The Rise and Potential of Large Language Model Based Agents: A Survey

Zhiheng Xi*†, Wenxiang Chen*, Xin Guo*, Wei He*, Yiwen Ding*, Boyang Hong*, Ming Zhang*, Junzhe Wang*, Senjie Jin*, Enyu Zhou*,

Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin,

Shihan Dou, Rongxiang Weng, Wensen Cheng,

Qi Zhang†, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang and Tao Gui†

Fudan NLP Group

Abstract

For a long time, humanity has pursued artificial intelligence (AI) equivalent to or surpassing the human level, with AI agents considered a promising vehicle for this pursuit. AI agents are artificial entities that sense their environment, make decisions, and take actions. Many efforts have been made to develop intelligent agents, but they mainly focus on advancement in algorithms or training strategies to enhance specific capabilities or performance on particular tasks. Actually, what the community lacks is a general and powerful model to serve as a starting point for designing AI agents that can adapt to diverse scenarios. Due to the versatile capabilities they demonstrate, large language models (LLMs) are regarded as potential sparks for Artificial General Intelligence (AGI), offering hope for building general AI agents. Many researchers have leveraged LLMs as the foundation to build AI agents and have achieved significant progress. In this paper, we perform a comprehensive survey on LLM-based agents. We start by tracing the concept of agents from its philosophical origins to its development in AI, and explain why LLMs are suitable foundations for agents. Building upon this, we present a general framework for LLM-based agents, comprising three main components: brain, perception, and action, and the framework can be tailored for different applications. Subsequently, we explore the extensive applications of LLM-based agents in three aspects: single-agent scenarios, multi-agent scenarios, and human-agent cooperation. Following this, we delve into agent societies, exploring the behavior and personality of LLM-based agents, the social phenomena that emerge from an agent society, and the insights they offer for human society. Finally, we discuss several key topics and open problems within the field. A repository for the related papers at https://github.com/WooooDyy/LLM-Agent-Paper-List.

† Correspondence to: zhxi22@m.fudan.edu.cn, {qz, tgui}@fudan.edu.cn
* Equal Contribution.

Contents

1 Introduction

> *"If they find a parrot who could answer to everything, I would claim it to be an intelligent being without hesitation."*
>
> —Denis Diderot, 1875

Artificial Intelligence (AI) is a field dedicated to designing and developing systems that can replicate human-like intelligence and abilities [1]. As early as the 18th century, philosopher Denis Diderot introduced the idea that if a parrot could respond to every question, it could be considered intelligent [2]. While Diderot was referring to living beings, like the parrot, his notion highlights the profound concept that a highly intelligent organism could resemble human intelligence. In the 1950s, Alan Turing expanded this notion to artificial entities and proposed the renowned Turing Test [3]. This test is a cornerstone in AI and aims to explore whether machines can display intelligent behavior comparable to humans. These AI entities are often termed "agents", forming the essential building blocks of AI systems. Typically in AI, an agent refers to an artificial entity capable of perceiving its surroundings using sensors, making decisions, and then taking actions in response using actuators [1; 4].

The concept of agents originated in Philosophy, with roots tracing back to thinkers like Aristotle and Hume [5]. It describes entities possessing desires, beliefs, intentions, and the ability to take actions [5]. This idea transitioned into computer science, intending to enable computers to understand users' interests and autonomously perform actions on their behalf [6; 7; 8]. As AI advanced, the term "agent" found its place in AI research to depict entities showcasing intelligent behavior and possessing qualities like autonomy, reactivity, pro-activeness, and social ability [4; 9]. Since then, the exploration and technical advancement of agents have become focal points within the AI community [1; 10]. AI agents are now acknowledged as a pivotal stride towards achieving Artificial General Intelligence (AGI) [1], as they encompass the potential for a wide range of intelligent activities [4; 11; 12].

From the mid-20th century, significant strides were made in developing smart AI agents as research delved deep into their design and advancement [13; 14; 15; 16; 17; 18]. However, these efforts have predominantly focused on enhancing specific capabilities, such as symbolic reasoning, or mastering particular tasks like Go or Chess [19; 20; 21]. Achieving a broad adaptability across varied scenarios remained elusive. Moreover, previous studies have placed more emphasis on the design of algorithms and training strategies, overlooking the development of the model's inherent general abilities like knowledge memorization, long-term planning, effective generalization, and efficient interaction [22; 23]. Actually, enhancing the inherent capabilities of the model is the pivotal factor for advancing the agent further, and the domain is in need of a powerful foundational model endowed with a variety of key attributes mentioned above to serve as a starting point for agent systems.

The development of large language models (LLMs) has brought a glimmer of hope for the further development of agents [24; 25; 26], and significant progress has been made by the community [22; 27; 28; 29]. According to the notion of World Scope (WS) [30] which encompasses five levels that depict the research progress from NLP to general AI (i.e., Corpus, Internet, Perception, Embodiment, and Social), the pure LLMs are built on the second level with internet-scale textual inputs and outputs. Despite this, LLMs have demonstrated powerful capabilities in knowledge acquisition, instruction comprehension, generalization, planning, and reasoning, while displaying effective natural language interactions with humans. These advantages have earned LLMs the designation of sparks for AGI [31], making them highly desirable for building intelligent agents to foster a world where humans and agents coexist harmoniously [22]. Starting from this, if we elevate LLMs to the status of agents and equip them with an expanded perception space and action space, they have the potential to reach the third and fourth levels of WS. Furthermore, these LLMs-based agents can tackle more complex tasks through cooperation or competition, and emergent social phenomena can be observed when placing them together, potentially achieving the fifth WS level. As shown in Figure 1, we envision a harmonious society composed of AI agents where human can also participate.

In this paper, we present a comprehensive and systematic survey focusing on LLM-based agents, attempting to investigate the existing studies and prospective avenues in this burgeoning field. To this end, we begin by delving into crucial **background** information (§ 2). In particular, we commence by tracing the origin of AI agents from philosophy to the AI domain, along with a brief overview of the

[1] Also known as Strong AI.

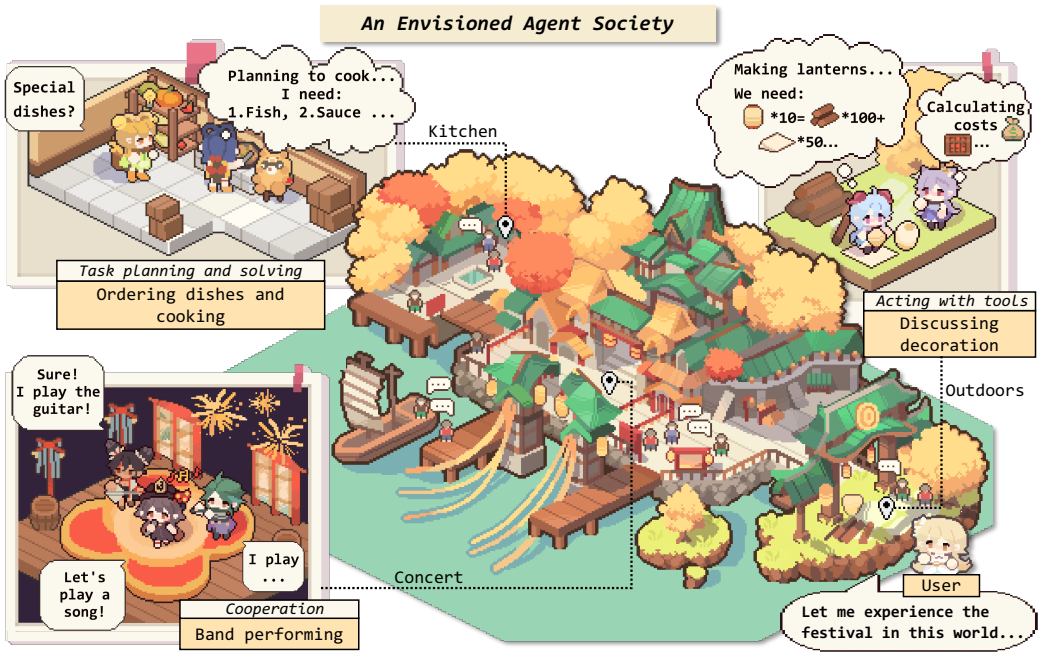

Figure 1: Scenario of an envisioned society composed of AI agents, in which humans can also participate. The above image depicts some specific scenes within society. In the kitchen, one agent orders dishes, while another agent is responsible for planning and solving the cooking task. At the concert, three agents are collaborating to perform in a band. Outdoors, two agents are discussing lantern-making, planning the required materials, and finances by selecting and using tools. Users can participate in any of these stages of this social activity.

debate surrounding the existence of artificial agents (§ 2.1). Next, we take the lens of technological trends to provide a concise historical review of the development of AI agents (§ 2.2). Finally, we delve into an in-depth introduction of the essential characteristics of agents and elucidate why large language models are well-suited to serve as the main component of brains or controllers for AI agents (§ 2.3).

