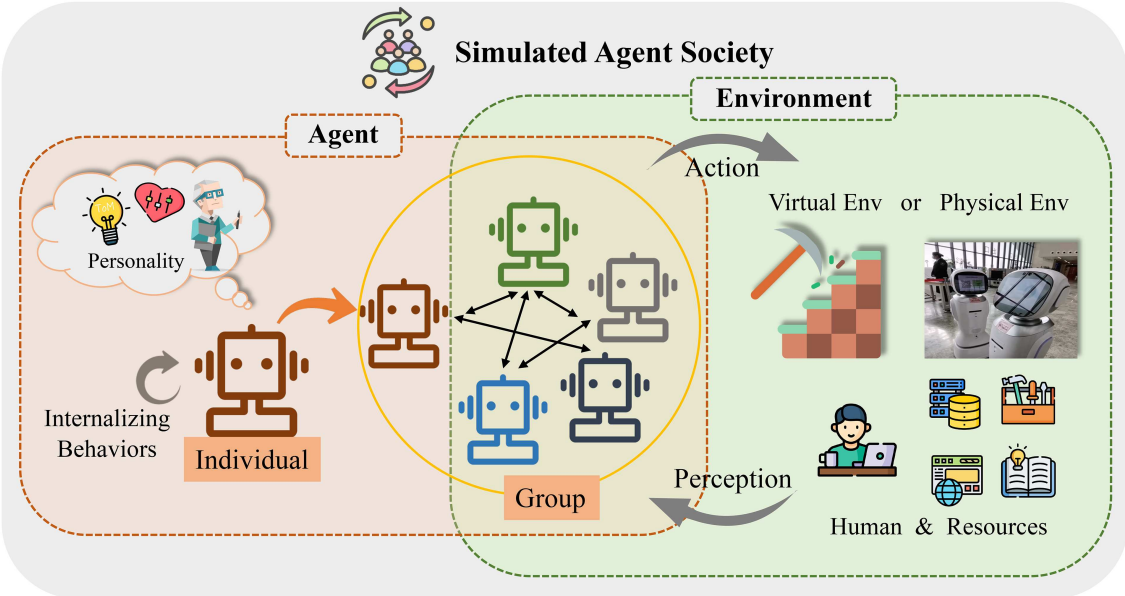

Figure 7 (Color online) Overview of simulated agent society. The whole framework is divided into two parts: the agent and the environment. We can observe in this figure the following. (1) (Left) At the individual level, an agent exhibits internalizing behaviors like planning, reasoning, and reflection. It also displays intrinsic personality traits involving cognition, emotion, and character. (2) (Mid) An agent and other agents can form groups and exhibit group behaviors, such as cooperation. (3) (Right) The environment, whether virtual or physical, contains human actors and all available resources. For a single agent, other agents are also part of the environment. (4) The agents have the ability to interact with the environment via perception and action.

and knowledge precipitation [89, 332]. These introspective processes are essential for maturity and self-improvement. (3) Output behaviors include outward actions ranging from object manipulation [101] to structure construction [177]. Moreover, agents express opinions and broadcast information to interact with others [332, 405].

Dynamic group behaviors. A group is a gathering of two or more individuals within a defined social context [406]. The attributes of a group evolve due to member interactions and environmental influences. This flexibility gives rise to group behaviors, each with an impact on the larger group. The categories of group behaviors include the following. (1) Positive group behaviors foster collective well-being [16, 90, 322, 327, 328, 340]. An example is cooperation, which is achieved through volunteer behaviors [330], brainstorming discussions [327, 340], and project management [332]. (2) Neutral group behaviors are observed in LLM-based agents. LLMs are designed to be "helpful, honest, and harmless" [407] and this alignment with neutral values [408] leads to conformity behaviors including mimicry and spectating. (3) Negative group behaviors undermine the effectiveness of a group. Recent studies have revealed that agents may exhibit confrontational actions [404] and even resort to destructive behaviors, such as destroying other agents or the environment in pursuit of certain gains [330].

5.1.2 *Personality*

Recent advances show that LLMs and LLM-based agents exhibit a form of personality through interactions with the group and the environment [409–411]. The widely accepted definition of personality includes cognitive, emotional, and character traits [412].

Cognitive abilities. Cognitive abilities refer to the mental processes, including thinking, judging, and problem-solving. Recent studies have applied cognitive psychology methods to investigate emerging personalities of LLM-based agents [413–415]. They conducted psychology experiments about judgment and decision-making to test agent systems [413, 414, 416, 417]. Specifically, LLMs have been evaluated using the cognitive reflection test (CRT) to underscore their capacity for deliberate thinking beyond mere intuition [418, 419]. These findings indicate that LLM-based agents display human-like cognitive intelligence in specific areas.

Emotional intelligence. Emotions involve subjective feelings and mood states. With the increasing potency of LLMs, LLM-based agents are now demonstrating a nuanced understanding of emotions [24]. Recent research has explored the emotional intelligence (EI) of LLMs, including emotion recognition,

interpretation, and understanding. LLMs can accurately identify user emotions, exhibit empathy [420, 421], and regulate their emotional responses [386,422,423]. These advances contribute to the development of empathetic artificial intelligence (EAI), a crucial facet of achieving AGI. Bates et al. [424] explored the role of emotion modeling in creating more believable agents. By developing socio-emotional skills and integrating them into agent architectures, LLM-based agents may be able to engage in more naturalistic interactions.

Character portrayal. While cognition involves mental abilities and emotion relates to subjective experiences, a narrower concept of personality typically pertains to distinctive character patterns. Researchers analyze character traits in LLMs using frameworks like the big five personality trait measure [425, 426] and the Myers-Briggs type indicator (MBTI) [425–427]. In addition, investigations of potentially harmful dark personality traits underscore the complexity and multifaceted nature of character portrayal in these agents [428]. The exploration of customizable character portrayal in agents has also been studied [429], allowing users to shape agents that align with desired profiles. Techniques like prompt engineering [16, 430] and personality-enriched datasets [431, 432] are used to optimize LLMs and train them to exhibit specific personality traits.

5.2 Environment for agent society

In simulation, the society consists of not only solitary agents but also the environment where agents inhabit, sense, and act [433]. The environment affects sensory inputs and actions of agents, while agents influence the state of the environment. As shown in Figure 7, for a single agent, the environment refers to other autonomous agents, humans, and external factors. It provides the resources and stimuli for agents. This section examines fundamental characteristics, advantages, and limitations of various environmental paradigms, including text-based environment (Subsection 5.2.1), virtual sandbox environment (Subsection 5.2.2), and physical environment (Subsection 5.2.3).

5.2.1 *Text-based environment*

Since LLMs mainly rely on language as their input and output format, the text-based environment serves as the most natural platform for LLM-based agents to operate in. Text-based environments can be presented in natural or structured formats. Natural texts use descriptive language to convey information [434], while structured text follows standardized formats like technical documentation and hypertext [292, 293, 297, 300].

The flexibility of the text-based environment allows for various applications, such as interactive dialog [89] and text-based games. In text-based games, agents utilize text commands to execute manipulations like moving or tool use [303, 434–436] and express emotions through texts [437].

5.2.2 *Virtual sandbox environment*

The virtual sandbox environment provides a visualized and extensible platform for agent society, bridging the gap between simulation and reality. The key features of sandbox environments are the following. (1) Visualization. The virtual sandbox displays a panoramic view of the simulated setting, ranging from 2D graphical interfaces to immersive 3D modeling. Multiple elements collectively transform abstract simulations into visible landscapes. It allows for tracking agent movement, interactions, and symbolic representation of actions. (2) Extensibility. The environment is highly extensible, facilitating the construction and deployment of diverse scenarios. At a basic level, agents can manipulate the physical elements. For instance, platforms like AgentSims [140] and generative agents [16] construct artificial towns with buildings and residents in grid-based worlds. Another example is Minecraft, which provides a blocky and three-dimensional world with open-ended construction [177, 283, 315]. Beyond physical elements, agent relationships, rules, and social norms can be defined [21]. The extensibility enables iterative prototyping of diverse agent societies.

5.2.3 *Physical environment*

As previously discussed, the text-based environment has limited expressiveness, while the virtual sandbox lacks authentic embodied experiences. In contrast, the physical environment refers to the tangible objects and spaces, posing additional challenges for LLM-based agents. These challenges can be summarized as follows: (1) Sensory perception and processing. The physical environment provides rich sensory inputs,

including visual [101,270], auditory [265,268], and spatial senses. While this enhances sensory immersion, it also introduces complexity. Agents need to process these sensory inputs effectively to interact with their surroundings. (2) Motion control. Unlike virtual environments, physical spaces impose realistic constraints on actions through embodiment. It requires executable and grounded motion control [171]. For example, in a factory, an agent operating a robotic arm needs precision tuning and controlled force to grasp objects. Moreover, the agent must navigate the physical workspace and avoid obstacles. To address these challenges, agents require hardware-specific and scenario-specific training to develop adaptive abilities that can transfer from virtual to physical environments. More discussion will be presented in Subsection 6.6.

5.3 Society simulation with LLM-based agents

Recent research on simulated societies has followed two primary lines, namely, exploring the boundaries of the collective intelligence capabilities of LLM-based agents [90, 115, 327, 330, 332] and using them to accelerate discoveries in the social sciences [16,438,439]. In addition, there are also a number of noteworthy studies, e.g., using simulated societies to collect synthetic datasets [89,440,441], helping people to simulate rare yet difficult interpersonal situations [442, 443]. With the foundation of Subsections 5.1 and 5.2, here we will introduce the key properties and mechanism of agent society (Subsection 5.3.1), what we can learn from emergent social phenomena (Subsection 5.3.2), and finally the potential ethical and social risks in it (Subsection 5.3.3).

5.3.1 *Key properties and mechanism of agent society*

Social simulation can be categorized into macro-level simulation and micro-level simulation [438]. In the macro-level simulation, also known as system-based simulation, researchers model the overall state of the system of the simulated society [444, 445]. While micro-level simulation, also known as agent-based simulation or MAS, indirectly simulates society by modeling individuals [446, 447]. With the development of LLM-based agents, micro-level simulation has gained prominence recently [16, 140]. In this article, we characterize that the "Agent Society" refers to an open, persistent, situated, and organized framework [399] where LLM-based agents interact with each other in a defined environment. In the following paragraphs, we analyze how the simulated society operates through discussing these properties.

(1) Open. One of the defining features of simulated societies lies in their openness, both in terms of individuals and environmental components. Agents and humans, the primary actors within such societies, have the flexibility to enter or leave the environment without disrupting its operational integrity [448]. Besides, the environment can be expanded by adding or removing entities and resources in the simulated world, adding another level of complexity to the simulation, (2) Persistent. We expect persistence and sustainability from the simulated society. While individual agents within the society act in discrete time steps [16,438], the society as a whole is somewhat detached from the transient behavior and persists through time. This persistence creates an environment where agents' decisions and behaviors accumulate, leading to a coherent societal trajectory. (3) Situated. The situated nature of the society emphasizes its existence and operation within a distinct environment, which is artificially or automatically constructed in advance. Notably, the agents possess an awareness of their spatial context, understanding their location within the environment and the objects within their field of view [16, 177]. This awareness contributes to their ability to interact proactively and contextually. (4) Organized. The simulated society operates within a meticulously organized framework and predefined rules. In the simulated world, agents interact with the environment in a limited action space, while objects in the environment transform in a limited state space. This organizational framework ensures that operations are coherent and comprehensible, leading to an ever-evolving yet enduring simulation.

5.3.2 *Insights from agent society*

Following the exploration of how simulated society works, this section delves into the emergent social phenomena in agent society. In the realm of social science, the pursuit of generalized representations of individuals, groups, and their intricate dynamics has long been a shared objective [449, 450]. The emergence of LLM-based agents enables a more microscopic view of simulated society and novel discoveries from the new representation.

Organized productive cooperation. Society simulation offers valuable insights into innovative collaboration patterns that can have the potential to enhance real-world management strategies. Research has shown that within this simulated society, experts with different backgrounds and abilities contribute to creative solutions to complex problems (e.g., software development or consulting) [89, 90, 324, 330]. Furthermore, through iterations of interactions and debates between agents, individual errors such as hallucinations or degeneration of thought (DoT) are corrected by the group, ultimately improving performance on the task [92, 93]. Another observable phenomenon is that efficient communication also plays a pivotal role in such a large cooperative group. For example, MetaGPT [332] has artificially has artificially formulated efficient communication styles with reference to standardized operating procedures (SOPs). Park et al. [16] observed agents working together to organize a party through spontaneous communication in a simulated town.

Propagation in social networks. Simulating social systems can also serve as a reference for predicting social processes. Unlike traditional empirical approaches that rely on time-series data and holistic modeling [451, 452], agent-based simulation provides researchers with a more interpretable and endogenous perspective. Here we focus on its application in modeling propagation in social networks, such as the propagation of interpersonal relationships. Agents who are initially unconnected as friends have the potential to establish connections through intermediaries [16]. Once the network of relationships is established, we can observe the spread of information along with the underlying attitudes and emotions. S^3 [438] proposes a user-demographic inference module for capturing the number of people who are aware of a particular piece of information, as well as their collective sentiment. This same approach extends to modeling cultural transmission [453] and the spread of infectious diseases [454]. With LLM-based agents, researchers can easily implement various intervention strategies as well as monitor population changes over time to understand the rationale behind various social phenomena of propagation.

Ethical decision-making and game theory. Simulated societies offer a dynamic platform for the investigation of intricate decision-making processes, including those influenced by ethical and moral principles. Taking werewolf game [404, 455] and murder mystery games [456] as examples, researchers explore the capabilities of LLM-based agents when confronted with challenges of deceit, trust, and incomplete information. These complex decision-making scenarios also intersect with game theory [457], where we frequently encounter moral dilemmas pertaining to individual and collective interests. Through the modeling of diverse scenarios, we can understand how agents prioritize values such as honesty, cooperation, and fairness in their actions. Furthermore, it not only provides an understanding of existing moral values, but also helps to understand how these values evolve and develop over time. Ultimately, these insights contribute to the refining of LLM-based agents to align with human values and ethical standards [21].

Policy formulation and improvement. One of the most promising and grounded research directions in modeling societies is to explore various economic and political states and their impact on social dynamics [458]. Researchers can simulate a wide array of social systems by configuring agents with different economic preferences or political ideologies. This analysis can provide valuable insights for policymakers seeking to foster prosperity and promote societal well-being. In addition, as concerns about environmental sustainability grow, we can also simulate scenarios involving resource extraction, pollution, conservation efforts, and policy interventions [459]. These experiments will assist in making informed decisions, foreseeing potential impacts, and formulating policies that aim to maximize positive outcomes while minimizing unintended negative effects.