Inspired by the definition of the agent, we present a general conceptual **framework** for the LLM-based agents with three key parts: **brain, perception, and action** (§ 3), and the framework can be tailored to suit different applications. We first introduce the brain, which is primarily composed of a large language model (§ 3.1). Similar to humans, the brain is the core of an AI agent because it not only stores crucial memories, information, and knowledge but also undertakes essential tasks of information processing, decision-making, reasoning, and planning. It is the key determinant of whether the agent can exhibit intelligent behaviors. Next, we introduce the perception module (§ 3.2). For an agent, this module serves a role similar to that of sensory organs for humans. Its primary function is to expand the agent's perceptual space from text-only to a multimodal space that includes diverse sensory modalities like text, sound, visuals, touch, smell, and more. This expansion enables the agent to better perceive information from the external environment. Finally, we present the action module for expanding the action space of an agent (§ 3.3). Specifically, we expect the agent to be able to possess textual output, take embodied actions, and use tools so that it can better respond to environmental changes and provide feedback, and even alter and shape the environment.

After that, we provide a detailed and thorough introduction to the **practical applications** of LLM-based agents and elucidate the foundational design pursuit—"Harnessing AI for good" (§ 4). To start, we delve into the current applications of a single agent and discuss their performance in text-based tasks and simulated exploration environments, with a highlight on their capabilities in handling specific tasks, driving innovation, and exhibiting human-like survival skills and adaptability (§ 4.1). Following that, we take a retrospective look at the development history of multi-agents. We introduce the interactions between agents in LLM-based multi-agent system applications, where they engage in

collaboration, negotiation, or competition. Regardless of the mode of interaction, agents collectively strive toward a shared objective (§ 4.2). Lastly, considering the potential limitations of LLM-based agents in aspects such as privacy security, ethical constraints, and data deficiencies, we discuss the human-agent collaboration. We summarize the paradigms of collaboration between agents and humans: the instructor-executor paradigm and the equal partnership paradigm, along with specific applications in practice (§ 4.3).

Building upon the exploration of practical applications of LLM-based agents, we now shift our focus to the concept of the "**Agent Society**", examining the intricate interactions between agents and their surrounding environments (§ 5). This section begins with an investigation into whether these agents exhibit human-like behavior and possess corresponding personality (§5.1). Furthermore, we introduce the social environments within which the agents operate, including text-based environment, virtual sandbox, and the physical world (§5.2). Unlike the previous section (§ 3.2), here we will focus on diverse types of the environment rather than how the agents perceive it. Having established the foundation of agents and their environments, we proceed to unveil the simulated societies that they form (§5.3). We will discuss the construction of a simulated society, and go on to examine the social phenomena that emerge from it. Specifically, we will emphasize the lessons and potential risks inherent in simulated societies.

Finally, we discuss a range of key **topics** (§ 6) and open problems within the field of LLM-based agents: (1) the mutual benefits and inspirations of the LLM research and the agent research, where we demonstrate that the development of LLM-based agents has provided many opportunities for both agent and LLM communities (§ 6.1); (2) existing evaluation efforts and some prospects for LLM-based agents from four dimensions, including utility, sociability, values and the ability to continually evolve (§ 6.2); (3) potential risks of LLM-based agents, where we discuss adversarial robustness and trustworthiness of LLM-based agents. We also include the discussion of some other risks like misuse, unemployment and the threat to the well-being of the human race (§ 6.3); (4) scaling up the number of agents, where we discuss the potential advantages and challenges of scaling up agent counts, along with the approaches of pre-determined and dynamic scaling (§ 6.4); (5) several open problems, such as the debate over whether LLM-based agents represent a potential path to AGI, challenges from virtual simulated environment to physical environment, collective Intelligence in AI agents, and Agent as a Service (§ 6.5). After all, we hope this paper could provide inspiration to the researchers and practitioners from relevant fields.

2 Background

In this section, we provide crucial background information to lay the groundwork for the subsequent content (§ 2.1). We first discuss the origin of AI agents, from philosophy to the realm of AI, coupled with a discussion of the discourse regarding the existence of artificial agents (§ 2.2). Subsequently, we summarize the development of AI agents through the lens of technological trends. Finally, we introduce the key characteristics of agents and demonstrate why LLMs are suitable to serve as the main part of the brains of AI agents (§ 2.3).

2.1 Origin of AI Agent

"Agent" is a concept with a long history that has been explored and interpreted in many fields. Here, we first explore its origins in philosophy, discuss whether artificial products can possess agency in a philosophical sense, and examine how related concepts have been introduced into the field of AI.

Agent in philosophy. The core idea of an agent has a historical background in philosophical discussions, with its roots traceable to influential thinkers such as Aristotle and Hume, among others [5]. 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 [5]. While in a narrow sense, "agency" is usually used to refer to the performance of intentional actions; and correspondingly, the term "agent" denotes entities that possess desires, beliefs, intentions, and the ability to act [32; 33; 34; 35]. Note that agents can encompass not only individual human beings but also other entities in both the physical and virtual world. Importantly, the concept of an agent involves individual autonomy, granting them the ability to exercise volition, make choices, and take actions, rather than passively reacting to external stimuli.

From the perspective of philosophy, is artificial entities capable of agency? In a general sense, if we define agents as entities with the capacity to act, AI systems do exhibit a form of agency [5]. However, the term agent is more usually used to refer to entities or subjects that possess consciousness, intentionality, and the ability to act [32; 33; 34]. Within this framework, it's not immediately clear whether artificial systems can possess agency, as it remains uncertain whether they possess internal states that form the basis for attributing desires, beliefs, and intentions. Some people argue that attributing psychological states like intention to artificial agents is a form of anthropomorphism and lacks scientific rigor [5; 36]. As Barandiaran et al. [36] stated, "Being specific about the requirements for agency has told us a lot about how much is still needed for the development of artificial forms of agency." In contrast, there are also researchers who believe that, in certain circumstances, employing the intentional stance (that is, interpreting agent behavior in terms of intentions) can provide a better description, explanation and abstraction of the actions of artificial agents, much like it is done for humans [11; 37; 38].

With the advancement of language models, the potential emergence of artificial intentional agents appears more promising [24; 25; 39; 40; 41]. In a rigorous sense, language models merely function as conditional probability models, using input to predict the next token [42]. Different from this, humans incorporate social and perceptual context, and speak according to their mental states [43; 44]. Consequently, some researchers argue that the current paradigm of language modeling is not compatible with the intentional actions of an agent [30; 45]. However, there are also researchers who propose that language models can, in a narrow sense, serve as models of agents [46; 47]. They argue that during the process of context-based next-word prediction, current language models can sometimes infer approximate, partial representations of the beliefs, desires, and intentions held by the agent who generated the context. With these representations, the language models can then generate utterances like humans. To support their viewpoint, they conduct experiments to provide some empirical evidence [46; 48; 49].

Introduction of agents into AI. 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 [50; 51; 52; 53]. As Wooldridge et al. [4] stated, 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 the field of AI, its meaning undergoes some changes. In the realm of Philosophy, an agent can be a human, an animal, or even a concept or entity with autonomy [5]. However, in the field of artificial intelligence, an agent is a computational entity [4; 7]. Due to the seemingly metaphysical nature of concepts like consciousness and desires for computational entities [11], and given that we can only observe the behavior of the machine, many AI researchers, including Alan Turing, suggest temporarily setting aside the question of whether an agent is "actually" thinking or literally possesses a "mind" [3]. Instead, researchers employ other attributes to help describe an agent, such as properties of autonomy, reactivity, pro-activeness and social ability [4; 9]. There are also researchers who held that intelligence is "in the eye of the beholder"; it is not an innate, isolated property [15; 16; 54; 55]. In essence, an AI agent is not equivalent to a philosophical agent; rather, it is a concretization of the philosophical concept of an agent in the context of AI. In this paper, we treat AI agents as artificial entities that are capable of perceiving their surroundings using sensors, making decisions, and then taking actions in response using actuators [1; 4].

2.2 Technological Trends in Agent Research

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

Symbolic Agents. In the early stages of artificial intelligence research, the predominant approach utilized was symbolic AI, characterized by its reliance on symbolic logic [56; 57]. This approach employed logical rules and symbolic representations to encapsulate knowledge and facilitate reasoning processes. Early AI agents were built based on this approach [58], and they primarily focused on two problems: the transduction problem and the representation/reasoning problem [59]. These agents are aimed 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 [13; 14; 60]. A classic example of this approach is knowledge-based expert systems. However, symbolic agents faced limitations in handling uncertainty and large-scale real-world problems [19; 20]. Additionally, due to the intricacies of symbolic reasoning algorithms, it was challenging to find an efficient algorithm capable of producing meaningful results within a finite timeframe [20; 61].

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 [15; 16; 20; 62; 63]. 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 [52]. 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 [17; 18; 64; 65]. 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 [21]. Initially, reinforcement learning (RL) agents were primarily based on fundamental techniques such as policy search and value function optimization, exemplified by Q-learning [66] and SARSA [67]. With the rise of deep learning, the integration of deep neural networks and reinforcement learning, known as Deep Reinforcement Learning (DRL), has emerged [68; 69]. This allows agents to learn intricate policies from high-dimensional inputs, leading to numerous significant accomplishments like AlphaGo [70] and DQN [71]. 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 [21].