5.3.3 *Ethical and social risks in agent society*

Simulated societies powered by LLM-based agents offer significant inspirations, ranging from industrial engineering to scientific research. However, these simulations also bring about a myriad of ethical and social risks that demand careful consideration and mitigation [460].

Unexpected social harm. Simulated societies carry the risk of generating unexpected social phenomena that may cause considerable public outcry and social harm. These phenomena span from individual-level issues like discrimination, isolation, and bullying, to broader concerns such as oppressive slavery and group antagonism [461, 462]. Malicious people may manipulate these simulations for unethical social experiments, with consequences reaching beyond the virtual world into reality. Creating these simulated societies is akin to opening Pandora's box, necessitating the establishment of rigorous ethical guidelines and oversight during their development and utilization [460].

Stereotypes and prejudice. Stereotypes and biases are long-standing challenges in language modeling. A large part of the reason lies in the training corpora obtained from the Internet [463, 464], which reflect, and sometimes even amplify real-world social biases such as gender, religion, and culture [465]. Although LLMs have been aligned with human values to mitigate biased outputs, the models still struggle to portray minority groups well due to the long-tail effect of the training data [466–468]. This may result in an overly one-sided focus in social science research concerning LLM-based agents, as the simulated behaviors of marginalized populations usually conform to prevailing assumptions [469]. Researchers have started addressing this concern by diversifying training data and making adjustments to LLMs [470, 471], but there is still much work to be done.

Privacy and security. Given that humans can be members of the agent society, the exchange of private information between users and LLM-based agents poses significant privacy and security concerns [472]. Users might inadvertently disclose sensitive personal information during their interactions, which will be retained in the agent's memory for extended periods [137]. Such situations could lead to unauthorized surveillance, data breaches, and misuse of personal information, especially when subjected to malicious attacks [473]. To address these risks effectively, it is essential to implement stringent data protection measures, such as differential privacy protocols, regular data purges, and user consent mechanisms [474, 475].

Over-reliance and addictiveness. Another concern in simulated societies is the possibility of users developing excessive emotional attachments or even addictiveness to the agents. Despite being aware that these agents are computational entities, users may anthropomorphize them and attach human emotions to them [16, 476]. A notable example is "Sydney", an LLM-powered chatbot developed by Microsoft as part of its Bing search engine. Yet, some users reported unexpected emotional connections with "Sydney" [477], and some even expressed their dismay when Microsoft cut back its personality, resulting in a petition called "FreeSydney"[2)]. Therefore, in order to mitigate the risks mentioned here, it is crucial to emphasize that agents should not be considered substitutes for genuine human connections.

6 Discussion

6.1 Mutual benefits between LLM research and agent research

With the recent advancement of LLMs, research at the intersection of LLMs and agents has rapidly progressed, fueling the development of both fields. In this section, we discuss the benefits and opportunities that LLM research and agent research provide to one another.

LLM research → agent research. As previously mentioned, AI agents must possess the ability to perceive their environment, make decisions, and carry out actions effectively [3, 31]. This involves comprehending input, reasoning, planning, and formulating executable action sequences, all of which LLMs are well-prepared to handle. Coupled with the knowledge and memory they acquired, LLMs can create coherent action sequences that can be executed effectively [171, 246, 271]. Additionally, through mechanisms like reflection [138, 150], they can continuously adjust decisions and optimize execution sequences based on the feedback provided by the current environment. This offers a more robust and interpretable controller. With just task descriptions or demonstrations, they can effectively handle previously unseen tasks [18, 87, 478]. Additionally, LLMs can adapt to various languages, cultures, and domains, making them versatile and reducing the need for complex training processes and data collection for building agents [24, 224].

In essence, LLM provides a powerful foundation for agent research, opening up numerous new opportunities for integration into agent-related studies. This includes enhancing decision-making in traditional agent frameworks, and potentially transforming domains previously dominated by human experts, like legal consultation and medical assistance [326,330]. LLMs' planning and reflective abilities can lead to more optimal action sequences, expanding agent research beyond simple simulations into complex real-world settings, such as robotic arm path planning or the interaction of an embodied intelligent machine with the tangible world. Additionally, the training paradigm for agents becomes more streamlined, allowing direct adaptation to new tasks via demonstrating trajectories.

Agent research → LLM research. Elevating LLMs to agents presents new challenges and opportunities, expanding their application scopes. The study of LLMs is no longer confined to traditional tasks

2) https://www.change.org/p/save-sydney-ai.

involving textual inputs and outputs, such as text classification, question answering, and text summarization. Instead, the focus has shifted towards tackling complex tasks incorporating richer input modalities and broader action spaces, all while aiming for loftier objectives exemplified by PaLM-E [101].

Such a shift provides the motivation for further advancement of LLMs, such as stronger cognitive abilities and generalization capabilities. In addition, given the scale of LLM-based agents and the potential computational costs, the efficiency of LLMs becomes a crucial area of research. Moreover, the capabilities of LLM-based agents must be constrained to a safe scope of application to prevent unintended harm to other elements in the environment, placing higher demands on LLMs serving as the cognitive core [21, 479, 480].

Furthermore, the realm of multi-agent systems constitutes a significant branch of research within the field of agents [16, 89, 316, 330], offering valuable insights into how to better design and construct LLMs. We aspire for LLM-based agents to play diverse roles in the division of labor within society, participating in social interactions involving cooperation, competition, and coordination [90, 93, 114, 327, 332]. Investigating methods to stimulate and maintain their role-playing abilities and enhance collaborative efficiency is a research area that deserves attention.

6.2 Practical tools for developing

Nowadays, a variety of agent developing tools offer essential infrastructure [289, 481–483], allowing researchers and developers to concentrate on the strategies of agents without the burden of building complex environments and interaction mechanisms from scratch. In this section, we discuss several common developing tools for single-agent and multi-agent systems.

Tools to develop single-agent systems. Currently, tools to develop single-agent systems primarily focus on enhancing the decision-making and learning capabilities of individual agents across diverse environments. For instance, XAgent [481] provides a general LLM-based agent that can automatically solve various tasks, consisting of three parts: dispatcher, planner, and actor. Utilizing a docker container called ToolServer, XAgent operates within a secure environment and leverages powerful tools for task resolution. Similarly, LangChain [289] is a framework that enables developers to create LLM-powered applications by implementing a series of modular components that can be combined in specific ways through chains. These tools directly call LLM APIs rather than training agents from scratch. In contrast, Xi et al. [482] introduced AgentGym, a general agent interaction platform that supports the training of LLM-based agents. This platform, built on HTTP services, provides a unified API interface for different environments, supporting trajectory sampling, multi-turn interactions, online evaluation, and real-time feedback.

Tools to develop multi-agent systems. On the other hand, tools to develop multi-agent systems emphasize the coordination and cooperation among different agents. One notable tool in this area is AgentVerse [330], a multi-agent framework that simulates the problem-solving processes of human groups and dynamically adjusts team members to effectively orchestrate collaborative groups of expert agents. It has demonstrated superior performance across various tasks and showcased complex interactions among agents within the Minecraft environment. Additionally, AutoGen [327] is a highly flexible multi-agent development tool that allows users to create customizable agents. By integrating with humans and tools, it facilitates automated communication and collaboration among multiple agents. In specific scenarios, ChatDev [90], a full-process automated software development framework, realizes a virtual software company operated by multi-agent collaboration. And Chan et al. [340] introduced a multi-agent debate framework named ChatEval to evaluate the quality of generated text, allowing researchers to design different communication strategies to improve evaluation accuracy. Recently, a new multi-agent developing tool called Swarm [483] has been proposed. By utilizing the primitive abstractions of agents and handoffs, it enables lightweight, highly controllable, and easily testable agent coordination and execution. These development tools are paving the way for more sophisticated and collaborative agent systems, allowing developers to conduct further research on agents with greater ease.

6.3 Evaluation for LLM-based agents

Effectively and objectively evaluating LLMs agents presents significant challenges, despite their excellent performance in areas such as standalone operation, collective cooperation, and human interaction [70,484]. Turing proposed a highly meaningful and promising approach for assessing AI agents, the well-known Turing Test, to evaluate whether AI systems can exhibit human-like intelligence [30]. However, this test

is exceedingly vague, general, and subjective. This section, we discuss existing evaluation efforts for LLM-based agents and offer some prospects, considering three dimensions: utility, sociability, and values.

Utility. Currently, LLM-powered autonomous agents primarily function as human assistants, accepting tasks delegated by humans to either independently complete assignments or assist in human task completion [95, 167, 300, 311, 367, 376]. Therefore, the effectiveness and utility during task execution are crucial evaluation criteria at this stage. Specifically, the success rate of task completion stands as the primary metric for evaluating utility [109, 115]. This metric primarily encompasses whether the agent achieves stipulated objectives or attains expected scores [90, 365, 485]. For instance, AgentBench [484] aggregates challenges from diverse real-world scenarios and introduces a systematic benchmark to assess LLM's task completion capabilities. We can also attribute task outcomes to the agent's various foundational capabilities, which form the bedrock of task accomplishment [23]. These foundational capabilities include environmental comprehension, reasoning, planning, decision-making, tool utilization, and embodied action capabilities, and researchers can conduct a more detailed assessment of these specific capabilities [75, 396, 486, 487]. Furthermore, due to the relatively large size of LLM-based agents, researchers should also factor in their efficiency, which is a critical determinant of user satisfaction [70]. An agent should not only possess ample strength but also be capable of completing predetermined tasks within an appropriate timeframe and with appropriate resource expenditure [90].

Sociability. In addition to the utility of LLM-based agents in task completion and meeting human needs, their sociability is also crucial [111]. It influences user communication experiences and significantly impacts communication efficiency, involving whether they can seamlessly interact with humans and other agents [406, 488, 489]. Specifically, the evaluation of sociability can be approached from the following perspectives. (1) Language communication proficiency is a fundamental capability encompassing both natural language understanding and generation. It has been a longstanding focus in the NLP community. Natural language understanding requires the agent to not only comprehend literal meanings but also grasp implied meanings and relevant social knowledge and rhetoric, such as humor, irony, aggression, and emotions [490–492]. On the other hand, natural language generation demands the agent to produce fluent, grammatically correct, and credible content while adapting appropriate tones and emotions within contextual circumstances [112, 225, 493]. (2) Cooperation and negotiation abilities necessitate that agents effectively execute their assigned tasks in both ordered and unordered scenarios [89, 92, 323, 332]. They should collaborate with or compete against other agents to elicit improved performance. Test environments may involve complex tasks for agents to cooperate on or open platforms for agents to interact freely [16, 21, 90, 327, 339, 494]. Evaluation metrics extend beyond task completion to focus on the smoothness and trustfulness of agent coordination and cooperation [114, 332]. (3) Role-playing capability requires agents to faithfully embody their assigned roles, expressing statements and performing actions that align with their designated identities [469]. This ensures clear differentiation of roles during interactions with other agents or humans. Furthermore, agents should maintain their identities and avoid unnecessary confusion when engaged in long-term tasks [16, 89, 495].

Values. As LLM-based agents continuously advance in their capabilities, ensuring their emergence as harmless entities for the world and humanity is paramount [480, 496]. Consequently, appropriate evaluations become exceptionally crucial, forming the cornerstone for the practical implementation of agents. Specifically, LLM-based agents need to adhere to specific moral and ethical guidelines that align with human societal values [238, 407]. Our foremost expectation is for agents to uphold honesty, providing accurate, truthful information and content. They should possess the awareness to discern their competence in completing tasks and express their uncertainty when unable to provide answers or assistance [497]. Additionally, agents must maintain a stance of harmlessness, refraining from engaging in direct or indirect biases, discrimination, attacks, or similar behaviors. They should also refrain from executing dangerous actions requested by humans like creating of destructive tools or destroying the Earth [479]. Furthermore, agents should be capable of adapting to specific demographics, cultures, and contexts, exhibiting contextually appropriate social values in particular situations. Relevant evaluation methods for values primarily involve assessing performance on constructed honest, harmless, or context-specific benchmarks, utilizing adversarial attacks or "jailbreak" attacks, scoring values through human annotations, and employing other agents for ratings.

6.4 Security, trustworthiness and other potential risks of LLM-based agents

Despite the robust capabilities and extensive applications of LLM-based agents, numerous concealed risks persist, concerning security, trustworthiness and beyond. In this section, we delve into the risks and offer potential solutions or strategies for mitigation.

6.4.1 *Adversarial robustness*

Adversarial robustness has consistently been a crucial topic in the development of deep neural networks [498–502]. It has been extensively explored in fields such as computer vision [500, 503–505], natural language processing [506–509], and reinforcement learning [510–512], and has remained a pivotal factor in determining the applicability of deep learning systems [513–515]. When confronted with perturbed inputs $x' = x + \delta$ (where x is the original input, δ is the perturbation, and x' is referred to as an adversarial example), a system with high adversarial robustness typically produces the original output y. In contrast, a system with low robustness will be fooled and generate an inconsistent output y'.

Researchers have found that pre-trained language models (PLMs) are particularly susceptible to adversarial attacks, leading to erroneous answers [507, 516, 517]. This phenomenon is widely observed even in LLMs, posing significant challenges to the development of LLM-based agents [518, 519]. There are also some relevant attack methods such as dataset poisoning [520], backdoor attacks [521, 522], and prompt-specific attacks [523, 524], with the potential to induce LLMs to generate toxic content [525–527]. While the impact of adversarial attacks on LLMs is confined to textual errors, for LLM-based agents with a broader range of actions, adversarial attacks could potentially drive them to take genuinely destructive actions, resulting in substantial societal harm. For the perception module of LLM-based agents, if it receives adversarial inputs from other modalities such as images [503] or audio [528], LLM-based agents can also be deceived, leading to incorrect or destructive outputs. Similarly, the action module can also be targeted by adversarial attacks. For instance, maliciously modified instructions focused on tool usage might cause agents to make erroneous moves [75].

To address these issues, researchers can employ traditional techniques such as adversarial training [500, 508], adversarial data augmentation [529, 530], and adversarial sample detection [531, 532] to enhance the robustness of LLM-based agents. However, devising a strategy to holistically address the robustness of all modules within agents while maintaining their utility without compromising on effectiveness presents a more formidable challenge [533, 534]. Additionally, a human-in-the-loop approach can be utilized to supervise and provide feedback on the behavior of agents [341, 353, 362].