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 [72; 73; 74; 75; 76]. Consequently, researchers have introduced transfer learning to expedite an agent's learning on new tasks [77; 78; 79]. 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 AI agents [80; 81; 82; 83; 84]. 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 [85]. 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 there may exist negative transfer [86; 87]. Additionally, the substantial amount of pre-training and large sample sizes required by meta learning make it hard to establish a universal learning policy [81; 88].

Large language model-based agents. As large language models have demonstrated impressive emergent capabilities and have gained immense popularity [24; 25; 26; 41], researchers have started to leverage these models to construct AI agents [22; 27; 28; 89]. Specifically, they employ LLMs as the primary component of brain or controller of these agents and expand their perceptual and action space through strategies such as multimodal perception and tool utilization [90; 91; 92; 93; 94]. These LLM-based agents can exhibit reasoning and planning abilities comparable to symbolic agents through techniques like Chain-of-Thought (CoT) and problem decomposition [95; 96; 97; 98; 99; 100; 101]. They can also acquire interactive capabilities with the environment, akin to reactive agents, by learning from feedback and performing new actions [102; 103; 104]. Similarly, large language models 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 [41; 105; 106; 107]. LLM-based agents have been applied to various real-world scenarios,

such as software development [108; 109] and scientific research [110]. 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 [108; 109; 111; 112]. Furthermore, research suggests that allowing multiple agents to coexist can lead to the emergence of social phenomena [22].

2.3 Why is LLM suitable as the primary component of an Agent's brain?

As mentioned before, researchers have introduced several properties to help describe and define agents in the field of AI. Here, we will delve into some key properties, elucidate their relevance to LLMs, and thereby expound on why LLMs are highly suited to serve as the main part of brains 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 [4; 113]. 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 [114; 115]. Moreover, they can dynamically adjust their outputs based on environmental input, reflecting a degree of adaptive autonomy [23; 27; 104]. Furthermore, they can showcase autonomy through exhibiting creativity like coming up with novel ideas, stories, or solutions that haven't been explicitly programmed into them [116; 117]. This implies a certain level of self-directed exploration and decision-making. Applications like Auto-GPT [114] 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 [9]. 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. However, researchers have demonstrated the potential to expand the perceptual space of LLMs using multimodal fusion techniques, enabling them to rapidly process visual and auditory information from the environment [25; 118; 119]. Similarly, it's also feasible to expand the action space of LLMs through embodiment techniques [120; 121] and tool usage [92; 94]. 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 textual form before eventually translating them into concrete actions. This intermediary process consumes time and reduces the response speed. However, this aligns closely with human behavioral patterns, where the principle of "think before you act" is observed [122; 123].

Pro-activeness. Pro-activeness denotes that agents don't merely react to their environments; they possess the capacity to display goal-oriented actions by proactively taking the initiative [9]. 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 [46; 48; 49]. LLMs have demonstrated a strong capacity for generalized reasoning and planning. By prompting large language models with instructions like "let's think step by step", we can elicit their reasoning abilities, such as logical and mathematical reasoning [95; 96; 97]. Similarly, large language models have shown the emergent ability of planning in forms of goal reformulation [99; 124], task decomposition [98; 125], and adjusting plans in response to environmental changes [100; 126].

Social ability. Social ability refers to an agent's capacity to interact with other agents, including humans, through some kind of agent-communication language [8]. Large language models exhibit strong natural language interaction abilities like understanding and generation [23; 127; 128]. Compared to structured languages or other communication protocals, such capability enables them to interact with other models or humans in an interpretable manner. This forms the cornerstone of social ability for LLM-based agents [22; 108]. Many researchers have demonstrated that LLM-based

agents can enhance task performance through social behaviors such as collaboration and competition [108; 111; 129; 130]. By inputting specific prompts, LLMs can also play different roles, thereby simulating the social division of labor in the real world [109]. Furthermore, when we place multiple agents with distinct identities into a society, emergent social phenomena can be observed [22].

3 The Birth of An Agent: Construction of LLM-based Agents

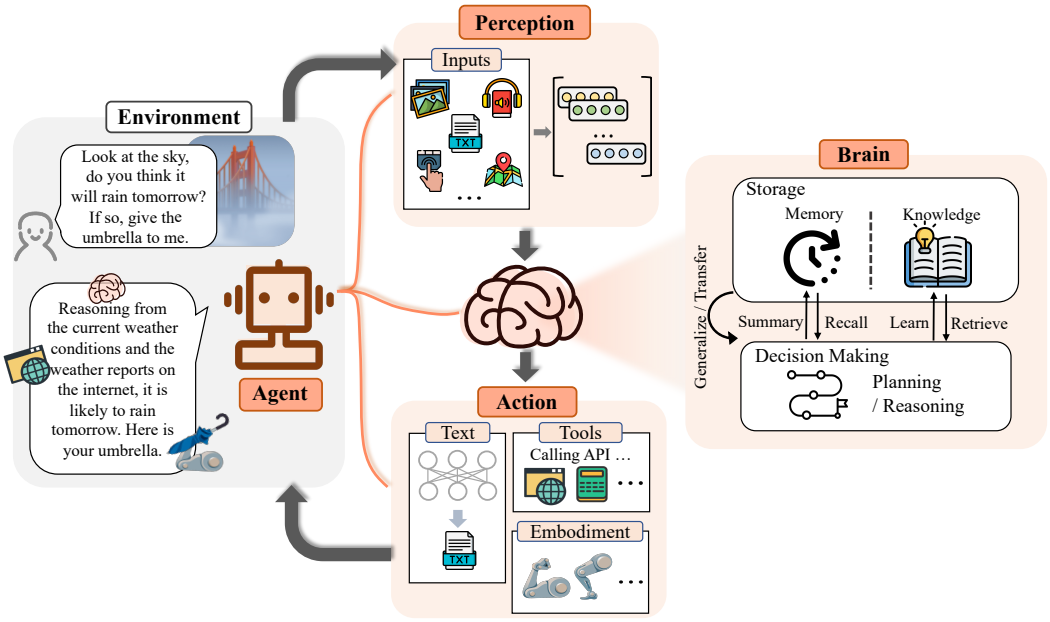

Figure 2: Conceptual framework of LLM-based agent with three components: brain, perception, and action. Serving as the controller, the brain module undertakes basic tasks like memorizing, thinking, and decision-making. The perception module perceives and processes multimodal information from the external environment, and the action module carries out the execution using tools and influences the surroundings. Here we give an example to illustrate the workflow: When a human asks whether it will rain, the perception module converts the instruction into an understandable representation for LLMs. 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, an agent can continuously get feedback and interact with the environment.

"Survival of the Fittest" [131] shows that if an individual wants to survive in the external environment, he must adapt to the surroundings efficiently. This requires him to be cognitive, able to perceive and respond to changes in the outside world, which is consistent with the definition of "agent" mentioned in §2.1. Inspired by this, we present a general conceptual framework of an LLM-based agent composed of three key parts: brain, perception, and action (see Figure 2). We first describe the structure and working mechanism of the brain, which is primarily composed of a large language model (§ 3.1). The brain is the core of an AI agent because it not only stores knowledge and memories but also undertakes indispensable functions like information processing and decision-making. It can present the process of reasoning and planning, and cope well with unseen tasks, exhibiting the intelligence of an agent. Next, we introduce the perception module (§ 3.2). Its core purpose is to broaden the agent's perception space from a text-only domain to a multimodal sphere that includes textual, auditory, and visual modalities. This extension 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 (§ 3.3). Specifically, we empower the agent with embodied action ability and tool-handling skills, enabling it to adeptly adapt to environmental changes, provide feedback, 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 such as the eyes and ears, 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 information processing activities such as thinking, decision-making, and operations with storage including memory and knowledge. Finally, the **action** module, corresponding to human limbs, carries out the execution with the assistance of tools 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

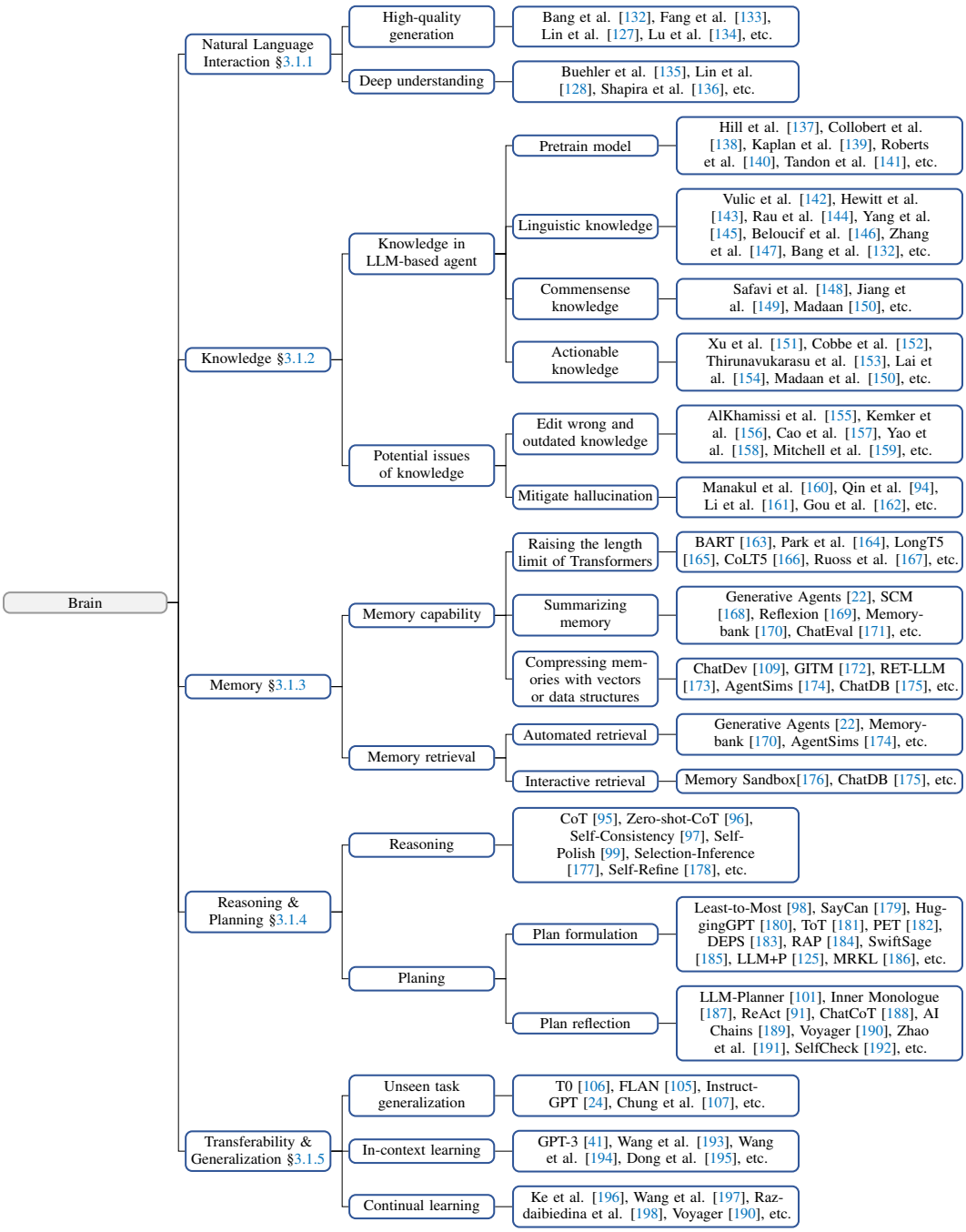

Figure 3: Typology of the brain module.

The human brain is a sophisticated structure comprised of a vast number of interconnected neurons, capable of processing various information, generating diverse thoughts, controlling different behaviors, and even creating art and culture [199]. Much like humans, the brain serves as the central nucleus of an AI agent, primarily composed of a large language model.

Operating mechanism. To ensure effective communication, the ability to engage in natural language interaction (§3.1.1) is paramount. After receiving the information processed by the perception module, the brain module first turns to storage, retrieving in knowledge (§3.1.2) and recalling from memory (§3.1.3). These outcomes aid the agent in devising plans, reasoning, and making informed decisions (§3.1.4). 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 transfer ability (§3.1.5). In the subsequent sections, we delve into a detailed exploration of these extraordinary facets of the brain module as depicted in Figure 3.

3.1.1 Natural Language Interaction

As a medium for communication, language contains a wealth of information. In addition to the intuitively expressed content, there may also be the speaker's beliefs, desires, and intentions hidden behind it [200]. Thanks to the powerful natural language understanding and generation capabilities inherent in LLMs [25; 201; 202; 203], agents can proficiently engage in not only basic interactive conversations [204; 205; 206] in multiple languages [132; 202] but also exhibit in-depth comprehension abilities, which allow humans to easily understand and interact with agents [207; 208]. Besides, LLM-based agents that communicate in natural language can earn more trust and cooperate more effectively with humans [130].

Multi-turn interactive conversation. The capability of multi-turn conversation is the foundation of effective and consistent communication. As the core of the brain module, LLMs, such as GPT series [40; 41; 201], LLaMA series [201; 209] and T5 series [107; 210], can understand natural language and generate coherent and contextually relevant responses, which helps agents to comprehend better and handle various problems [211]. However, even humans find it hard to communicate without confusion in one sitting, so multiple rounds of dialogue are necessary. Compared with traditional text-only reading comprehension tasks like SQuAD [212], multi-turn conversations (1) are interactive, involving multiple speakers, and lack continuity; (2) may involve multiple topics, and the information of the dialogue may also be redundant, making the text structure more complex [147]. In general, the multi-turn conversation is mainly divided into three steps: (1) Understanding the history of natural language dialogue, (2) Deciding what action to take, and (3) Generating natural language responses. LLM-based agents are capable of continuously refining outputs using existing information to conduct multi-turn conversations and effectively achieve the ultimate goal [132; 147].

High-quality natural language generation. Recent LLMs show exceptional natural language generation capabilities, consistently producing high-quality text in multiple languages [132; 213]. The coherency [214] and grammatical accuracy [133] of LLM-generated content have shown steady enhancement, evolving progressively from GPT-3 [41] to InstructGPT [24], and culminating in GPT-4 [25]. See et al. [214] empirically affirm that these language models can "adapt to the style and content of the conditioning text" [215]. And the results of Fang et al. [133] suggest that ChatGPT excels in grammar error detection, underscoring its powerful language capabilities. In conversational contexts, LLMs also perform well in key metrics of dialogue quality, including content, relevance, and appropriateness [127]. Importantly, they do not merely copy training data but display a certain degree of creativity, generating diverse texts that are equally novel or even more novel than the benchmarks crafted by humans [216]. Meanwhile, human oversight remains effective through the use of controllable prompts, ensuring precise control over the content generated by these language models [134].

Intention and implication understanding. Although models trained on the large-scale corpus are already intelligent enough to understand instructions, most are still incapable of emulating human dialogues or fully leveraging the information conveyed in language [217]. Understanding the implied meanings is essential for effective communication and cooperation with other intelligent agents [135],

and enables one to interpret others' feedback. The emergence of LLMs highlights the potential of foundation models to understand human intentions, but when it comes to vague instructions or other implications, it poses a significant challenge for agents [94; 136]. For humans, grasping the implied meanings from a conversation comes naturally, whereas for agents, they should formalize implied meanings into a reward function that allows them to choose the option in line with the speaker's preferences in unseen contexts [128]. One of the main ways for reward modeling is inferring rewards based on feedback, which is primarily presented in the form of comparisons [218] (possibly supplemented with reasons [219]) and unconstrained natural language [220]. Another way involves recovering rewards from descriptions, using the action space as a bridge [128]. Jeon et al. [221] suggests that human behavior can be mapped to a choice from an implicit set of options, which helps to interpret all the information in a single unifying formalism. By utilizing their understanding of context, agents can take highly personalized and accurate action, tailored to specific requirements.

3.1.2 Knowledge

Due to the diversity of the real world, many NLP researchers attempt to utilize data that has a larger scale. This data usually is unstructured and unlabeled [137; 138], yet it contains enormous knowledge that language models could learn. In theory, language models can learn more knowledge as they have more parameters [139], and it is possible for language models to learn and comprehend everything in natural language. Research [140] shows that language models trained on a large-scale dataset can encode a wide range of knowledge into their parameters and respond correctly to various types of queries. Furthermore, the knowledge can assist LLM-based agents in making informed decisions [222]. All of this knowledge can be roughly categorized into the following types:

- **Linguistic knowledge.** Linguistic knowledge [142; 143; 144] is represented as a system of constraints, a grammar, which defines all and only the possible sentences of the language. It includes morphology, syntax, semantics [145; 146], and pragmatics. Only the agents that acquire linguistic knowledge can comprehend sentences and engage in multi-turn conversations [147]. Moreover, these agents can acquire multilingual knowledge [132] by training on datasets that contain multiple languages, eliminating the need for extra translation models.
- **Commonsense knowledge.** Commonsense knowledge [148; 149; 150] 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, the models lacking the corresponding commonsense knowledge may fail to grasp or misinterpret the intended meaning [141]. 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 [151; 154; 150], mathematics [152], medicine [153], etc. It is essential for models to effectively solve problems within a particular domain [223]. For example, models designed to perform programming tasks need to possess programming knowledge, such as code format. Similarly, models intended for diagnostic purposes should possess medical knowledge like the names of specific diseases and prescription drugs.

Although LLMs demonstrate excellent performance in acquiring, storing, and utilizing knowledge [155], there remain potential issues and unresolved problems. For example, the knowledge acquired by models during training could become outdated or even be incorrect from the start. A simple way to address this is retraining. However, it requires advanced data, extensive time, and computing resources. Even worse, it can lead to catastrophic forgetting [156]. Therefore, some researchers[157; 158; 159] try editing LLMs to locate and modify specific knowledge stored within the models. This involved unloading incorrect knowledge while simultaneously acquiring new knowledge. Their experiments show that this method can partially edit factual knowledge, but its underlying mechanism still requires further research. Besides, LLMs may generate content that conflicts with the source or factual information [224], a phenomenon often referred to as hallucinations [225]. It is one of the critical reasons why LLMs can not be widely used in factually rigorous tasks. To tackle this issue, some researchers [160] proposed a metric to measure the level of hallucinations and provide developers with an effective reference to evaluate the trustworthiness of LLM outputs. Moreover, some researchers[161; 162] enable LLMs to utilize external tools[94; 226; 227] to avoid incorrect

knowledge. Both of these methods can alleviate the impact of hallucinations, but further exploration of more effective approaches is still needed.