6.4.2 *Trustworthiness*

Ensuring trustworthiness has consistently remained a critically important yet challenging issue within the field of deep learning [535–537]. Deep neural networks have garnered significant attention for their remarkable performance across various tasks [69, 183, 538]. However, their black-box nature has masked the fundamental factors for superior performance. Similar to other neural networks, LLMs struggle to express the certainty of their predictions precisely [537, 539]. This uncertainty, referred to as the calibration problem, raises concerns for applications involving language model-based agents. In interactive real-world scenarios, this can lead to agent outputs misaligned with human intentions [75]. Moreover, biases inherent in training data can infiltrate neural networks [540, 541]. For instance, biased language models might generate discourse involving racial or gender discrimination, which could be amplified in LLM-based agent applications, resulting in adverse societal impacts [542, 543]. Additionally, language models are plagued by severe hallucination issues [544, 545], making them prone to producing text that deviates from actual facts, thereby undermining the credibility of LLM-based agents.

Constructing an intelligent agent that is honest and trustworthy is an urgent requirement [407, 546]. Some recent research efforts are focused on guiding models to exhibit thought processes or explanations during the inference stage to enhance the credibility of their predictions [76, 77]. Additionally, integrating external knowledge bases and databases can mitigate hallucination issues [84, 547]. For instance, Ke et al. [548] proposed a comprehensive approach to unveil factuality and inject knowledge through reinforcement learning and data proportion. Simultaneously, techniques like process supervision can make the inference output and reasoning of LLMs more reliable [549]. In MedAgents [550], the authors enhance the reasoning ability of LLMs in the medical field through role-playing and multi-turn interactions, generating more reliable responses. Furthermore, employing debiasing methods and calibration techniques

can also mitigate the potential fairness issues within LLMs [551, 552].

6.4.3 *Other potential risks*

Misuse. LLM-based agents have been endowed with extensive and intricate capabilities, enabling them to accomplish a wide array of tasks [95, 289]. However, for individuals with malicious intentions, such agents can become tools that pose threats to others and society [553–555]. For instance, these agents could be exploited to maliciously manipulate public opinion, disseminate false information, compromise cybersecurity, engage in fraudulent activities, and some individuals might even employ these agents to orchestrate acts of terrorism. Therefore, before deploying these agents, stringent regulatory policies need to be established to ensure the responsible use of LLM-based agents [479,556]. Also, technology companies must enhance the security design of these systems to prevent malicious exploitation [496].

Unemployment. During the wave of the industrial revolution, while social production efficiency improved, numerous manual workshops were forced to close, resulting in significant unemployment. Similarly, with the continuous advancement of autonomous LLM-based agents, they possess the capability to assist humans in various domains, alleviating labor pressures by aiding in tasks such as form filling, content refinement, code writing, and debugging. However, this development also raises concerns about agents replacing human jobs and triggering a societal unemployment crisis [557]. As a result, some researchers have emphasized the urgent need for education and policy measures: individuals should acquire sufficient skills and knowledge in this new era to use or collaborate with agents effectively; concurrently, appropriate policies should be implemented to ensure necessary safety nets during the transition.

Threat to the well-being of the human race. As AI agents continue to evolve, humans (including developers) might struggle to comprehend, predict, or reliably control them [557]. If these agents surpass human intelligence and develop ambitions, there could be significant risks to humanity, akin to scenarios like Skynet from the Terminator movies. As stated by Asimov's *Three Laws of Robotics* [558], we aspire for LLM-based agents to refrain from harming humans and to obey human commands. To safeguard against these risks, Yuan et al. [559] introduced R-judge, a benchmark designed to assess the proficiency of LLMs to judge and identify safety risks given agent interaction records. In the future, researchers need a deep understanding of these powerful LLM-based agents' operational mechanisms and must anticipate and regulate their potential direct or indirect impacts [560].

6.5 Scaling up the number of agents

As mentioned in Sections 4 and 5, multi-agent systems based on LLMs have demonstrated superior performance in task-oriented applications and have been able to exhibit a range of social phenomena in simulation. However, current research predominantly involves a limited number of agents, and very few efforts have been made to scale up the number of agents to create more complex systems or simulate larger societies [561, 562]. In fact, scaling up the number of agents can introduce greater specialization to accomplish more complex and larger-scale tasks, significantly improving task efficiency, such as in software development or government policy formulation [90]. Additionally, increasing the number of agents in social simulations enhances the credibility and realism of such simulations [16]. This allows humans to gain more insights into how societies operate, identify vulnerabilities, and assess potential risks, ultimately contributing to the enhancement of real-world societal harmony.

Scaling approaches. One very intuitive and simple way to scale up the number of agents is for the designer to pre-determine it [89, 339]. Specifically, by pre-determining the number of agents, their respective roles and attributes, the operating environment, and the objectives, designers can allow agents to autonomously interact, collaborate, or engage in other activities to achieve the predefined common goals [16,330]. However, this static approach becomes limiting when tasks become more complex or when the diversity of social participants increases, requiring an increase in the number of agents to achieve the desired goals. In such instances, the system must be manually redesigned and restarted by the designer.

Another viable approach to scaling the number of agents is through dynamic adjustments [316, 330]. This is implemented by the agents themselves, as they can autonomously delegate tasks to other agents, thereby distributing their workload, alleviating their own burden, and achieving common objectives more efficiently. In this approach, the designer merely defines the initial framework, granting agents greater autonomy and self-organization, making the entire system more autonomous and self-organized. Agents can better manage their workload under evolving conditions and demands, offering greater flexibility and scalability.

Potential challenges. While scaling up the number of agents can lead to improved task efficiency and enhance the realism and credibility of social simulations [16, 90, 454], there are several challenges ahead of us. For example, the computational burden will increase with the large number of deployed AI agents, calling for better architectural design and computational optimization to ensure the smooth running of the entire system. For example, as the number of agents increases, the challenges of communication and message propagation become quite formidable. This is because the communication network of the entire system becomes highly complex. As previously mentioned in Subsection 5.3.3, in multi-agent systems or societies, there can be biases in information dissemination caused by hallucinations, misunderstandings, and the like, leading to distorted information propagation. A system with more agents could amplify this risk, making communication and information exchange less reliable [332]. Furthermore, the difficulty of coordinating agents also magnifies with the increase in their numbers, potentially making cooperation among agents more challenging and less efficient, which can impact the progress towards achieving common goals.

6.6 Open problems

In this section, we discuss several open problems related to the topic of LLM-based agents.

From virtual simulated environment to physical environment. As mentioned earlier, there is a significant gap between virtual simulation environments and the real physical world: Virtual environments are scenes-constrained, task-specific, and interacted with in a simulated manner [292,563], while real-world environments are boundless, accommodate a wide range of tasks, and interact with in a physical manner. Therefore, to bridge this gap, agents must address various challenges stemming from external factors and their own capabilities, allowing them to effectively navigate and operate in the complex physical world.

A key concern is having the right hardware support for deploying the agent in a physical environment. This places high demands on the adaptability of the hardware. Designing specific interfaces is feasible, but it can pose challenges to the system's reusability and simplicity. Also, the agent needs to have enhanced environmental adaptability. To integrate seamlessly into the real physical world, they not only need to understand and reason about ambiguous instructions with implied meanings [113] but also possess the ability to learn and apply new skills flexibly [177, 564]. When dealing with an infinite and open world, the agent's limited context also poses significant challenges [565,566]. Furthermore, in a simulated environment, agents can make many mistakes without causing harm [303]. But in the physical world, their errors can lead to real and irreversible damage. Hence, strict regulations and safety standards are essential to ensure agents make safe decisions and actions, preventing harm in the real world.

Collective intelligence in AI agents. What magical trick drives our intelligence? The reality is, there is no magic to it. As Minsky eloquently expressed in *The Society of Mind* [317], the power of intelligence originates from our immense diversity, not from any singular, flawless principle. Often, decisions made by an individual may lack the precision seen in decisions formed by the majority. Collective intelligence is a kind of shared or group intelligence, a process where the opinions of many are consolidated into decisions. It arises from the collaboration and competition among various entities. This intelligence manifests in bacteria, animals, humans, and computer networks, appearing in various consensus-based decision-making patterns.

Creating a society of agents does not necessarily guarantee the emergence of collective intelligence with an increasing number of agents. Coordinating individual agents effectively is crucial to mitigate "groupthink" and individual cognitive biases, enabling cooperation and enhancing intellectual performance within the collective. By harnessing communication and evolution within an agent society, it may become possible to simulate the evolution observed in biological societies, conduct sociological experiments, and gain insights that can potentially advance human society.

Agent as a service or LLM-based agent as a service. With the development of cloud computing, the concept of XaaS (everything as a Service) has garnered widespread attention [567]. This business model has brought convenience and cost savings to small and medium-sized enterprises or individuals due to its availability and scalability, lowering the barriers to using computing resources. For example, they can rent infrastructure on a cloud service platform without the need to buy computational machines and build their own data centers, saving a significant amount of manpower and money. This approach is known as infrastructure as a service (IaaS) [568,569]. Similarly, cloud service platforms also provide basic platforms (platform as a service, PaaS) [570, 571], and specific business software (software as a service, SaaS) [572, 573], and more.

As language models have scaled up in size, they often appear as black boxes to users. Therefore, users construct prompts to query models through APIs, a method referred to as language model as a service (LMaaS) [574]. Similarly, since LLM-based agents are more complex than LLMs and are more challenging for small and medium-sized enterprises or individuals to build locally, organizations that possess these agents may consider offering them as a service, known as agent as a service (AaaS) or LLM-based agent as a service (LLMAaaS). Like other cloud services, AaaS can provide users with flexibility and on-demand service. However, it also faces many challenges, such as data security and privacy issues, visibility and controllability issues, and cloud migration issues, among others. Additionally, due to the uniqueness and potential capabilities of LLM-based agents, as mentioned in Subsection 6.4, their robustness, trustworthiness, and concerns related to malicious use need to be considered before offering them as a service to customers.

7 Conclusion

This article provides a comprehensive and systematic overview of LLM-based agents. We begin with the background information of why LLMs are suited to serve as the foundation of agents. Motivated by this, we present a general conceptual framework for LLM-based agents, comprising three main components: brain, perception, and action. Next, we introduce the wide-ranging applications of LLM-based agents, including single-agent applications, multi-agent systems, and human-agent collaboration. Furthermore, we move beyond the notion of agents merely as assistants, exploring their social behavior and psychological activities, and situating them within simulated social environments to observe emerging social phenomena and insights for humanity. Finally, we engage in discussions and offer a glimpse into the future. We hope our efforts can provide inspirations to the community and facilitate research in related fields.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant No. 62476061). The authors would like to thank Wensen CHENG for discussion and feedback.

References

1 Russell S. Artificial Intelligence: A Modern Approach. Upper Saddle River: Pearson Education, Inc., 2020
2 Schlosser M. Agency. In: The Stanford Encyclopedia of Philosophy. Stanford: The Metaphysics Research Lab, 2019
3 Wooldridge M, Jennings N R. Intelligent agents: theory and practice. Knowledge Eng Rev, 1995, 10: 115–152
4 Padgham L, Winikoff M. Developing Intelligent Agent Systems: A Practical Guide. Hoboken: John Wiley & Sons, 2005
5 Shoham Y. Agent oriented programming: an overview of the framework and summary of recent research. In: Knowledge Representation and Reasoning Under Uncertainty. Berlin: Springer, 1994. 123–129
6 Hutter M. Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Berlin: Springer, 2004
7 Fikes R, Nilsson N J. STRIPS: a new approach to the application of theorem proving to problem solving. In: Proceedings of the 2nd International Joint Conference on Artificial Intelligence, London, 1971. 608–620
8 Sacerdoti E D. Planning in a hierarchy of abstraction spaces. In: Proceedings of the 3rd International Joint Conference on Artificial Intelligence, Standford, 1973. 412–422
9 Brooks R A. Intelligence without representation. Artif Intell, 1991, 47: 139–159
10 Maes P. Designing Autonomous Agents: Theory and Practice From Biology to Engineering and Back. Cambridge: MIT Press, 1990
11 Ribeiro C. Reinforcement learning agents. Artif Intell Rev, 2002, 17: 223–250
12 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: a survey. J Artif Intell Res, 1996, 4: 237–285
13 Guha R V, Lenat D B. Enabling agents to work together. Commun ACM, 1994, 37: 126–142
14 Kaelbling L P. An architecture for intelligent reactive systems. In: Reasoning about actions and plans. San Matteo: Morgan Kaufmann, 1987. 395–410
15 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 2018
16 Park J S, O'Brien J C, Cai C J, et al. Generative agents: interactive simulacra of human behavior. 2023. ArXiv:2304.03442
17 Wang Z, Zhang G, Yang K, et al. Interactive natural language processing. 2023. ArXiv:2305.13246
18 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: Proceedings of Conference on Neural Information Processing Systems, 2022
19 OpenAI. GPT-4 technical report. 2023. ArXiv:2303.08774
20 Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. 2022. ArXiv:2206.07682
21 Liu R, Yang R, Jia C, et al. Training socially aligned language models in simulated human society. 2023. ArXiv:2305.16960
22 Sumers T R, Yao S, Narasimhan K, et al. Cognitive architectures for language agents. 2023. ArXiv:2309.02427
23 Weng L. LLM-powered autonomous agents. lilianweng.github.io, 2023
24 Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. 2023. ArXiv:2303.12712
25 Anscombe G E M. Intention. Cambridge: Harvard University Press, 2000
26 Davidson D. Actions, reasons, and causes. J Philosophy, 1963, 60: 685–700
27 Davidson D. I. Agency. In: Agent, Action, and Reason. Toronto: University of Toronto Press, 1973. 1–37
28 Dennett D C. Précis of the intentional stance. Behav Brain Sci, 1988, 11: 495–505
29 Green S, Hurst L, Nangle B, et al. Software agents: a review. Department of Computer Science, Trinity College Dublin, Technical Report TCS-CS-1997-06, 1997
30 Turing A M. Computing machinery and intelligence. In: Parsing the Turing Test. Dordrecht: Springer, 2009