3.1.3 Memory

In our framework, "memory" stores sequences of the agent's past observations, thoughts and actions, which is akin to the definition presented by Nuxoll et al. [228]. Just as the human brain relies on memory systems to retrospectively harness prior experiences for strategy formulation and decision-making, agents necessitate specific memory mechanisms to ensure their proficient handling of a sequence of consecutive tasks [229; 230; 231]. When faced with complex problems, memory mechanisms help the agent to revisit and apply antecedent strategies effectively. Furthermore, these memory mechanisms enable individuals to adjust to unfamiliar environments by drawing on past experiences.

With the expansion of interaction cycles in LLM-based agents, two primary challenges arise. The first pertains to the sheer length of historical records. LLM-based agents process prior interactions in natural language format, appending historical records to each subsequent input. As these records expand, they might surpass the constraints of the Transformer architecture that most LLM-based agents rely on. When this occurs, the system might truncate some content. The second challenge is the difficulty in extracting relevant memories. As agents amass a vast array of historical observations and action sequences, they grapple with an escalating memory burden. This makes establishing connections between related topics increasingly challenging, potentially causing the agent to misalign its responses with the ongoing context.

Methods for better memory capability. Here we introduce several methods to enhance the memory of LLM-based agents.

- **Raising the length limit of Transformers.** The first method tries to address or mitigate the inherent sequence length constraints. The Transformer architecture struggles with long sequences due to these intrinsic limits. As sequence length expands, computational demand grows exponentially due to the pairwise token calculations in the self-attention mechanism. Strategies to mitigate these length restrictions encompass text truncation [163; 164; 232], segmenting inputs [233; 234], and emphasizing key portions of text [235; 236; 237]. Some other works modify the attention mechanism to reduce complexity, thereby accommodating longer sequences [238; 165; 166; 167].
- **Summarizing memory.** The second strategy for amplifying memory efficiency hinges on the concept of memory summarization. This ensures agents effortlessly extract pivotal details from historical interactions. Various techniques have been proposed for summarizing memory. Using prompts, some methods succinctly integrate memories [168], while others emphasize reflective processes to create condensed memory representations [22; 239]. Hierarchical methods streamline dialogues into both daily snapshots and overarching summaries [170]. Notably, specific strategies translate environmental feedback into textual encapsulations, bolstering agents' contextual grasp for future engagements [169]. Moreover, in multi-agent environments, vital elements of agent communication are captured and retained [171].
- **Compressing memories with vectors or data structures.** By employing suitable data structures, intelligent agents boost memory retrieval efficiency, facilitating prompt responses to interactions. Notably, several methodologies lean on embedding vectors for memory sections, plans, or dialogue histories [109; 170; 172; 174]. Another approach translates sentences into triplet configurations [173], while some perceive memory as a unique data object, fostering varied interactions [176]. Furthermore, ChatDB [175] and DB-GPT [240] integrate the LLMrollers with SQL databases, enabling data manipulation through SQL commands.

Methods for memory retrieval. When an agent interacts with its environment or users, it is imperative to retrieve the most appropriate content from its memory. This ensures that the agent accesses relevant and accurate information to execute specific actions. An important question arises: How can an agent select the most suitable memory? Typically, agents retrieve memories in an automated manner [170; 174]. A significant 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 [22].

Some research introduces the concept of interactive memory objects, which are representations of dialogue history that can be moved, edited, deleted, or combined through summarization. Users can view and manipulate these objects, influencing how the agent perceives the dialogue [176]. Similarly, other studies allow for memory operations like deletion based on specific commands provided by users [175]. Such methods ensure that the memory content aligns closely with user expectations.

3.1.4 Reasoning and Planning

Reasoning. Reasoning, underpinned by evidence and logic, is fundamental to human intellectual endeavors, serving as the cornerstone for problem-solving, decision-making, and critical analysis [241; 242; 243]. Deductive, inductive, and abductive are the primary forms of reasoning commonly recognized in intellectual endeavor [244]. For LLM-based agents, like humans, reasoning capacity is crucial for solving complex tasks [25].

Differing academic views exist regarding the reasoning capabilities of large language models. Some argue language models possess reasoning during pre-training or fine-tuning [244], while others believe it emerges after reaching a certain scale in size [26; 245]. Specifically, the representative Chain-of-Thought (CoT) method [95; 96] has been demonstrated to elicit the reasoning capacities of large language models by guiding LLMs to generate rationales before outputting the answer. Some other strategies have also been presented to enhance the performance of LLMs like self-consistency [97], self-polish [99], self-refine [178] and selection-inference [177], among others. Some studies suggest that the effectiveness of step-by-step reasoning can be attributed to the local statistical structure of training data, with locally structured dependencies between variables yielding higher data efficiency than training on all variables [246].

Planning. Planning is a key strategy humans employ when facing complex challenges. For humans, planning helps organize thoughts, set objectives, and determine the steps to achieve those objectives [247; 248; 249]. Just as with humans, the ability to plan is crucial for agents, and central to this planning module is the capacity for reasoning [250; 251; 252]. This offers a structured thought process for agents based on LLMs. Through reasoning, agents deconstruct complex tasks into more manageable sub-tasks, devising appropriate plans for each [253; 254]. Moreover, as tasks progress, agents can employ introspection to modify their plans, ensuring they align better with real-world circumstances, leading to adaptive and successful task execution.

Typically, planning comprises two stages: plan formulation and plan reflection.

- **Plan formulation.** During the process of plan formulation, agents generally decompose an overarching task into numerous sub-tasks, and various approaches have been proposed in this phase. Notably, some works advocate for LLM-based agents to decompose problems comprehensively in one go, formulating a complete plan at once and then executing it sequentially [98; 179; 255; 256]. 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 [95; 96; 257]. Additionally, some methods emphasize hierarchical planning [182; 185], 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 [97; 181; 184; 258; 184]. 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 [125; 130; 186; 259].

- **Plan reflection.** Upon formulating a plan, it's imperative to reflect upon and evaluate its merits. LLM-based agents leverage internal feedback mechanisms, often drawing insights from pre-existing models, to hone and enhance their strategies and planning approaches [169; 178; 188; 192]. To better align with human values and preferences, agents actively engage with humans, allowing them to rectify some misunderstandings and assimilate this tailored feedback into their planning methodology [108; 189; 190]. Furthermore, they could draw feedback from tangible or virtual surroundings, such as cues from task accomplishments or post-action observations, aiding them in revising and refining their plans [91; 101; 187; 191; 260].

3.1.5 Transferability and Generalization

Intelligence shouldn't be limited to a specific domain or task, but rather encompass a broad range of cognitive skills and abilities [31]. The remarkable nature of the human brain is largely attributed to its high degree of plasticity and adaptability. It can continuously adjust its structure and function 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 on large-scale corpora can learn universal language representations [36; 261; 262]. Leveraging the power of pre-trained models, with only a small amount of data for fine-tuning, LLMs can demonstrate excellent performance in downstream tasks [263]. There is no need to train new models from scratch, which saves a lot of computation resources. However, through this task-specific fine-tuning, the models lack versatility and struggle to be generalized 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 [24; 105; 106].

Unseen task generalization. Studies show that instruction-tuned LLMs exhibit zero-shot generalization without the need for task-specific fine-tuning [24; 25; 105; 106; 107]. With the expansion of model size and corpus size, LLMs gradually exhibit remarkable emergent abilities in unfamiliar tasks [132]. Specifically, LLMs can complete new tasks they do not encounter in the training stage by following the instructions based on their own understanding. One of the implementations is multi-task learning, for example, FLAN [105] finetunes language models on a collection of tasks described via instructions, and T0 [106] introduces a unified framework that converts every language problem into a text-to-text format. Despite being purely a language model, GPT-4 [25] demonstrates remarkable capabilities in a variety of domains and tasks, including abstraction, comprehension, vision, coding, mathematics, medicine, law, understanding of human motives and emotions, and others [31]. It is noticed that the choices in prompting are critical for appropriate predictions, and training directly on the prompts can improve the models' robustness in generalizing to unseen tasks [264]. Promisingly, such generalization capability can further be enhanced by scaling up both the model size and the quantity or diversity of training instructions [94; 265].

In-context learning. Numerous studies indicate that LLMs can perform a variety of complex tasks through in-context learning (ICL), which refers to the models' ability to learn from a few examples in the context [195]. Few-shot in-context learning enhances the predictive performance of language models by concatenating the original input with several complete examples as prompts to enrich the context [41]. The key idea of ICL is learning from analogy, which is similar to the learning process of humans [266]. Furthermore, since the prompts are written in natural language, the interaction is interpretable and changeable, making it easier to incorporate human knowledge into LLMs [95; 267]. Unlike the supervised learning process, ICL doesn't involve fine-tuning or parameter updates, which could greatly reduce the computation costs for adapting the models to new tasks. Beyond text, researchers also explore the potential ICL capabilities in different multimodal tasks [193; 194; 268; 269; 270; 271], making it possible for agents to be applied to large-scale real-world tasks.

Continual learning. Recent studies [190; 272] have highlighted the potential of LLMs' planning capabilities in facilitating continuous learning [196; 197] for agents, which involves continuous acquisition and update of skills. A core challenge in continual learning is catastrophic forgetting [273]: as a model learns new tasks, it tends to lose knowledge from previous tasks. Numerous efforts have been devoted to addressing the above challenge, which can be broadly separated into three groups, introducing regularly used terms in reference to the previous model [274; 275; 276; 277], approximating prior data distributions [278; 279; 280], and designing architectures with task-adaptive parameters [281; 198]. LLM-based agents have emerged as a novel paradigm, leveraging the planning capabilities of LLMs to combine existing skills and address more intricate challenges. Voyager [190] attempts to solve progressively harder tasks proposed by the automatic curriculum devised by GPT-4 [25]. By synthesizing complex skills from simpler programs, the agent not only rapidly enhances its capabilities but also effectively counters catastrophic forgetting.

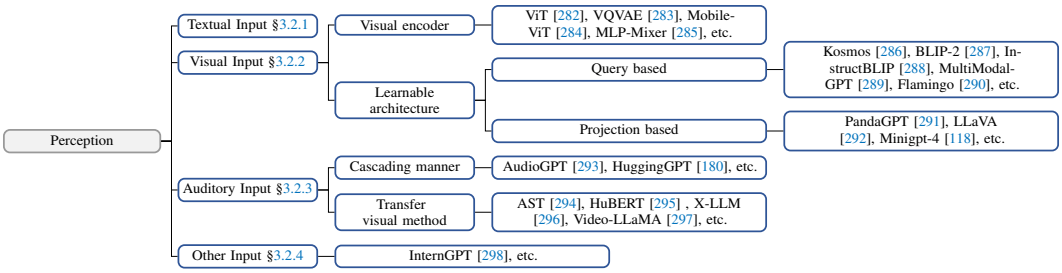

Figure 4: Typology of the perception module.

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 [299; 300], allowing us to perceive and interact with the world. Similarly, it's 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 (§ 3.2.1), visual (§ 3.2.2), and auditory inputs (§ 3.2.3). We also consider other potential input forms (§ 3.2.4) such as tactile feedback, gestures, and 3D maps to enrich the agent's perception domain and enhance its versatility.3). The typology diagram for the LLM-based agent perception is depicted in Figure 4.

3.2.1 Textual Input

Text is a way to carry data, information, and knowledge, making text communication one of the most important ways humans interact with the world. An LLM-based agent already has the fundamental ability to communicate with humans through textual input and output [114]. In a user's textual input, aside from the explicit content, there are also beliefs, desires, and intentions hidden behind it. 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, as discussed in § 3.1.1, understanding implied meanings within textual input remains challenging for the current LLM-based agent. For example, some works [128; 218; 219; 220] employ 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 new tasks. Understanding text instructions for unknown tasks places higher demands on the agent's text perception abilities. As described in § 3.1.5, an LLM that has undergone instruction tuning [105] 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 exhibit outstanding performance in language comprehension [25; 301] and multi-turn conversations [302], 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 [120], 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 [303; 304; 305; 306; 307]. Captions can be directly linked with standard text instructions and fed into the agent. This approach is highly interpretable and doesn't require additional training for caption generation, which can save a significant number of computational resources. However, caption

generation is a low-bandwidth method [120; 308], and it may lose a lot of potential information during the conversion process. Furthermore, the agent's focus on images may introduce biases.

Inspired by the excellent performance of transformers [309] in natural language processing, researchers have extended their use to the field of computer vision. Representative works like ViT/VQVAE [282; 283; 284; 285; 310] have successfully encoded visual information using transformers. Researchers first divide an image into fixed-size patches and then treat these patches, after linear projection, as input tokens for Transformers [292]. In the end, by calculating self-attention between tokens, they are able to integrate information across the entire image, resulting in a highly effective way to perceive visual content. Therefore, some works [311] try to combine the image encoder and LLM directly to train the entire model in an end-to-end way. While the agent can achieve remarkable visual perception abilities, it comes at the cost of substantial computational resources.

Extensively pre-trained visual encoders and LLMs can greatly enhance the agent's visual perception and language expression abilities [286; 312]. Freezing one or both of them during training is a widely adopted paradigm that achieves a balance between training resources and model performance [287]. However, LLMs cannot directly understand the output of a visual encoder, so it's 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 [287] and InstructBLIP [288] use the Querying Transformer(Q-Former) module as an intermediate layer between the visual encoder and the LLM [288]. Q-Former is a transformer that employs learnable query vectors [289], giving it the capability to extract language-informative visual representations. It can provide the most valuable information to the LLM, reducing the agent's burden of learning visual-language alignment and thereby mitigating the issue of catastrophic forgetting. At the same time, some researchers adopt a computationally efficient method by using a single projection layer to achieve visual-text alignment, reducing the need for training additional parameters [118; 291; 312]. Moreover, the projection layer can effectively integrate with the learnable interface to adapt the dimensions of its outputs, making them compatible with LLMs [296; 297; 313; 314].

Video input consists of a series of continuous image frames. As a result, the methods used by agents to perceive images [287] may be applicable to the realm of videos, allowing the agent to have good perception of video inputs as well. Compared to image information, video information adds a temporal dimension. Therefore, the agent's understanding of the relationships between different frames in time is crucial for perceiving video information. Some works like Flamingo [290; 315] ensure temporal order when understanding videos using a mask mechanism. The mask mechanism restricts the agent's view to only access visual information from frames that occurred earlier in time when it perceives a specific frame in the video.

3.2.3 Auditory Input

Undoubtedly, auditory information is 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. Indeed, there are numerous well-established models and approaches [293; 316; 317] for processing audio as a standalone modality. However, these models often excel at specific tasks. Given the excellent tool-using capabilities of LLMs (which will be discussed in detail in §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. For instance, AudioGPT [293], makes full use of the capabilities of models like FastSpeech [317], GenerSpeech [316], Whisper [316], and others [318; 319; 320; 321; 322] which have achieved excellent results in tasks such as Text-to-Speech, Style Transfer, and Speech Recognition.

An audio spectrogram provides an intuitive representation of the frequency spectrum of an audio signal as it changes over time [323]. 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 [294; 295] efforts aim to migrate perceptual methods from the visual domain to audio. AST (Audio Spectrogram Transformer) [294] 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. Moreover, some researchers [296; 297] have drawn inspiration from the idea of freezing encoders to reduce training

time and computational costs. They align audio encoding with data encoding from other modalities by adding the same learnable interface layer.

3.2.4 Other Input

As mentioned earlier, many studies have looked into perception units for text, visual, and audio. However, LLM-based agents might 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 have a clear sense of the temperature, humidity, and brightness in their surroundings, enabling them to take environment-aware actions. Moreover, by efficiently integrating basic perceptual abilities like vision, text, and light sensitivity, agents can develop various user-friendly perception modules for humans. InternGPT [298] introduces pointing instructions. Users can interact with specific, hard-to-describe portions of an image by using gestures or moving the cursor to select, drag, or draw. The addition of pointing instructions helps provide more precise specifications for individual text instructions. Building upon this, agents have the potential to perceive more complex user inputs. For example, technologies such as eye-tracking in AR/VR devices, body motion capture, and even brainwave signals in brain-computer interaction.

Finally, a human-like LLM-based agent should possess awareness of a broader overall environment. At present, numerous mature and widely adopted hardware devices can assist agents in accomplishing this. Lidar [324] can create 3D point cloud maps to help agents detect and identify objects in their surroundings. GPS [325] can provide accurate location coordinates and can be integrated with map data. Inertial Measurement Units (IMUs) can measure and record the three-dimensional motion of objects, offering details about an object's speed and direction. However, these sensory data are complex and cannot be directly understood by LLM-based agents. Exploring how agents can perceive more comprehensive input is a promising direction for the future.

3.3 Action

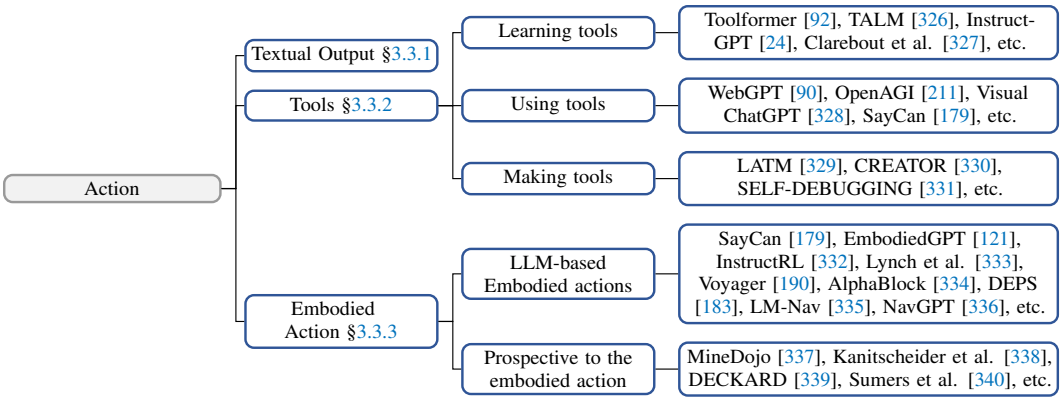

Figure 5: Typology of the action module.