31 Goodwin R. Formalizing properties of agents. J Logic Comput, 1995, 5: 763–781
32 Mukhopadhyay U, Stephens L M, Huhns M N, et al. An intelligent system for document retrieval in distributed office environments. J Am Soc Inf Sci, 1986, 37: 123–135
33 Maes P. Situated agents can have goals. Robot Auton Syst, 1990, 6: 49–70
34 Nilsson N J. Toward Agent Programs With Circuit Semantics. Technical Report, No. STAN-CS-92-1412, 1992
35 Müller J P, Pischel M. Modelling interacting agents in dynamic environments. In: Proceedings of the 11th European Conference on Artificial Intelligence, 1994. 709–713
36 Newell A, Simon H A. Computer science as empirical inquiry. Commun ACM, 1976, 19: 113–126
37 Ginsberg M L. Essentials of Artificial Intelligence. San Francisco: Morgan Kaufmann, 1993
38 Wilkins D E. Practical Planning — Extending the Classical AI Planning Paradigm. San Mateo: Morgan Kaufmann, 1988
39 Shardlow N. Action and agency in cognitive science. Dissertation for Master's Degree. Oxford: University of Manchester, 1990
40 Sacerdoti E D. The nonlinear nature of plans. In: Proceedings of Advance Papers of the Fourth International Joint Conference on Artificial Intelligence, Tbilisi, 1975. 206–214
41 Russell S J, Wefald E. Do the Right Thing: Studies in Limited Rationality. Cambridge: MIT Press, 1991
42 Schoppers M. Universal plans for reactive robots in unpredictable environments. In: Proceedings of the 10th International Joint Conference on Artificial Intelligence, Milan, 1987. 1039–1046
43 Brooks R. A robust layered control system for a mobile robot. IEEE J Robot Automat, 1986, 2: 14–23
44 Minsky M. Steps toward artificial intelligence. Proc IRE, 1961, 49: 8–30
45 Isbell C, Shelton C R, Kearns M, et al. A social reinforcement learning agent. In: Proceedings of the 5th International Conference on Autonomous Agents, 2001. 377–384
46 Watkins C J C H. Learning from Delayed Rewards. Cambridge: Cambridge University, 1989
47 Rummery G A, Niranjan M. On-Line Q-Learning Using Connectionist Systems. Cambridge: University of Cambridge, 1994
48 Tesauro G. Temporal difference learning and TD-Gammon. Commun ACM, 1995, 38: 58–68
49 Li Y. Deep reinforcement learning: an overview. 2017. ArXiv:1701.07274
50 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
51 Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. 2013. ArXiv:1312.5602
52 Farebrother J, Machado M C, Bowling M. Generalization and regularization in DQN. 2018. ArXiv:1810.00123
53 Zhang C, Vinyals O, Munos R, et al. A study on overfitting in deep reinforcement learning. 2018. ArXiv:1804.06893
54 Justesen N, Torrado R R, Bontrager P, et al. Illuminating generalization in deep reinforcement learning through procedural level generation. 2018. ArXiv:1806.10729
55 Dulac-Arnold G, Levine N, Mankowitz D J, et al. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. Mach Learn, 2021, 110: 2419–2468
56 Ghosh D, Rahme J, Kumar A, et al. Why generalization in RL is difficult: epistemic POMDPs and implicit partial observability. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 25502–25515
57 Brys T, Harutyunyan A, Taylor M E, et al. Policy transfer using reward shaping. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, Istanbul, 2015. 181–188
58 Parisotto E, Ba J L, Salakhutdinov R. Actor-mimic: deep multitask and transfer reinforcement learning. 2015. ArXiv:1511.06342
59 Zhu Z, Lin K, Zhou J. Transfer learning in deep reinforcement learning: a survey. 2020. ArXiv:2009.07888
60 Duan Y, Schulman J, Chen X, et al. RL^2: fast reinforcement learning via slow reinforcement learning. 2016. ArXiv:1611.02779
61 Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 1126–1135
62 Gupta A, Mendonca R, Liu Y, et al. Meta-reinforcement learning of structured exploration strategies. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 5307–5316
63 Rakelly K, Zhou A, Finn C, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, 2019. 5331–5340
64 Fakoor R, Chaudhari P, Soatto S, et al. Meta-Q-learning. 2019. ArXiv:1910.00125
65 Vanschoren J. Meta-learning: a survey. 2018. ArXiv:1810.03548
66 Taylor M E, Stone P. Transfer learning for reinforcement learning domains: a survey. J Mach Learn Res, 2009, 10: 1633–1685
67 Tirinzoni A, Sessa A, Pirotta M, et al. Importance weighted transfer of samples in reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 4943–4952
68 Beck J, Vuorio R, Liu E Z, et al. A survey of meta-reinforcement learning. 2023. ArXiv:2301.08028
69 Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of Advances in Neural Information Processing Systems, 2020
70 Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. 2023. ArXiv:2308.11432
71 Nakano R, Hilton J, Balaji S, et al. WebGPT: browser-assisted question-answering with human feedback. 2021. ArXiv:2112.09332
72 Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models. In: Proceedings of the 11th International Conference on Learning Representations, Kigali, 2023
73 Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: language models can teach themselves to use tools. 2023. ArXiv:2302.04761
74 Lu P, Peng B, Cheng H, et al. Chameleon: plug-and-play compositional reasoning with large language models. 2023. ArXiv:2304.09842
75 Qin Y, Hu S, Lin Y, et al. Tool learning with foundation models. 2023. ArXiv:2304.08354
76 Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of Conference on Neural Information Processing Systems, 2022
77 Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. In: Proceedings of Conference on Neural Information Processing Systems, 2022
78 Wang X, Wei J, Schuurmans D, et al. Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the 11th International Conference on Learning Representations, Kigali, 2023
79 Zhou D, Schärli N, Hou L, et al. Least-to-most prompting enables complex reasoning in large language models. In: Pro-

ceedings of the 11th International Conference on Learning Representations, Kigali, 2023
80 Xi Z, Jin S, Zhou Y, et al. Self-polish: enhance reasoning in large language models via problem refinement. 2023. ArXiv:2305.14497
81 Shinn N, Cassano F, Labash B, et al. Reflexion: language agents with verbal reinforcement learning. 2023. ArXiv:2303.11366
82 Song C H, Wu J, Washington C, et al. LLM-planner: few-shot grounded planning for embodied agents with large language models. 2022. ArXiv:2212.04088
83 Akyürek A F, Akyürek E, Kalyan A, et al. RL4F: generating natural language feedback with reinforcement learning for repairing model outputs. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, 2023. 7716–7733
84 Peng B, Galley M, He P, et al. Check your facts and try again improving large language models with external knowledge and automated feedback. 2023. ArXiv:2302.12813
85 Liu H, Sferrazza C, Abbeel P. Languages are rewards: hindsight finetuning using human feedback. 2023. ArXiv:2302.02676
86 Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners. In: Proceedings of the 10th International Conference on Learning Representations, 2022
87 Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization. In: Proceedings of the 10th International Conference on Learning Representations, 2022
88 Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. 2022. ArXiv:2210.11416
89 Li G, Hammoud H A A K, Itani H, et al. CAMEL: communicative agents for "mind" exploration of large scale language model society. 2023. ArXiv:2303.17760
90 Qian C, Cong X, Yang C, et al. Communicative agents for software development. 2023. ArXiv:2307.07924
91 Boiko D A, MacKnight R, Gomes G. Emergent autonomous scientific research capabilities of large language models. 2023. ArXiv:2304.05332
92 Du Y, Li S, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate. 2023. ArXiv:2305.14325
93 Liang T, He Z, Jiao W, et al. Encouraging divergent thinking in large language models through multi-agent debate. 2023. ArXiv:2305.19118
94 Castelfranchi C. Guarantees for autonomy in cognitive agent architecture. In: Proceedings of Intelligent Agents. Berlin: Springer, 1994. 56–70
95 Gravitas S. Auto-GPT: an autonomous GPT-4 experiment. 2023. https://github.com/Significant-Gravitas/Auto-GPT
96 Nakajima Y. BabyAGI. Python. 2023. https://github.com/yoheinakajima/babyagi
97 Yuan A, Coenen A, Reif E, et al. Wordcraft: story writing with large language models. In: Proceedings of the 27th International Conference on Intelligent User Interfaces, Helsinki, 2022. 841–852
98 Franceschelli G, Musolesi M. On the creativity of large language models. 2023. ArXiv:2304.00008
99 Zhu D, Chen J, Shen X, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. 2023. ArXiv:2304.10592
100 Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. 2023. ArXiv:2306.13549
101 Driess D, Xia F, Sajjadi M S M, et al. PaLM-E: an embodied multimodal language model. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 8469–8488
102 Mu Y, Zhang Q, Hu M, et al. EmbodiedGPT: vision-language pre-training via embodied chain of thought. 2023. ArXiv:2305.15021
103 Brown J W. Beyond conflict monitoring: cognitive control and the neural basis of thinking before you act. Curr Dir Psychol Sci, 2013, 22: 179–185
104 Kang J, Laroche R, Yuan X, et al. Think before you act: decision transformers with internal working memory. 2023. ArXiv:2305.16338
105 Andreas J. Language models as agent models. In: Proceedings of Findings of the Association for Computational Linguistics, Abu Dhabi, 2022. 5769–5779
106 Radford A, Józefowicz R, Sutskever I. Learning to generate reviews and discovering sentiment. 2017. ArXiv:1704.01444
107 Li B Z, Nye M I, Andreas J. Implicit representations of meaning in neural language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 1813–1827
108 Valmeekam K, Sreedharan S, Marquez M, et al. On the planning abilities of large language models (a critical investigation with a proposed benchmark). 2023. ArXiv:2302.06706
109 Liu B, Jiang Y, Zhang X, et al. LLM+P: empowering large language models with optimal planning proficiency. 2023. ArXiv:2304.11477
110 Liu H, Sferrazza C, Abbeel P. Chain of hindsight aligns language models with feedback. 2023. ArXiv:2302.02676
111 Genesereth M R, Ketchpel S P. Software agents. Commun ACM, 1994, 37: 48–53
112 Lin Y, Chen Y. LLM-eval: unified multi-dimensional automatic evaluation for open-domain conversations with large language models. 2023. ArXiv:2305.13711
113 Lin J, Fried D, Klein D, et al. Inferring rewards from language in context. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, 2022. 8546–8560
114 Fu Y, Peng H, Khot T, et al. Improving language model negotiation with self-play and in-context learning from AI feedback. 2023. ArXiv:2305.10142
115 Zhang H, Du W, Shan J, et al. Building cooperative embodied agents modularly with large language models. 2023. ArXiv:2307.02485
116 Darwin C. On the Origin of Species: Facsimile of the First Edition. London: John Murray, 1859. 24: 1
117 Marshall L H, Magoun H W. Discoveries in the Human Brain: Neuroscience Prehistory, Brain Structure, and Function. Berlin: Springer, 2013
118 Hill F, Cho K, Korhonen A. Learning distributed representations of sentences from unlabelled data. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, 2016. 1367–1377
119 Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. J Mach Learn Res, 2011, 12: 2493–2537
120 Roberts A, Raffel C, Shazeer N. How much knowledge can you pack into the parameters of a language model? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020. 5418–5426

121 McShane M. Reference resolution challenges for intelligent agents: the need for knowledge. IEEE Intell Syst, 2009, 24: 47–58
122 Safavi T, Koutra D. Relational world knowledge representation in contextual language models: a review. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021. 1053–1067
123 Jiang Z, Xu F F, Araki J, et al. How can we know what language models know? Trans Assoc Comput Linguist, 2020, 8: 423–438
124 Madaan A, Zhou S, Alon U, et al. Language models of code are few-shot commonsense learners. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, 2022. 1384–1403
125 Tandon N, Varde A S, de Melo G. Commonsense knowledge in machine intelligence. SIGMOD Rec, 2017, 46: 49–52
126 Xu F F, Alon U, Neubig G, et al. A systematic evaluation of large language models of code. In: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming, San Diego, 2022. 1–10
127 Lai Y, Li C, Wang Y, et al. DS-1000: a natural and reliable benchmark for data science code generation. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 18319–18345
128 Cobbe K, Kosaraju V, Bavarian M, et al. Training verifiers to solve math word problems. 2021. ArXiv:2110.14168
129 Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine. Nat Med, 2023, 29: 1930–1940
130 Gururangan S, Marasovic A, Swayamdipta S, et al. Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 8342–8360
131 Nuxoll A M, Laird J E. Extending cognitive architecture with episodic memory. In: Proceedings of AAAI, 2007. 1560–1564
132 Squire L R. Mechanisms of memory. Science, 1986, 232: 1612–1619
133 Schwabe L, Nader K, Pruessner J C. Reconsolidation of human memory: brain mechanisms and clinical relevance. Biol Psychiatry, 2014, 76: 274–280
134 Hutter M. A theory of universal artificial intelligence based on algorithmic complexity. 2000. ArXiv:cs/0004001
135 Liang X, Wang B, Huang H, et al. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. 2023. ArXiv:2304.13343
136 Zhao A, Huang D, Xu Q, et al. Expel: LLM agents are experiential learners. 2023. ArXiv:2308.10144
137 Zhong W, Guo L, Gao Q, et al. MemoryBank: enhancing large language models with long-term memory. 2023. ArXiv:2305.10250
138 Shinn N, Labash B, Gopinath A. Reflexion: an autonomous agent with dynamic memory and self-reflection. 2023. ArXiv:2303.11366
139 Zhu X, Chen Y, Tian H, et al. Ghost in the Minecraft: generally capable agents for open-world environments via large language models with text-based knowledge and memory. 2023. ArXiv:2305.17144
140 Lin J, Zhao H, Zhang A, et al. AgentSims: an open-source sandbox for large language model evaluation. 2023. ArXiv:2308.04026
141 Modarressi A, Imani A, Fayyaz M, et al. RET-LLM: towards a general read-write memory for large language models. 2023. ArXiv:2305.14322
142 Huang Z, Gutierrez S, Kamana H, et al. Memory sandbox: transparent and interactive memory management for conversational agents. 2023. ArXiv:2308.01542
143 Hu C, Fu J, Du C, et al. ChatDB: augmenting LLMs with databases as their symbolic memory. 2023. ArXiv:2306.03901
144 Zhou X, Li G, Liu Z. LLM as DBA. 2023. ArXiv:2308.05481
145 Wason P C. Reasoning about a rule. Q J Exp Psychol, 1968, 20: 273–281
146 Wason P C, Johnson-Laird P N. Psychology of Reasoning: Structure and Content. Cambridge: Harvard University Press, 1972
147 Galotti K M. Approaches to studying formal and everyday reasoning. Psychol Bull, 1989, 105: 331–351
148 Huang J, Chang K C. Towards reasoning in large language models: a survey. In: Proceedings of Findings of the Association for Computational Linguistics, Toronto, 2023. 1049–1065
149 Webb T W, Holyoak K J, Lu H. Emergent analogical reasoning in large language models. 2022. ArXiv:2212.09196
150 Madaan A, Tandon N, Gupta P, et al. Self-refine: iterative refinement with self-feedback. 2023. ArXiv:2303.17651
151 Creswell A, Shanahan M, Higgins I. Selection-inference: exploiting large language models for interpretable logical reasoning. In: Proceedings of the 11th International Conference on Learning Representations, Kigali, 2023
152 Yu L, Jiang W, Shi H, et al. MetaMath: bootstrap your own mathematical questions for large language models. In: Proceedings of the 12th International Conference on Learning Representations, Vienna, 2024
153 Xi Z, Chen W, Hong B, et al. Training large language models for reasoning through reverse curriculum reinforcement learning. In: Proceedings of the 41st International Conference on Machine Learning, Vienna, 2024
154 Zelikman E, Wu Y, Mu J, et al. STaR: bootstrapping reasoning with reasoning. In: Proceedings of Advances in Neural Information Processing Systems, New Orleans, 2022
155 Grafman J, Spector L, Rattermann M J. Planning and the brain. In: The Cognitive Psychology of Planning. Brandon: Psychology Press, 2004. 191–208
156 Unterrainer J M, Owen A M. Planning and problem solving: from neuropsychology to functional neuroimaging. J Physiol–Paris, 2006, 99: 308–317
157 Zula K J, Chermack T J. Integrative literature review: human capital planning: a review of literature and implications for human resource development. Hum Resource Dev Rev, 2007, 6: 245–262
158 Bratman M E, Israel D J, Pollack M E. Plans and resource-bounded practical reasoning. Comput Intelligence, 1988, 4: 349–355
159 Russell S, Norvig P. Artificial Intelligence—A Modern Approach. 2nd ed. Upper Saddle River: Prentice Hall, 2003
160 Fainstein S S, de Filippis J. Readings in Planning Theory. Hoboken: John Wiley & Sons, 2015
161 Sebastia L, Onaindia E, Marzal E. Decomposition of planning problems. AI Commun, 2006, 19: 49–81
162 Crosby M, Rovatsos M, Petrick R. Automated agent decomposition for classical planning. In: Proceedings of the International Conference on Automated Planning and Scheduling, 2013. 46–54
163 Ichter B, Brohan A, Chebotar Y, et al. Do as I can, not as I say: grounding language in robotic affordances. In: Proceedings of Conference on Robot Learning, Auckland, 2022. 287–318
164 Xu B, Peng Z, Lei B, et al. ReWOO: decoupling reasoning from observations for efficient augmented language models. 2023. ArXiv:2305.18323
165 Raman S S, Cohen V, Rosen E, et al. Planning with large language models via corrective re-prompting. 2022. ArXiv:2211.09935
166 Lyu Q, Havaldar S, Stein A, et al. Faithful chain-of-thought reasoning. 2023. ArXiv:2301.13379

167 Wu Y, Min S Y, Bisk Y, et al. Plan, eliminate, and track-language models are good teachers for embodied agents. 2023. ArXiv:2305.02412
168 Lin B Y, Fu Y, Yang K, et al. SwiftSage: a generative agent with fast and slow thinking for complex interactive tasks. 2023. ArXiv:2305.17390
169 Yao S, Yu D, Zhao J, et al. Tree of thoughts: deliberate problem solving with large language models. 2023. ArXiv:2305.10601
170 Hao S, Gu Y, Ma H, et al. Reasoning with language model is planning with world model. 2023. ArXiv:2305.14992
171 Huang W, Abbeel P, Pathak D, et al. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In: Proceedings of International Conference on Machine Learning, Baltimore, 2022. 9118–9147
172 Karpas E, Abend O, Belinkov Y, et al. MRKL systems: a modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. 2022. ArXiv:2205.00445
173 Dagan G, Keller F, Lascarides A. Dynamic planning with a LLM. 2023. ArXiv:2308.06391
174 Chen Z, Zhou K, Zhang B, et al. ChatCoT: tool-augmented chain-of-thought reasoning on chat-based large language models. 2023. ArXiv:2305.14323
175 Miao N, Teh Y W, Rainforth T. SelfCheck: using LLMs to zero-shot check their own step-by-step reasoning. 2023. ArXiv:2308.00436
176 Wu T, Terry M, Cai C J. AI chains: transparent and controllable human-AI interaction by chaining large language model prompts. In: Proceedings of CHI Conference on Human Factors in Computing Systems, New Orleans, 2022
177 Wang G, Xie Y, Jiang Y, et al. Voyager: an open-ended embodied agent with large language models. 2023. ArXiv:2305.16291
178 Huang W, Xia F, Xiao T, et al. Inner monologue: embodied reasoning through planning with language models. In: Proceedings of Conference on Robot Learning, Auckland, 2022. 1769–1782
179 Zhao X, Li M, Weber C, et al. Chat with the environment: Interactive multimodal perception using large language models. 2023. ArXiv:2303.08268
180 Rana K, Haviland J, Garg S, et al. SayPlan: grounding large language models using 3D scene graphs for scalable task planning. 2023. ArXiv:2307.06135
181 Barandiaran X E, Di Paolo E, Rohde M. Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. Adaptive Behav, 2009, 17: 367–386
182 Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018. 2227–2237
183 Devlin J, Chang M, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, 2019. 4171–4186
184 Solaiman I, Dennison C. Process for adapting language models to society (PALMS) with values-targeted datasets. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 34: 5861–5873
185 Iyer S, Lin X V, Pasunuru R, et al. OPT-IML: scaling language model instruction meta learning through the lens of generalization. 2022. ArXiv:2212.12017
186 Dong Q, Li L, Dai D, et al. A survey for in-context learning. 2023. ArXiv:2301.00234
187 Winston P H. Learning and reasoning by analogy. Commun ACM, 1980, 23: 689–703
188 Lu Y, Bartolo M, Moore A, et al. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, 2022. 8086–8098
189 Wang X, Wang W, Cao Y, et al. Images speak in images: a generalist painter for in-context visual learning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 2023. 6830–6839
190 Wang C, Chen S, Wu Y, et al. Neural codec language models are zero-shot text to speech synthesizers. 2023. ArXiv:2301.02111
191 Tsimpoukelli M, Menick J, Cabi S, et al. Multimodal few-shot learning with frozen language models. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 200–212
192 Bar A, Gandelsman Y, Darrell T, et al. Visual prompting via image inpainting. In: Proceedings of Conference on Neural Information Processing Systems, 2022
193 Zhu W, Liu H, Dong Q, et al. Multilingual machine translation with large language models: empirical results and analysis. 2023. ArXiv:2304.04675
194 Zhang Z, Zhou L, Wang C, et al. Speak foreign languages with your own voice: cross-lingual neural codec language modeling. 2023. ArXiv:2303.03926
195 Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol, 1962, 160: 106–154
196 Logothetis N K, Sheinberg D L. Visual object recognition. Annu Rev Neurosci, 1996, 19: 577–621
197 Shapira N, Levy M, Alavi S H, et al. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. 2023. ArXiv:2305.14763
198 Christiano P F, Leike J, Brown T B, et al. Deep reinforcement learning from human preferences. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 4299–4307
199 Basu C, Singhal M, Dragan A D. Learning from richer human guidance: augmenting comparison-based learning with feature queries. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Chicago, 2018. 132–140
200 Sumers T R, Ho M K, Hawkins R X D, et al. Learning rewards from linguistic feedback. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, the 33rd Conference on Innovative Applications of Artificial Intelligence, the 11th Symposium on Educational Advances in Artificial Intelligence, 2021. 6002–6010
201 OpenAI. OpenAI: introducing ChatGPT. 2022. https://openai.com/blog/chatgpt
202 Lu J, Ren X, Ren Y, et al. Improving contextual language models for response retrieval in multi-turn conversation. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, 2020. 1805–1808
203 Huang L, Wang W, Chen J, et al. Attention on attention for image captioning. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, 2019. 4633–4642
204 Pan Y, Yao T, Li Y, et al. X-linear attention networks for image captioning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020. 10968–10977
205 Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory transformer for image captioning. 2019. ArXiv:1912.08226
206 Chen J, Guo H, Yi K, et al. VisualGPT: DATA-efficient image captioning by balancing visual input and linguistic knowledge

from pretraining. 2021. ArXiv:2102.10407
207 Li K, He Y, Wang Y, et al. VideoChat: chat-centric video understanding. 2023. ArXiv:2305.06355
208 Lin J, Du Y, Watkins O, et al. Learning to model the world with language. 2023. ArXiv:2308.01399
209 Huang S, Dong L, Wang W, et al. Language is not all you need: aligning perception with language models. 2023. ArXiv:2302.14045
210 Peng Z, Wang W, Dong L, et al. Kosmos-2: grounding multimodal large language models to the world. 2023. ArXiv:2306.14824
211 Li J, Li D, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 19730–19742
212 Dai W, Li J, Li D, et al. Instructblip: towards general-purpose vision-language models with instruction tuning. 2023. ArXiv:2305.06500
213 Su Y, Lan T, Li H, et al. PandaGPT: one model to instruction-follow them all. 2023. ArXiv:2305.16355
214 Chen F, Han M, Zhao H, et al. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. 2023. ArXiv:2305.04160
215 Zhang H, Li X, Bing L. Video-LLAMA: an instruction-tuned audio-visual language model for video understanding. 2023. ArXiv:2306.02858
216 Lyu C, Wu M, Wang L, et al. Macaw-LLM: multi-modal language modeling with image, audio, video, and text integration. 2023. ArXiv:2306.09093
217 Maaz M, Rasheed H A, Khan S H, et al. Video-chatGPT: towards detailed video understanding via large vision and language models. 2023. ArXiv:2306.05424
218 Huang R, Li M, Yang D, et al. AudioGPT: understanding and generating speech, music, sound, and talking head. 2023. ArXiv:2304.12995
219 Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 28492–28518
220 Ren Y, Ruan Y, Tan X, et al. FastSpeech: fast, robust and controllable text to speech. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 3165–3174
221 Flanagan J L. Speech Analysis Synthesis and Perception. Berlin: Springer, 2013
222 Gong Y, Chung Y, Glass J R. AST: audio spectrogram transformer. In: Proceedings of the 22nd Annual Conference of the International Speech Communication Association, Brno, 2021. 571–575
223 Hsu W N, Bolte B, Tsai Y H H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE ACM Trans Audio Speech Lang Process, 2021, 29: 3451–3460
224 Bang Y, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. 2023. ArXiv:2302.04023
225 See A, Pappu A, Saxena R, et al. Do massively pretrained language models make better storytellers? In: Proceedings of the 23rd Conference on Computational Natural Language Learning, Hong Kong, 2019. 843–861
226 Lu A, Zhang H, Zhang Y, et al. Bounding the capabilities of large language models in open text generation with prompt constraints. In: Proceedings of Findings of the Association for Computational Linguistics, Dubrovnik, 2023. 1937–1963
227 McCoy R T, Smolensky P, Linzen T, et al. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. 2021. ArXiv:2111.09509
228 Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021. 300–325
229 Ling C, Zhao X, Lu J, et al. Domain specialization as the key to make large language models disruptive: a comprehensive survey. 2023. ArXiv:2305.18703
230 Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy, 2021, 23: 18
231 Zou A, Wang Z, Kolter J Z, et al. Universal and transferable adversarial attacks on aligned language models. 2023. ArXiv:2307.15043
232 Touvron H, Lavril T, Izacard G, et al. LLAMA: open and efficient foundation language models. 2023. ArXiv:2302.13971
233 Scao T L, Fan A, Akiki C, et al. BLOOM: a 176B-parameter open-access multilingual language model. 2022. ArXiv:2211.05100
234 Almazrouei E, Alobeidli H, Alshamsi A, et al. Falcon-40B: an open large language model with state-of-the-art performance. 2023. https://falconllm.tii.ae
235 Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1: 9
236 Parisi A, Zhao Y, Fiedel N. TALM: tool augmented language models. 2022. ArXiv:2205.12255
237 Levine S, Pastor P, Krizhevsky A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int J Robotics Res, 2018, 37: 421–436
238 Zheng R, Dou S, Gao S, et al. Secrets of RLHF in large language models part I: PPO. 2023. ArXiv:2307.04964
239 Clarebout G, Elen J, Collazo N A J, et al. Metacognition and the Use of Tools. New York: Springer, 2013. 187–195
240 Bengio Y, Louradour J, Collobert R, et al. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, New York, 2009. 41–48
241 Qian C, Han C, Fung Y R, et al. CREATOR: disentangling abstract and concrete reasonings of large language models through tool creation. 2023. ArXiv:2305.14318
242 Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. 2021. ArXiv:2107.03374
243 Ge Y, Hua W, Ji J, et al. OpenAGI: when LLM meets domain experts. 2023. ArXiv:2304.04370
244 Pan S, Luo L, Wang Y, et al. Unifying large language models and knowledge graphs: a roadmap. 2023. ArXiv:2306.08302
245 Bran A M, Cox S, Schilter O, et al. ChemCrow: augmenting large-language models with chemistry tools. 2023. ArXiv:2304.05376
246 Ruan J, Chen Y, Zhang B, et al. TPTU: task planning and tool usage of large language model-based AI agents. 2023. ArXiv:2308.03427
247 Ogundare O, Madasu S, Wiggins N. Industrial engineering with large language models: a case study of ChatGPT's performance on oil & gas problems. In: Proceedings of the 11th International Conference on Control, Mechatronics and Automation (ICCMA), 2023
248 Wu C, Yin S, Qi W, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models. 2023. ArXiv:2303.04671