After humans perceive their environment, their brains integrate, analyze, and reason with the perceived information and 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 the construction of the agent, the action module receives action sequences sent by the brain module and carries out actions to interact with the environment. As Figure 5 shows, this section begins with textual output (§ 3.3.1), which is the inherent capability of LLM-based agents. Next we talk about the tool-using capability of LLM-based agents (§ 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 (§ 3.3.3).

3.3.1 Textual Output

As discussed in § 3.1.1, the rise and development of Transformer-based generative large language models have endowed LLM-based agents with inherent language generation capabilities [132; 213]. The text quality they generate excels in various aspects such as fluency, relevance, diversity, controllability [127; 214; 134; 216]. Consequently, LLM-based agents can be exceptionally strong language generators.

3.3.2 Tool Using

Tools are extensions of the capabilities of tool users. 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 [94].

LLM-based agents have *limitations* in some aspects, and the use of tools can *strengthen the agents' capabilities*. First, although LLM-based agents have a strong knowledge base and expertise, they don't have the ability to memorize every piece of training data [341; 342]. They may also fail to steer to correct knowledge due to the influence of contextual prompts [226], or even generate hallucinate knowledge [208]. Coupled with the lack of corpus, training data, and tuning for specific fields and scenarios, agents' expertise is also limited when specializing in specific domains [343]. Specialized tools enable LLMs to enhance their expertise, adapt domain knowledge, and be more suitable for domain-specific needs in a pluggable form. Furthermore, the decision-making process of LLM-based agents lacks transparency, making them less trustworthy in high-risk domains such as healthcare and finance [344]. Additionally, LLMs are susceptible to adversarial attacks [345], and their robustness against slight input modifications is inadequate. In contrast, agents that accomplish tasks with the assistance of tools exhibit stronger interpretability and robustness. The execution process of tools can reflect the agents' approach to addressing complex requirements and enhance the credibility of their decisions. Moreover, for the reason that tools are specifically designed for their respective usage scenarios, agents utilizing such tools are better equipped to handle slight input modifications and are more resilient against adversarial attacks [94].

LLM-based agents not only require the use of tools, but are also *well-suited* for tool integration. Leveraging the rich world knowledge accumulated through the pre-training process and CoT prompting, LLMs have demonstrated remarkable reasoning and decision-making abilities in complex interactive environments [97], which help agents break down and address tasks specified by users in an appropriate way. What's more, LLMs show significant potential in intent understanding and other aspects [25; 201; 202; 203]. When agents are combined with tools, the threshold for tool utilization can be lowered, thereby fully unleashing the creative potential of human users [94].

Understanding tools. A prerequisite for an agent to use tools effectively is a comprehensive understanding of the tools' application scenarios and invocation methods. Without this understanding, the process of the agent using tools will become untrustworthy and fail to genuinely enhance the agent's capabilities. Leveraging the powerful zero-shot and few-shot learning abilities of LLMs [40; 41], agents can acquire knowledge about tools by utilizing *zero-shot prompts* that describe tool functionalities and parameters, or *few-shot prompts* that provide demonstrations of specific tool usage scenarios and corresponding methods [92; 326]. These learning approaches parallel human methods of learning by consulting tool manuals or observing others using tools [94]. A single tool is often insufficient when facing complex tasks. Therefore, the agents should first decompose the complex task into subtasks in an appropriate manner, and their understanding of tools play a significant role in task decomposition.

Learning to use tools. The methods for agents to learn to utilize tools primarily consist of *learning from demonstrations* and *learning from feedback*. This involves mimicking the behavior of human experts [346; 347; 348], as well as understanding the consequences of their actions and making adjustments based on feedback received from both the environment and humans [24; 349; 350]. 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 [94].

If an agent rigidly applies tools without *adaptability*, it cannot achieve acceptable performance in all scenarios. Agents need to generalize their tool usage skills learned in specific contexts to more general situations, such as transferring a model trained on Yahoo search to Google search. To accomplish this, it's necessary for agents to grasp the common principles or patterns in tool usage strategies, which can potentially be achieved through meta-tool learning [327]. Enhancing the agent's understanding of relationships between simple and complex tools, such as how complex tools are built on simpler ones, can contribute to the agents' capacity to generalize tool usage. This allows agents to effectively discern nuances across various application scenarios and transfer previously learned knowledge to new tools [94]. Curriculum learning [351], which allows an agent to start from simple tools and progressively learn complex ones, aligns with the requirements. Moreover, benefiting from the understanding of user intent reasoning and planning abilities, agents can better design methods of tool utilization and collaboration and then provide higher-quality outcomes.

Making tools for self-sufficiency. Existing tools are often designed for human convenience, which might not be optimal for agents. To make agents use tools better, there's a need for tools specifically designed for agents. These tools should be more modular and have input-output formats that are more suitable for agents. If instructions and demonstrations are provided, LLM-based agents also possess the ability to create tools by generating executable programs, or integrating existing tools into more powerful ones [94; 330; 352]. and they can learn to perform self-debugging [331]. Moreover, if the agent that serves as a tool maker successfully creates a tool, it can produce packages containing the tool's code and demonstrations for other agents in a multi-agent system, in addition to using the tool itself [329]. Speculatively, in the future, agents might become self-sufficient and exhibit a high degree of autonomy in terms of tools.

Tools can expand 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 [92]. 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 [211; 353]. 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 [90; 175]. What's 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 [354; 355]. For multi-agent systems, communication tools (e.g., emails) may serve as a means for agents to interact with each other under strict security constraints, facilitating their *collaboration*, and showing autonomy and flexibility [94].

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. For example, image processing and generation can be accomplished by an agent that draws on a visual model [328]. In aerospace engineering, agents are being explored for modeling physics and solving complex differential equations [356]; in the field of robotics, agents are required to plan physical operations and control the robot execution [179]; and so on. Agents that are capable of dynamically interacting with the environment or the world through tools, or in a multimodal manner, can be referred to as digitally embodied [94]. The *embodiment* of agents has been a central focus of embodied learning research. We will make a deep discussion on agents' embodied action in §3.3.3.

3.3.3 Embodied Action

In the pursuit of Artificial General Intelligence (AGI), the embodied agent is considered a pivotal paradigm while it strives to integrate model intelligence with the physical world. *The Embodiment hypothesis* [357] 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. Similarly, unlike traditional deep learning models that learn explicit capabilities from the internet datasets to solve domain problems, people anticipate that LLM-based agents' behaviors will no longer be limited to pure text output or calling exact tools to perform

particular domain tasks [358]. Instead, they should be capable of actively perceiving, comprehending, and interacting with physical environments, making decisions, and generating specific behaviors to modify the environment based on LLM's extensive 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. Before the widespread rise of LLMs, researchers tended to use methods like reinforcement learning to explore the embodied actions of agents. Despite the extensive success of RL-based embodiment [359; 360; 361], it does have certain limitations in some aspects. In brief, RL algorithms face limitations in terms of data efficiency, generalization, and complex problem reasoning due to challenges in modeling the dynamic and often ambiguous real environment, or their heavy reliance on precise reward signal representations [362]. Recent studies have indicated that leveraging the rich internal knowledge acquired during the pre-training of LLMs can effectively alleviate these issues [120; 187; 258; 363].

- **Cost efficiency.** Some on-policy algorithms struggle with sample efficiency as they require fresh data for policy updates while gathering enough embodied data for high-performance training is costly and noisy. The constraint is also found in some end-to-end models [364; 365; 366]. By leveraging the intrinsic knowledge from LLMs, agents like PaLM-E [120] 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 section §3.1.5, an agent's competence should extend beyond specific tasks. When faced with intricate, uncharted real-world environments, it's imperative that the agent exhibits dynamic learning and generalization capabilities. However, the majority of RL algorithms are designed to train and evaluate relevant skills for specific tasks [101; 367; 368; 369]. In contrast, fine-tuned by diverse forms and rich task types, LLMs have showcased remarkable cross-task generalization capabilities [370; 371]. For instance, PaLM-E exhibits surprising zero-shot or one-shot generalization capabilities to new objects or novel combinations of existing objects [120]. Further, language proficiency represents a distinctive advantage of LLM-based agents, serving both as a means to interact with the environment and as a medium for transferring foundational skills to new tasks [372]. SayCan [179] 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 [101]. To address this, Voyager [190] 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 constitutes a pivotal strategy employed by humans in response to complex problems as well as LLM-based agents. Before LLMs exhibited remarkable reasoning abilities, researchers introduced Hierarchical Reinforcement Learning (HRL) methods while the high-level policy constraints sub-goals for the low-level policy and the low-level policy produces appropriate action signals [373; 374; 375]. Similar to the role of high-level policies, LLMs with emerging reasoning abilities [26] can be seamlessly applied to complex tasks in a zero-shot or few-shot manner [95; 97; 98; 99]. In addition, external feedback from the environment can further enhance LLM-based agents' planning performance. Based on the current environmental feedback, some work [101; 91; 100; 376] dynamically generate, maintain, and adjust high-level action plans in order to minimize dependency on prior knowledge in partially observable environments, thereby grounding the plan. Feedback can also come from models or humans, which can usually be referred to as the critics, assessing task completion based on the current state and task prompts [25; 190].