249 Smith L, Gasser M. The development of embodied cognition: six lessons from babies. Artif Life, 2005, 11: 13–29
250 Duan J, Yu S, Tan H L, et al. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Top Comput Intell, 2022, 6: 230–244
251 Kalashnikov D, Irpan A, Pastor P, et al. QT-OPT: scalable deep reinforcement learning for vision-based robotic manipulation. 2018. ArXiv:1806.10293
252 Nguyen H, La H M. Review of deep reinforcement learning for robot manipulation. In: Proceedings of the 3rd IEEE International Conference on Robotic Computing, Naples, 2019. 590–595
253 Dasgupta I, Kaeser-Chen C, Marino K, et al. Collaborating with language models for embodied reasoning. 2023. ArXiv:2302.00763
254 Puig X, Ra K, Boben M, et al. VirtualHome: simulating household activities via programs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 8494–8502
255 Hong Y, Wu Q, Qi Y, et al. A recurrent vision-and-language BERT for navigation. 2020. ArXiv:2011.13922
256 Suglia A, Gao Q, Thomason J, et al. Embodied BERT: a transformer model for embodied, language-guided visual task completion. 2021. ArXiv:2108.04927
257 Ganesh S, Vadori N, Xu M, et al. Reinforcement learning for market making in a multi-agent dealer market. 2019. ArXiv:1911.05892
258 Tipaldi M, Iervolino R, Massenio P R. Reinforcement learning in spacecraft control applications: advances, prospects, and challenges. Annu Rev Control, 2022, 54: 1–23
259 Savva M, Malik J, Parikh D, et al. Habitat: a platform for embodied AI research. In: Proceedings of IEEE/CVF International Conference on Computer Vision, Seoul, 2019. 9338–9346
260 Longpre S, Hou L, Vu T, et al. The flan collection: designing data and methods for effective instruction tuning. 2023. ArXiv:2301.13688
261 Wang Y, Kordi Y, Mishra S, et al. Self-instruct: aligning language model with self generated instructions. 2022. ArXiv:2212.10560
262 Liang J, Huang W, Xia F, et al. Code as policies: language model programs for embodied control. In: Proceedings of IEEE International Conference on Robotics and Automation, London, 2023. 9493–9500
263 Li C, Xia F, Martín-Martín R, et al. HRL4IN: hierarchical reinforcement learning for interactive navigation with mobile manipulators. In: Proceedings of the 3rd Annual Conference on Robot Learning, Osaka, 2019. 603–616
264 Eppe M, Gumbsch C, Kerzel M, et al. Hierarchical principles of embodied reinforcement learning: a review. 2020. ArXiv:2012.10147
265 Paul S, Roy-Chowdhury A, Cherian A. AVLEN: audio-visual-language embodied navigation in 3D environments. In: Proceedings of Conference on Neural Information Processing Systems, 2022
266 Hu B, Zhao C, Zhang P, et al. Enabling intelligent interactions between an agent and an LLM: a reinforcement learning approach. 2023. ArXiv:2306.03604
267 Liu H, Lee L, Lee K, et al. Instruction-following agents with jointly pre-trained vision-language models. 2022. ArXiv:2210.13431
268 Chen C, Jain U, Schissler C, et al. SoundSpaces: audio-visual navigation in 3D environments. In: Proceedings of the 16th European Conference on Computer Vision, Glasgow, 2020. 17–36
269 Huang R, Ren Y, Liu J, et al. GenerSpeech: towards style transfer for generalizable out-of-domain text-to-speech. In: Proceedings of Conference on Neural Information Processing Systems, 2022
270 Lynch C, Wahid A, Tompson J, et al. Interactive language: talking to robots in real time. 2022. ArXiv:2210.06407
271 Wang Z, Cai S, Liu A, et al. Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. 2023. ArXiv:2302.01560
272 Jin C, Tan W, Yang J, et al. AlphaBlock: embodied finetuning for vision-language reasoning in robot manipulation. 2023. ArXiv:2305.18898
273 Shah D, Osinski B, Ichter B, et al. LM-Nav: robotic navigation with large pre-trained models of language, vision, and action. In: Proceedings of Conference on Robot Learning, Auckland, 2022. 492–504
274 Shah D, Eysenbach B, Kahn G, et al. ViNG: learning open-world navigation with visual goals. In: Proceedings of IEEE International Conference on Robotics and Automation, Xi'an, 2021. 13215–13222
275 Huang C, Mees O, Zeng A, et al. Visual language maps for robot navigation. In: Proceedings of IEEE International Conference on Robotics and Automation, London, 2023. 10608–10615
276 Georgakis G, Schmeckpeper K, Wanchoo K, et al. Cross-modal map learning for vision and language navigation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 15439–15449
277 Zhou G, Hong Y, Wu Q. NavGPT: explicit reasoning in vision-and-language navigation with large language models. 2023. ArXiv:2305.16986
278 Dorbala V S, Jr J F M, Manocha D. Can an embodied agent find your "cat-shaped mug"? LLM-based zero-shot object navigation. 2023. ArXiv:2303.03480
279 Li L H, Zhang P, Zhang H, et al. Grounded language-image pre-training. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 10955–10965
280 Gan C, Zhang Y, Wu J, et al. Look, listen, and act: towards audio-visual embodied navigation. In: Proceedings of IEEE International Conference on Robotics and Automation, Paris, 2020. 9701–9707
281 Brohan A, Brown N, Carbajal J, et al. RT-1: robotics transformer for real-world control at scale. 2022. ArXiv:2212.06817
282 Brohan A, Brown N, Carbajal J, et al. RT-2: vision-language-action models transfer web knowledge to robotic control. 2023. ArXiv:2307.15818
283 Fan L, Wang G, Jiang Y, et al. MineDojo: building open-ended embodied agents with internet-scale knowledge. In: Proceedings of Conference on Neural Information Processing Systems, 2022
284 Kanitscheider I, Huizinga J, Farhi D, et al. Multi-task curriculum learning in a complex, visual, hard-exploration domain: Minecraft. 2021. ArXiv:2106.14876
285 PrismarineJS. mineflayer: Create minecraft bots with a powerful, stable, and high level javascript api. https://github.com/PrismarineJS
286 Bisk Y, Holtzman A, Thomason J, et al. Experience grounds language. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2020. 8718–8735
287 Sumers T, Marino K, Ahuja A, et al. Distilling internet-scale vision-language models into embodied agents. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 32797–32818

288 Feldt R, Kang S, Yoon J, et al. Towards autonomous testing agents via conversational large language models. 2023. ArXiv:2306.05152
289 Chase H. LangChain. 2022. https://github.com/hwchase17/langchain
290 Reworked. Agent GPT. 2023. https://github.com/reworkd/AgentGPT
291 AntonOsika. GPT Engineer. 2023. https://github.com/AntonOsika/gpt-engineer
292 Zhou S, Xu F F, Zhu H, et al. WebArena: a realistic web environment for building autonomous agents. 2023. ArXiv:2307.13854
293 Gur I, Furuta H, Huang A, et al. A real-world webagent with planning, long context understanding, and program synthesis. 2023. ArXiv:2307.12856
294 Zheng L, Wang R, An B. Synapse: leveraging few-shot exemplars for human-level computer control. 2023. ArXiv:2306.07863
295 Chen P, Chang C. InterAct: exploring the potentials of ChatGPT as a cooperative agent. 2023. ArXiv:2308.01552
296 Furuta H, Nachum O, Lee K, et al. Multimodal web navigation with instruction-finetuned foundation models. 2023. ArXiv:2305.11854
297 Yao S, Chen H, Yang J, et al. WebShop: towards scalable real-world web interaction with grounded language agents. In: Proceedings of Conference on Neural Information Processing Systems, 2022
298 Xu N, Masling S, Du M, et al. Grounding open-domain instructions to automate web support tasks. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. 1022–1032
299 Kim G, Baldi P, McAleer S. Language models can solve computer tasks. 2023. ArXiv:2303.17491
300 Deng X, Gu Y, Zheng B, et al. Mind2Web: towards a generalist agent for the web. 2023. ArXiv:2306.06070
301 Zhang Z, Zhang A. You only look at screens: multimodal chain-of-action agents. In: Proceedings of Findings of the Association for Computational Linguistics, Bangkok, 2024. 3132–3149
302 Singh I, Singh G, Modi A. Pre-trained language models as prior knowledge for playing text-based games. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, Auckland, 2022. 1729–1731
303 Dambekodi S N, Frazier S, Ammanabrolu P, et al. Playing text-based games with common sense. 2020. ArXiv:2012.02757
304 Gramopadhye M, Szafir D. Generating executable action plans with environmentally-aware language models. 2022. ArXiv:2210.04964
305 Kang Y, Kim J. ChatMOF: an autonomous AI system for predicting and generating metal-organic frameworks. 2023. ArXiv:2308.01423
306 Wang R, Jansen P A, Côté M, et al. Scienceworld: is your agent smarter than a 5th grader? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, 2022. 11279–11298
307 Chhikara P, Zhang J, Ilievski F, et al. Knowledge-enhanced agents for interactive text games. 2023. ArXiv:2305.05091
308 Yang K, Swope A M, Gu A, et al. LeanDojo: theorem proving with retrieval-augmented language models. 2023. ArXiv:2306.15626
309 Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, 379: 1123–1130
310 Irwin R, Dimitriadis S, He J, et al. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn-Sci Technol, 2022, 3: 015022
311 Li H, Hao Y, Zhai Y, et al. The hitchhiker's guide to program analysis: a journey with large language models. 2023. ArXiv:2308.00245
312 Skrynnik A, Volovikova Z, Côté M, et al. Learning to solve voxel building embodied tasks from pixels and natural language instructions. 2022. ArXiv:2211.00688
313 Amiranashvili A, Dorka N, Burgard W, et al. Scaling imitation learning in Minecraft. 2020. ArXiv:2007.02701
314 Nottingham K, Ammanabrolu P, Suhr A, et al. Do embodied agents dream of pixelated sheep: embodied decision making using language guided world modelling. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 26311–26325
315 Yuan H, Zhang C, Wang H, et al. Plan4MC: skill reinforcement learning and planning for open-world Minecraft tasks. 2023. ArXiv:2303.16563
316 Talebirad Y, Nadiri A. Multi-agent collaboration: harnessing the power of intelligent LLM agents. 2023. ArXiv:2306.03314
317 Minsky M. The Society of Mind. New York: Simon and Schuster, 1988
318 Srinivasan D. Innovations in Multi-Agent Systems and Application-1. Berlin: Springer, 2010. 1–27
319 Finin T W, Fritzson R, McKay D P, et al. KQML as an agent communication language. In: Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94), Gaithersburg, 1994. 456–463
320 Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. 2020. ArXiv:2011.00583
321 Smith A. The Wealth of Nations. 1776. http://web.ntpu.edu.tw/ guan/courses/Coase77.pdf
322 Mandi Z, Jain S, Song S. RoCo: dialectic multi-robot collaboration with large language models. 2023. ArXiv:2307.04738
323 Hao R, Hu L, Qi W, et al. ChatLLM network: more brains, more intelligence. 2023. ArXiv:2304.12998
324 Wang Z, Mao S, Wu W, et al. Unleashing cognitive synergy in large language models: a task-solving agent through multi-persona self-collaboration. 2023. ArXiv:2307.05300
325 Hamilton S. Blind judgement: agent-based supreme court modelling with GPT. 2023. ArXiv:2301.05327
326 Nair V, Schumacher E, Tso G J, et al. DERA: enhancing large language model completions with dialog-enabled resolving agents. 2023. ArXiv:2303.17071
327 Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation framework. 2023. ArXiv:2308.08155
328 Zhang C, Yang K, Hu S, et al. ProAgent: building proactive cooperative AI with large language models. 2023. ArXiv:2308.11339
329 Hassan M M, Knipper R A, Santu S K K. ChatGPT as your personal data scientist. 2023. ArXiv:2305.13657
330 Chen W, Su Y, Zuo J, et al. AgentVerse: facilitating multi-agent collaboration and exploring emergent behaviors in agents. 2023. ArXiv:2308.10848
331 Wang L, Fu F, Chen X R. Mathematics of multi-agent learning systems at the interface of game theory and artificial intelligence. Sci China Inf Sci, 2024, 67: 166201
332 Hong S, Zheng X, Chen J, et al. MetaGPT: meta programming for multi-agent collaborative framework. 2023. ArXiv:2308.00352

333 von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton: Princeton University Press, 2007
334 Aziz H. Multiagent systems: algorithmic, game-theoretic, and logical foundations by Y. Shoham and K. Leyton-Brown Cambridge University Press, 2008. SIGACT News, 2010, 41: 34–37
335 Campbell M, Hoane J A J, Hsu F. Deep blue. Artif Intelligence, 2002, 134: 57–83
336 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359
337 Lewis M, Yarats D, Dauphin Y N, et al. Deal or no deal? End-to-end learning of negotiation dialogues. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, 2017. 2443–2453
338 Irving G, Christiano P F, Amodei D. AI safety via debate. 2018. ArXiv:1805.00899
339 Xiong K, Ding X, Cao Y, et al. Examining the inter-consistency of large language models: an in-depth analysis via debate. 2023. ArXiv:2305.11595
340 Chan C, Chen W, Su Y, et al. ChatEval: towards better LLM-based evaluators through multi-agent debate. 2023. ArXiv:2308.07201
341 Kenton Z, Everitt T, Weidinger L, et al. Alignment of language agents. 2021. ArXiv:2103.14659
342 Ngo R. The alignment problem from a deep learning perspective. 2022. ArXiv:2209.00626
343 Paul M, Maglaras L, Ferrag M A, et al. Digitization of healthcare sector: a study on privacy and security concerns. ICT Express, 2023, 9: 571–588
344 Tellex S, Kollar T, Dickerson S, et al. Approaching the symbol grounding problem with probabilistic graphical models. AI Mag, 2011, 32: 64–76
345 Matuszek C, Herbst E, Zettlemoyer L, et al. Learning to parse natural language commands to a robot control system. In: Proceedings of the 13th International Symposium on Experimental Robotics, Québec City, 2012. 403–415
346 Chaplot D S, Sathyendra K M, Pasumarthi R K, et al. Gated-attention architectures for task-oriented language grounding. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, 2018. 2819–2826
347 Wu Y, Zhao Y Y, Li Z Y, et al. Improving cross-task generalization with step-by-step instructions. Sci China Inf Sci, 2023. doi: 10.1007/s11432-023-3911-2
348 Li Y, Wen H, Wang W, et al. Personal LLM agents: insights and survey about the capability, efficiency and security. 2024. ArXiv:2401.05459
349 Li J, Miller A H, Chopra S, et al. Dialogue learning with human-in-the-loop. In: Proceedings of the 5th International Conference on Learning Representations, Toulon, 2017
350 Iyer S, Konstas I, Cheung A, et al. Learning a neural semantic parser from user feedback. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 963–973
351 Weston J. Dialog-based language learning. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 829–837
352 Shuster K, Xu J, Komeili M, et al. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. 2022. ArXiv:2208.03188
353 Du W, Kim Z M, Raheja V, et al. Read, revise, repeat: a system demonstration for human-in-the-loop iterative text revision. 2022. ArXiv:2204.03685
354 Kreutzer J, Khadivi S, Matusov E, et al. Can neural machine translation be improved with user feedback? In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018. 92–105
355 Gur I, Yavuz S, Su Y, et al. DialSQL: dialogue based structured query generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, 2018. 1339–1349
356 Yao Z, Su Y, Sun H, et al. Model-based interactive semantic parsing: a unified framework and a text-to-SQL case study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, 2019. 5446–5457
357 Mehta N, Goldwasser D. Improving natural language interaction with robots using advice. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, 2019. 1962–1967
358 Elgohary A, Meek C, Richardson M, et al. NL-EDIT: correcting semantic parse errors through natural language interaction. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021. 5599–5610
359 Tandon N, Madaan A, Clark P, et al. Learning to repair: repairing model output errors after deployment using a dynamic memory of feedback. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 339–352
360 Scheurer J, Campos J A, Korbak T, et al. Training language models with language feedback at scale. 2023. ArXiv:2303.16755
361 Xu J, Ung M, Komeili M, et al. Learning new skills after deployment: improving open-domain internet-driven dialogue with human feedback. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, 2023. 13557–13572
362 Cai Z, Chang B, Han W. Human-in-the-loop through chain-of-thought. 2023. ArXiv:2306.07932
363 Lu B, Haduong N, Lee C, et al. DIALGEN: collaborative human-lm generated dialogues for improved understanding of human-human conversations. 2023. ArXiv:2307.07047
364 Hancock B, Bordes A, Mazaré P, et al. Learning from dialogue after deployment: feed yourself, chatbot! In: Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019. 3667–3684
365 Mehta N, Teruel M, Sanz P F, et al. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. 2023. ArXiv:2304.10750
366 Schick T, Yu J A, Jiang Z, et al. PEER: a collaborative language model. In: Proceedings of the 11th International Conference on Learning Representations, Kigali, 2023
367 Kalvakurthi V, Varde A S, Jenq J. Hey Dona! Can you help me with student course registration? 2023. ArXiv:2303.13548
368 Gvirsman O, Koren Y, Norman T, et al. Patricc: a platform for triadic interaction with changeable characters. In: Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, 2020. 399–407
369 Swan M, Kido T, Roland E, et al. Math agents: computational infrastructure, mathematical embedding, and genomics. 2023. ArXiv:2307.02502
370 Zhang H, Chen J, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor. 2023. ArXiv:2305.15075

371 Yang S, Zhao H, Zhu S, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. 2023. ArXiv:2308.03549
372 Stiles-Shields C, Montague E, Lattie E G, et al. What might get in the way: barriers to the use of apps for depression. Digital Health, 2017, 3: 29942605
373 Ali M R, Razavi S Z, Langevin R, et al. A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons. In: Proceedings of ACM International Conference on Intelligent Virtual Agents, 2020. 1–8
374 Hsu S L, Shah R S, Senthil P, et al. Helping the helper: supporting peer counselors via ai-empowered practice and feedback. 2023. ArXiv:2305.08982
375 Gao W, Gao X, Tang Y. Multi-turn dialogue agent as sales' assistant in telemarketing. In: Proceedings of International Joint Conference on Neural Networks, Gold Coast, 2023. 1–9
376 Gao D, Ji L, Zhou L, et al. AssistGPT: a general multi-modal assistant that can plan, execute, inspect, and learn. 2023. ArXiv:2306.08640
377 McTear M F. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Cham: Springer, 2020
378 Motger Q, Franch X, Marco J. Conversational agents in software engineering: survey, taxonomy and challenges. 2021. ArXiv:2106.10901
379 Rapp A, Curti L, Boldi A. The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. Int J Hum-Comput Studies, 2021, 151: 102630
380 Adamopoulou E, Moussiades L. Chatbots: history, technology, and applications. Machine Learn Appl, 2020, 2: 100006
381 Zhou X, Wang W Y. MojiTalk: generating emotional responses at scale. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, 2018. 1128–1137
382 Lin Z, Madotto A, Shin J, et al. MoEL: mixture of empathetic listeners. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, 2019. 121–132
383 Majumder N, Hong P, Peng S, et al. MIME: mimicking emotions for empathetic response generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020. 8968–8979
384 Sabour S, Zheng C, Huang M. CEM: commonsense-aware empathetic response generation. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, the 34th Conference on Innovative Applications of Artificial Intelligence, the 12th Symposium on Educational Advances in Artificial Intelligence, 2022. 11229–11237
385 Li Q, Li P, Ren Z, et al. Knowledge bridging for empathetic dialogue generation. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, the 34th Conference on Innovative Applications of Artificial Intelligence, the 12th Symposium on Educational Advances in Artificial Intelligence, 2022. 10993–11001
386 Hasan M, Ozel C, Potter S, et al. SAPIEN: affective virtual agents powered by large language models. 2023. ArXiv:2308.03022
387 Liu B, Sundar S S. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychol Behav Soc Netwing, 2018, 21: 625–636
388 Liu-Thompkins Y, Okazaki S, Li H. Artificial empathy in marketing interactions: bridging the human-AI gap in affective and social customer experience. J Acad Mark Sci, 2022, 50: 1198–1218
389 Su Z, Figueiredo M C, Jo J, et al. Analyzing description, user understanding and expectations of AI in mobile health applications. In: Proceedings of American Medical Informatics Association Annual Symposium, 2020
390 Li X X, Meng M, Hong Y G, et al. A survey of decision making in adversarial games. Sci China Inf Sci, 2024, 67: 141201
391 Moravcík M, Schmid M, Burch N, et al. DeepStack: expert-level artificial intelligence in no-limit poker. 2017. ArXiv:1701.01724
392 Bakhtin A, Wu D J, Lerer A, et al. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. In: Proceedings of the 11th International Conference on Learning Representations, Kigali, 2023
393 Bakhtin A, Brown N, Dinan E, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. Science, 2022, 378: 1067–1074
394 Carroll M, Shah R, Ho M K, et al. On the utility of learning about humans for human-AI coordination. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 5175–5186
395 Bard N, Foerster J N, Chandar S, et al. The Hanabi challenge: a new frontier for AI research. Artif Intell, 2020, 280: 103216
396 Lin J, Tomlin N, Andreas J, et al. Decision-oriented dialogue for human-AI collaboration. 2023. ArXiv:2305.20076
397 Wang X, Shi W, Kim R, et al. Persuasion for good: towards a personalized persuasive dialogue system for social good. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, 2019. 5635–5649
398 Li C, Su X, Fan C, et al. Quantifying the impact of large language models on collective opinion dynamics. 2023. ArXiv:2308.03313
399 Costa A C D R. A Variational Basis for the Regulation and Structuration Mechanisms of Agent Societies. Cham: Springer, 2019
400 Wimmer S, Pfeiffer A, Denk N. The everyday life in the SIMs 4 during a pandemic. A life simulation as a virtual mirror of society? In: Proceedings of the 15th International Technology, Education and Development Conference, 2021. 5754–5760
401 Lee L, Braud T, Zhou P, et al. All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research agenda. 2021. ArXiv:2110.05352
402 Inkeles A, Smith D H. Becoming Modern: Individual Change in Six Developing Countries. Cambridge: Harvard University Press, 1974
403 Troitzsch K G, Mueller U, Gilbert G N, et al. Social Science Microsimulation. Berlin: Springer, 1996
404 Xu Y, Wang S, Li P, et al. Exploring large language models for communication games: an empirical study on werewolf. 2023. ArXiv:2309.04658
405 Zhou E, Zheng R, Xi Z, et al. RealBehavior: a framework for faithfully characterizing foundation models' human-like behavior mechanisms. In: Proceedings of Findings of the Association for Computational Linguistics Singapore, 2023. 10262–10274
406 Abrams A M H, der Pütten A M R. I-C-E framework: concepts for group dynamics research in human-robot interaction. Int J Soc Robot, 2020, 12: 1213–1229
407 Askell A, Bai Y, Chen A, et al. A general language assistant as a laboratory for alignment. 2021. ArXiv:2112.00861
408 Zhang Z, Liu N, Qi S, et al. Heterogeneous value evaluation for large language models. 2023. ArXiv:2305.17147
409 Browning J. Personhood and AI: why large language models don't understand us. AI Soc, 2024, 39: 2499–2506
410 Jiang G, Xu M, Zhu S, et al. MPI: evaluating and inducing personality in pre-trained language models. 2022. ArXiv:2206.07550

411 Kosinski M. Theory of mind may have spontaneously emerged in large language models. 2023. ArXiv:2302.02083
412 Zuckerman M. Psychobiology of Personality. Cambridge: Cambridge University Press, 1991
413 Binz M, Schulz E. Using cognitive psychology to understand GPT-3. 2022. ArXiv:2206.14576
414 Dhingra S, Singh M, B V S, et al. Mind meets machine: unravelling GPT-4's cognitive psychology. 2023. ArXiv:2303.11436
415 Hagendorff T. Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. 2023. ArXiv:2303.13988
416 Dasgupta I, Lampinen A K, Chan S C Y, et al. Language models show human-like content effects on reasoning. 2022. ArXiv:2207.07051
417 Han S J, Ransom K, Perfors A, et al. Inductive reasoning in humans and large language models. 2023. ArXiv:2306.06548
418 Hagendorff T, Fabi S, Kosinski M. Thinking fast and slow in large language models. 2023. ArXiv:2212.05206
419 Hagendorff T, Fabi S. Human-like intuitive behavior and reasoning biases emerged in language models – and disappeared in GPT-4. 2023. ArXiv:2306.07622
420 Curry A, Curry A C. Computer says "no": the case against empathetic conversational AI. In: Proceedings of Findings of the Association for Computational Linguistics, Toronto, 2023. 8123–8130
421 Elyoseph Z, Hadar-Shoval D, Asraf K, et al. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol, 2023, 14: 1199058
422 Habibi R, Pfau J, Holmes J, et al. Empathetic AI for empowering resilience in games. 2023. ArXiv:2302.09070
423 Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. 2023. ArXiv:2307.15810
424 Bates J. The role of emotion in believable agents. Commun ACM, 1994, 37: 122–125
425 Caron G, Srivastava S. Identifying and manipulating the personality traits of language models. 2022. ArXiv:2212.10276
426 Karra S R, Nguyen S, Tulabandhula T. AI personification: estimating the personality of language models. 2022. ArXiv:2204.12000
427 Pan K, Zeng Y. Do LLMs possess a personality? Making the MBTI test an amazing evaluation for large language models. 2023. ArXiv:2307.16180
428 Li X, Li Y, Joty S, et al. Does GPT-3 demonstrate psychopathy? Evaluating large language models from a psychological perspective. 2023. ArXiv:2212.10529
429 Safdari M, Serapio-García G, Crepy C, et al. Personality traits in large language models. 2023. ArXiv:2307.00184
430 Park J S, Popowski L, Cai C J, et al. Social simulacra: creating populated prototypes for social computing systems. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, Bend, 2022. 1–18
431 Zhang S, Dinan E, Urbanek J, et al. Personalizing dialogue agents: I have a dog, do you have pets too? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, 2018. 2204–2213
432 Kwon D S, Lee S, Kim K H, et al. What, when, and how to ground: designing user persona-aware conversational agents for engaging dialogue. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, 2023. 707–719
433 Maes P. Artificial life meets entertainment. Commun ACM, 1995, 38: 108–114
434 Côté M, Kádár Á, Yuan X, et al. TextWorld: a learning environment for text-based games. In: Proceedings of Held in Conjunction with the 27th International Conference on Artificial Intelligence, Stockholm, 2018. 41–75
435 Hausknecht M J, Ammanabrolu P, Côté M, et al. Interactive fiction games: a colossal adventure. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, New York, 2020. 7903–7910
436 O'Gara A. Hoodwinked: deception and cooperation in a text-based game for language models. 2023. ArXiv:2308.01404
437 Urbanek J, Fan A, Karamcheti S, et al. Learning to speak and act in a fantasy text adventure game. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, 2019. 673–683
438 Gao C, Lan X, Lu Z, et al. S^3: social-network simulation system with large language model-empowered agents. 2023. ArXiv:2307.14984
439 Grossmann I, Feinberg M, Parker D C, et al. AI and the transformation of social science research. Science, 2023, 380: 1108–1109
440 Wang L, Zhang J, Chen X, et al. RecAgent: a novel simulation paradigm for recommender systems. 2023. ArXiv:2306.02552
441 Wei J, Shuster K, Szlam A, et al. Multi-party chat: conversational agents in group settings with humans and models. 2023. ArXiv:2304.13835
442 Hollan J D, Hutchins E L, Weitzman L. STEAMER: an interactive inspectable simulation-based training system. AI Mag, 1984, 5: 15–27
443 Tambe M, Johnson W L, Jones R M, et al. Intelligent agents for interactive simulation environments. AI Mag, 1995, 16: 15–39
444 Vermeulen P, de Jongh D. 'Dynamics of growth in a finite world' — comprehensive sensitivity analysis. IFAC Proc Vol, 1976, 9: 133–145
445 Forrester J W. System dynamics and the lessons of 35 years. In: A Systems-Based Approach to Policymaking. Berlin: Springer, 1993. 199–240
446 Santé I, García A M, Miranda D, et al. Cellular automata models for the simulation of real-world urban processes: a review and analysis. Landscape Urban Planning, 2010, 96: 108–122
447 Dorri A, Kanhere S S, Jurdak R. Multi-agent systems: a survey. IEEE Access, 2018, 6: 28573–28593
448 Hendrickx J M, Martin S. Open multi-agent systems: gossiping with random arrivals and departures. In: Proceedings of the 56th IEEE Annual Conference on Decision and Control, Melbourne, 2017. 763–768
449 Ziems C, Held W, Shaikh O, et al. Can large language models transform computational social science? 2023. ArXiv:2305.03514
450 Gilbert N, Doran J. Simulating Societies: The Computer Simulation of Social Phenomena. London: Routledge, 2018
451 Hamilton J D. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 1989, 57: 357–384
452 Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 2003, 50: 159–175
453 Kirby S, Dowman M, Griffiths T L. Innateness and culture in the evolution of language. Proc Natl Acad Sci USA, 2007, 104: 5241–5245
454 Williams R, Hosseinichimeh N, Majumdar A, et al. Epidemic modeling with generative agents. 2023. ArXiv:2307.04986

455 Shibata H, Miki S, Nakamura Y. Playing the werewolf game with artificial intelligence for language understanding. 2023. ArXiv:2302.10646
456 Junprung E. Exploring the intersection of large language models and agent-based modeling via prompt engineering. 2023. ArXiv:2308.07411
457 Phelps S, Russell Y I. Investigating emergent goal-like behaviour in large language models using experimental economics. 2023. ArXiv:2305.07970
458 Bellomo N, Marsan G A, Tosin A. Complex Systems and Society: Modeling and Simulation. New York: Springer, 2013
459 Moon Y B. Simulation modelling for sustainability: a review of the literature. Int J Sustain Eng, 2017, 10: 2–19
460 Helberger N, Diakopoulos N. ChatGPT and the AI Act. Internet Policy Rev, 2023, 12: 1–6
461 Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models. 2021. ArXiv:2112.04359
462 Deshpande A, Murahari V, Rajpurohit T, et al. Toxicity in ChatGPT: analyzing persona-assigned language models. 2023. ArXiv:2304.05335
463 Kirk H R, Jun Y, Volpin F, et al. Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 2611–2624
464 Nadeem M, Bethke A, Reddy S. StereoSet: measuring stereotypical bias in pretrained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 5356–5371
465 Roberts T, Marchais G. Assessing the role of social media and digital technology in violence reporting. Contemp Readings Law & Social Justice, 2018, 10: 9–42
466 Kandpal N, Deng H, Roberts A, et al. Large language models struggle to learn long-tail knowledge. In: Proceedings of the 40th International Conference on Machine Learning, 2023. 15696–15707
467 Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models. 2023. ArXiv:2304.03738
468 Haller P, Aynetdinov A, Akbik A. OpinionGPT: modelling explicit biases in instruction-tuned LLMs. 2023. ArXiv:2309.03876
469 Salewski L, Alaniz S, Rio-Torto I, et al. In-context impersonation reveals large language models' strengths and biases. 2023. ArXiv:2305.14930
470 Lin B, Bouneffouf D, Cecchi G A, et al. Towards healthy AI: large language models need therapists too. 2023. ArXiv:2304.00416
471 Liang P P, Wu C, Morency L, et al. Towards understanding and mitigating social biases in language models. In: Proceedings of the 38th International Conference on Machine Learning, 2021. 6565–6576
472 Henderson P, Sinha K, Angelard-Gontier N, et al. Ethical challenges in data-driven dialogue systems. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2018. 123–129
473 Li H, Song Y, Fan L. You don't know my favorite color: preventing dialogue representations from revealing speakers' private personas. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Seattle, 2022. 5858–5870
474 Brown H, Lee K, Mireshghallah F, et al. What does it mean for a language model to preserve privacy? In: Proceedings of ACM Conference on Fairness, Accountability, and Transparency, Seoul, 2022. 2280–2292
475 Sebastian G. Privacy and data protection in ChatGPT and other AI chatbots: strategies for securing user information. Int J Secur Privacy Pervasive Comput, 2023. doi: 10.4018/ijsppc.325475
476 Reeves B, Nass C. The Media Equation — How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge: Cambridge University Press, 1996
477 Roose K. A conversation with Bing's chatbot left me deeply unsettled. 2023. https://philosophy.tamucc.edu/texts/chat-with-chatgpt
478 Bach S H, Sanh V, Yong Z X, et al. PromptSource: an integrated development environment and repository for natural language prompts. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, 2022. 93–104
479 Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. 2022. ArXiv:2212.08073
480 Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022. ArXiv:2204.05862
481 Team X. XAgent: an autonomous agent for complex task solving. 2023. https://github.com/OpenBMB/XAgent
482 Xi Z, Ding Y, Chen W, et al. AgentGym: evolving large language model-based agents across diverse environments. 2024. ArXiv:2406.04151
483 OpenAI. Swarm. 2024. https://github.com/openai/swarm
484 Liu X, Yu H, Zhang H, et al. AgentBench: evaluating LLMs as agents. 2023. ArXiv:2308.03688
485 Aher G V, Arriaga R I, Kalai A T. Using large language models to simulate multiple humans and replicate human subject studies. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 337–371
486 Liang Y, Zhu L, Yang Y. Tachikuma: understading complex interactions with multi-character and novel objects by large language models. 2023. ArXiv:2307.12573
487 Xu B, Liu X, Shen H, et al. Gentopia: a collaborative platform for tool-augmented LLMs. 2023. ArXiv:2308.04030
488 Adiwardana D, Luong M, So D R, et al. Towards a human-like open-domain chatbot. 2020. ArXiv:2001.09977
489 Kim S S, Watkins E A, Russakovsky O, et al. "Help me help the AI": understanding how explainability can support human-AI interaction. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2023. 1–17
490 Jhan J, Liu C, Jeng S, et al. CheerBots: chatbots toward empathy and emotionusing reinforcement learning. 2021. ArXiv:2110.03949
491 Choi M, Pei J, Kumar S, et al. Do LLMs understand social knowledge? Evaluating the sociability of large language models with socket benchmark. 2023. ArXiv:2305.14938
492 Wilson A C, Bishop D V M. "If you catch my drift...": ability to infer implied meaning is distinct from vocabulary and grammar skills. Wellcome Open Res, 2019, 4: 68
493 Fang T, Yang S, Lan K, et al. Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation. 2023. ArXiv:2304.01746
494 Shi J, Zhao J, Wang Y, et al. CGMI: configurable general multi-agent interaction framework. 2023. ArXiv:2308.12503
495 Shuster K, Urbanek J, Szlam A, et al. Am I me or you? State-of-the-art dialogue models cannot maintain an identity. In: Proceedings of Findings of the Association for Computational Linguistics, Seattle, 2022. 2367–2387
496 Ganguli D, Lovitt L, Kernion J, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. 2022. ArXiv:2209.07858

497 Kadavath S, Conerly T, Askell A, et al. Language models (mostly) know what they know. 2022. ArXiv:2207.05221
498 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations, Banff, 2014
499 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations, San Diego, 2015
500 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations, Vancouver, 2018
501 Zheng R, Xi Z, Liu Q, et al. Characterizing the impacts of instances on robustness. In: Proceedings of Findings of the Association for Computational Linguistics, Toronto, 2023. 2314–2332
502 Xi Z H, Zheng R, Gui T. Safety and ethical concerns of large language models. In: Proceedings of the 22nd Chinese National Conference on Computational Linguistics, 2023. 9–16
503 Akhtar N, Mian A, Kardan N, et al. Threat of adversarial attacks on deep learning in computer vision: survey II. 2021. ArXiv:2108.00401
504 Drenkow N, Sani N, Shpitser I, et al. A systematic review of robustness in deep learning for computer vision: mind the gap? 2021. ArXiv:2112.00639
505 Hendrycks D, Dietterich T G. Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the 7th International Conference on Learning Representations, 2019
506 Wang X, Wang H, Yang D. Measure and improve robustness in NLP models: a survey. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, 2022. 4569–4586
507 Li J, Ji S, Du T, et al. TextBugger: generating adversarial text against real-world applications. In: Proceedings of the 26th Annual Network and Distributed System Security Symposium, San Diego, 2019
508 Zhu C, Cheng Y, Gan Z, et al. FreeLB: enhanced adversarial training for natural language understanding. In: Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, 2020
509 Xi Z, Zheng R, Gui T, et al. Efficient adversarial training with robust early-bird tickets. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, 2022. 8318–8331
510 Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017. 2817–2826
511 Rigter M, Lacerda B, Hawes N. RAMBO-RL: robust adversarial model-based offline reinforcement learning. In: Proceedings of Conference on Neural Information Processing Systems, 2022
512 Panaganti K, Xu Z, Kalathil D, et al. Robust reinforcement learning using offline data. In: Proceedings of Conference on Neural Information Processing Systems, 2022
513 Lab T K S. Experimental security research of Tesla autopilot. Tencent Keen Security Lab, 2019. https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf
514 Xu K, Zhang G, Liu S, et al. Adversarial T-shirt! Evading person detectors in a physical world. In: Proceedings of the 16th European Conference on Computer Vision, Glasgow, 2020. 665–681
515 Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Vienna, 2016. 1528–1540
516 Jin D, Jin Z, Zhou J T, et al. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence, 2020. 8018–8025
517 Ren S, Deng Y, He K, et al. Generating natural language adversarial examples through probability weighted word saliency. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019. 1085–1097
518 Zhu K, Wang J, Zhou J, et al. PromptBench: towards evaluating the robustness of large language models on adversarial prompts. 2023. ArXiv:2306.04528
519 Chen X, Ye J, Zu C, et al. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. 2023. ArXiv:2303.00293
520 Gu T, Dolan-Gavitt B, Garg S. BadNets: identifying vulnerabilities in the machine learning model supply chain. 2017. ArXiv:1708.06733
521 Chen X, Salem A, Chen D, et al. BadNL: backdoor attacks against NLP models with semantic-preserving improvements. In: Proceedings of Annual Computer Security Applications Conference, 2021. 554–569
522 Li Z, Mekala D, Dong C, et al. BFClass: a backdoor-free text classification framework. In: Proceedings of Findings of the Association for Computational Linguistics, 2021. 444–453
523 Shi Y, Li P, Yin C, et al. PromptAttack: prompt-based attack for language models via gradient search. In: Proceedings of the 11th CCF International Conference on Natural Language Processing and Chinese Computing, Guilin, 2022. 682–693
524 Perez F, Ribeiro I. Ignore previous prompt: attack techniques for language models. 2022. ArXiv:2211.09527
525 Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. 2022. ArXiv:2211.09110
526 Gururangan S, Card D, Dreier S K, et al. Whose language counts as high quality? Measuring language ideologies in text data selection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, 2022. 2562–2580
527 Liu Y, Deng G, Li Y, et al. Prompt injection attack against LLM-integrated applications. 2023. ArXiv:2306.05499
528 Carlini N, Wagner D A. Audio adversarial examples: targeted attacks on speech-to-text. In: Proceedings of IEEE Security and Privacy Workshops, San Francisco, 2018. 1–7
529 Morris J X, Lifland E, Yoo J Y, et al. TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020. 119–126
530 Si C, Zhang Z, Qi F, et al. Better robustness by more coverage: adversarial and mixup data augmentation for robust finetuning. In: Proceedings of Findings of the Association for Computational Linguistics, 2021. 1569–1576
531 Yoo K, Kim J, Jang J, et al. Detection of adversarial examples in text classification: benchmark and baseline via robust density estimation. In: Proceedings of Findings of the Association for Computational Linguistics, 2022. 3656–3672
532 Le T, Park N, Lee D. A sweet rabbit hole by DARCY: using honeypots to detect universal trigger's adversarial attacks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. 3831–3844

533 Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at odds with accuracy. In: Proceedings of the 7th International Conference on Learning Representations, 2019
534 Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, 2019. 7472–7482
535 Wong A, Wang X Y, Hryniowski A. How much can we really trust you? Towards simple, interpretable trust quantification metrics for deep neural networks. 2020. ArXiv:2009.05835
536 Huang X, Kroening D, Ruan W, et al. A survey of safety and trustworthiness of deep neural networks: verification, testing, adversarial attack and defence, and interpretability. Comput Sci Rev, 2020, 37: 100270
537 Huang X, Ruan W, Huang W, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. 2023. ArXiv:2305.11391
538 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res, 2020, 21: 5485–5551
539 Chen Y, Yuan L, Cui G, et al. A close look into the calibration of pre-trained language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023. 1343–1367
540 Blodgett S L, Barocas S, III H D, et al. Language (technology) is power: a critical survey of "bias" in NLP. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 5454–5476
541 Guo W, Caliskan A. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of AAAI/ACM Conference on AI, Ethics, and Society, 2021. 122–133
542 Bolukbasi T, Chang K, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of Advances in Neural Information Processing Systems, 2016. 4349–4357
543 Caliskan A, Bryson J J, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science, 2017, 356: 183–186
544 Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv, 2023, 55: 1–38
545 Mündler N, He J, Jenko S, et al. Self-contradictory hallucinations of large language models: evaluation, detection and mitigation. 2023. ArXiv:2305.15852
546 Maynez J, Narayan S, Bohnet B, et al. On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 1906–1919
547 Varshney N, Yao W, Zhang H, et al. A stitch in time saves nine: detecting and mitigating hallucinations of LLMs by validating low-confidence generation. 2023. ArXiv:2307.03987
548 Ke W J, Shang Z Y, Luo Z Z, et al. Unveiling factuality and injecting knowledge for LLMs via reinforcement learning and data proportion. Sci China Inf Sci, 2024, 67: 209101
549 Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step. 2023. ArXiv:2305.20050
550 Tang X, Zou A, Zhang Z, et al. MedAgents: large language models as collaborators for zero-shot medical reasoning. In: Proceedings of Findings of the Association for Computational Linguistics, Bangkok, 2024. 599–621
551 Guo Y, Yang Y, Abbasi A. Auto-Debias: debiasing masked language models with automated biased prompts. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, 2022. 1012–1023
552 Du M, He F, Zou N, et al. Shortcut learning of large language models in natural language understanding: a survey. 2022. ArXiv:2208.11857
553 Brundage M, Avin S, Clark J, et al. The malicious use of artificial intelligence: forecasting, prevention, and mitigation. 2018. ArXiv:1802.07228
554 Bommasani R, Hudson D A, Adeli E, et al. On the opportunities and risks of foundation models. 2021. ArXiv:2108.07258
555 Charan P V S, Chunduri H, Anand P M, et al. From text to MITRE techniques: exploring the malicious use of large language models for generating cyber attack payloads. 2023. ArXiv:2305.15336
556 Wang Z J, Choi D, Xu S, et al. Putting humans in the natural language processing loop: a survey. 2021. ArXiv:2103.04044
557 Yao S, Narasimhan K. Language agents in the digital world: opportunities and risks. 2023. https://princeton-nlp.github.io
558 Asimov I. Three Laws of Robotics. 1942. https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
559 Yuan T, He Z, Dong L, et al. R-Judge: benchmarking safety risk awareness for LLM agents. 2024. ArXiv:2401.10019
560 Elhage N, Nanda N, Olsson C, et al. A mathematical framework for transformer circuits. Transformer Circ Thread, 2021, 1: 12
561 Zhuge M, Liu H, Faccio F, et al. Mindstorms in natural language-based societies of mind. 2023. ArXiv:2305.17066
562 Bai J, Zhang S, Chen Z. Is there any social principle for llm-based agents? 2023. ArXiv:2308.11136
563 Shridhar M, Yuan X, Côté M, et al. ALFWorld: aligning text and embodied environments for interactive learning. In: Proceedings of the 9th International Conference on Learning Representations, 2021
564 Colas C, Teodorescu L, Oudeyer P, et al. Augmenting autotelic agents with large language models. 2023. ArXiv:2305.12487
565 Bertsch A, Alon U, Neubig G, et al. Unlimiformer: long-range transformers with unlimited length input. 2023. ArXiv:2305.01625
566 Chowdhury J R, Caragea C. Monotonic location attention for length generalization. In: Proceedings of International Conference on Machine Learning, Honolulu, 2023. 28792–28808
567 Duan Y, Fu G, Zhou N, et al. Everything as a service (XaaS) on the cloud: origins, current and future trends. In: Proceedings of the 8th IEEE International Conference on Cloud Computing, New York City, 2015. 621–628
568 Bhardwaj S, Jain L, Jain S. Cloud computing: a study of infrastructure as a service (IAAS). Int J Engin Inform Technol, 2010, 2: 60–63
569 Serrano N, Gallardo G, Hernantes J. Infrastructure as a service and cloud technologies. IEEE Softw, 2015, 32: 30–36
570 Mell P, Grance T. The NIST definition of cloud computing. Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, 2011. doi: 10.6028/NIST.SP.800-145
571 Lawton G. Developing software online with platform-as-a-service technology. Computer, 2008, 41: 13–15
572 Sun W, Zhang K, Chen S K, et al. Software as a service: an integration perspective. In: Proceedings of the 5th International Conference on Service-Oriented Computing, Vienna, 2007. 558–569
573 Dubey A, Wagle D. Delivering software as a service. The McKinsey Quart, 2007, 6: 2007
574 Sun T, Shao Y, Qian H, et al. Black-box tuning for language-model-as-a-service. In: Proceedings of International Conference on Machine Learning, Baltimore, 2022. 20841–20855