Embodied actions for LLM-based agents. Depending on the agents' level of autonomy in a task or the complexity of actions, there are several fundamental LLM-based embodied actions, primarily including observation, manipulation, and navigation.

- **Observation.** Observation constitutes the primary ways by which the agent acquires environmental information and updates states, playing a crucial role in enhancing the efficiency of subsequent embodied actions. As mentioned in §3.2, observation by embodied agents primarily occurs in environments with various inputs, which 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 and special tokens are marked to denote the positions of multimodal data [120; 332; 121]. Soundspaces [377] proposes the identification of physical spatial geometric

elements guided by reverberant audio input, enhancing the agent's observations with a more comprehensive perspective [375]. In recent times, even more research takes audio as a modality for embedded observation. Apart from the widely employed cascading paradigm [293; 378; 316], audio information encoding similar to ViT further enhances the seamless integration of audio with other modalities of inputs [294]. The agent's observation of the environment can also be derived from real-time linguistic instructions from humans, while human feedback helps the agent in acquiring detail information that may not be readily obtained or parsed [333; 190].

- **Manipulation.** In general, manipulation tasks for embodied agents include object rearrangements, tabletop manipulation, and mobile manipulation [23; 120]. 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, as well as cleaning the tabletop [179]. Besides precise observation, this involves combining a series of subgoals by leveraging LLM. Consequently, maintaining synchronization between the agent's state and the subgoals is of significance. DEPS [183] utilizes an LLM-based interactive planning approach to maintain this consistency and help error correction from agent's feedback throughout the multi-step, long-haul reasoning process. In contrast to these, AlphaBlock [334] focuses on more challenging manipulation tasks (e.g. making a smiley face using building blocks), which requires the agent to have a more grounded understanding of the instructions. Unlike the existing open-loop paradigm, AlphaBlock constructs a dataset comprising 35 complex high-level tasks, along with corresponding multi-step planning and observation pairs, and then fine-tunes a multimodal model to enhance its comprehension of high-level cognitive instructions.
- **Navigation.** Navigation permits agents to dynamically alter their positions within the environment, which often involves multi-angle and multi-object observations, as well as long-horizon manipulations based on current exploration [23]. Before navigation, it is essential for embodied agents to establish prior internal maps about the external environment, which are typically in the form of a topological map, semantic map or occupancy map [358]. For example, LM-Nav [335] utilizes the VNM [379] to create an internal topological map. It further leverages the LLM and VLM for decomposing input commands and analyzing the environment to find the optimal path. Furthermore, some [380; 381] highlight the importance of spatial representation to achieve the precise localization of spatial targets rather than conventional point or object-centric navigation actions by leveraging the pre-trained VLM model to combine visual features from images with 3D reconstructions of the physical world [358]. Navigation is usually a long-horizon task, where the upcoming states of the agent are influenced by its past actions. A memory buffer and summary mechanism are needed to serve as a reference for historical information [336], which is also employed in Smallville and Voyager [22; 190; 382; 383]. Additionally, as mentioned in §3.2, some works have proposed the audio input is also of great significance, but integrating audio information presents challenges in associating it with the visual environment. A basic framework includes a dynamic path planner that uses visual and auditory observations along with spatial memories to plan a series of actions for navigation [375; 384].

By integrating these, the agent can accomplish more complex tasks, such as embodied question answering, whose primary objective is autonomous exploration of the environment, and responding to pre-defined multimodal questions, such as *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 [358].

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 for achieving specific sub-goals. The low-level policy can be a robotic transformer [120; 385; 386], which takes images and instructions as inputs and produces control commands for the end effector as well as robotic arms in particular embodied tasks [179]. Recently, in virtual embodied environments, the high-level strategies are utilized to control agents in gaming [172; 183; 190; 337] or simulated worlds [22; 108; 109]. For instance, Voyager [190] calls the Mineflayer [387] API interface to continuously acquire various skills and explore the world.

Prospective future of the embodied action. LLM-based embodied actions are seen as the bridge between virtual intelligence and the physical world, enabling agents to perceive and modify the environment much like humans. However, there remain several constraints such as high costs of physical-world robotic operators and the scarcity of embodied datasets, which foster a growing

interest in investigating agents' embodied actions within simulated environments like Minecraft [183; 338; 337; 190; 339]. By utilizing the Mineflayer [387] API, these investigations enable cost-effective examination of a wide range of embodied agents' operations including exploration, planning, self-improvement, and even lifelong learning [190]. Despite notable progress, achieving optimal embodied actions remains a challenge due to the significant disparity between simulated platforms and the physical world. To enable the effective deployment of embodied agents in real-world scenarios, an increasing demand exists for embodied task paradigms and evaluation criteria that closely mirror real-world conditions [358]. On the other hand, learning to ground language for agents is also an obstacle. For example, expressions like "jump down like a cat" primarily convey a sense of lightness and tranquility, but this linguistic metaphor requires adequate world knowledge [30]. [340] endeavors to amalgamate text distillation with Hindsight Experience Replay (HER) to construct a dataset as the supervised signal for the training process. Nevertheless, additional investigation on grounding embodied datasets still remains necessary while embodied action plays an increasingly pivotal role across various domains in human life.

4 Agents in Practice: Harnessing AI for Good

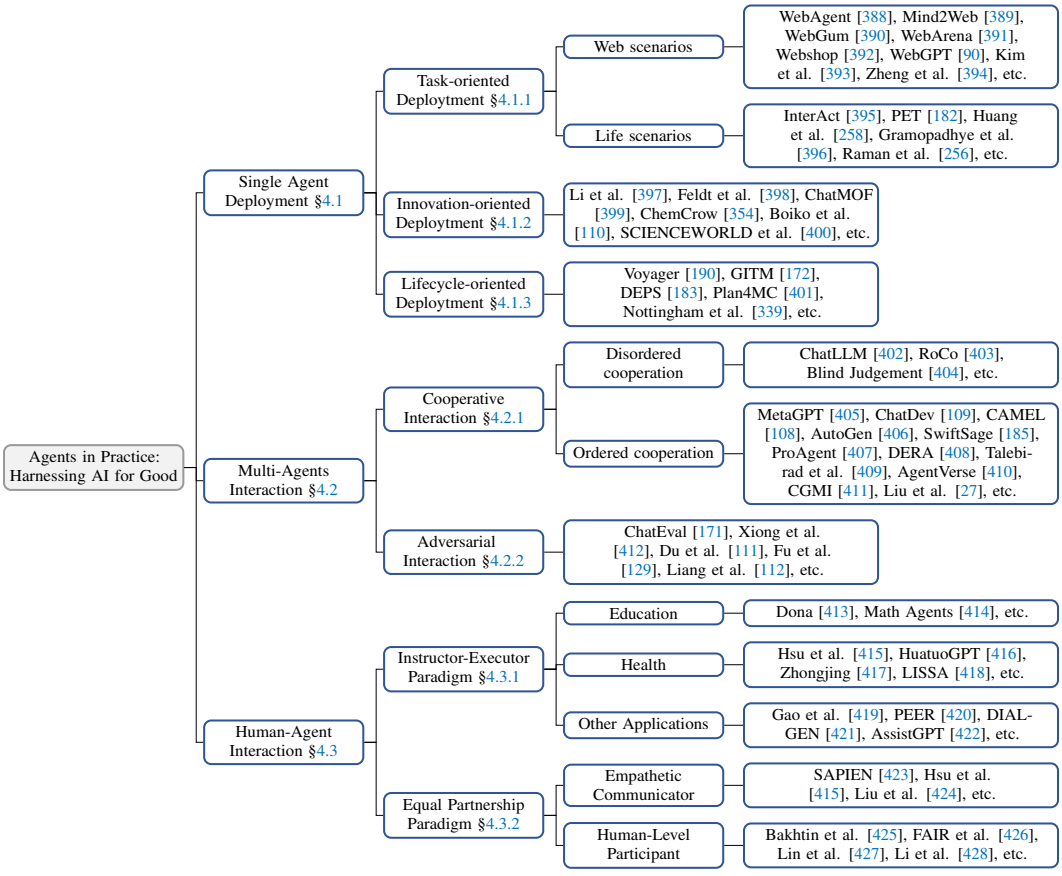

Figure 6: Typology of applications of LLM-based agents.

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. We can state with great confidence that, the possibility of having a personal agent capable of assisting users with typical daily tasks is larger than ever before [398]. As an LLM-based agent, its design objective should always be beneficial to humans, i.e., humans can *harness AI for good*. Specifically, we expect the agent to achieve the following objectives:

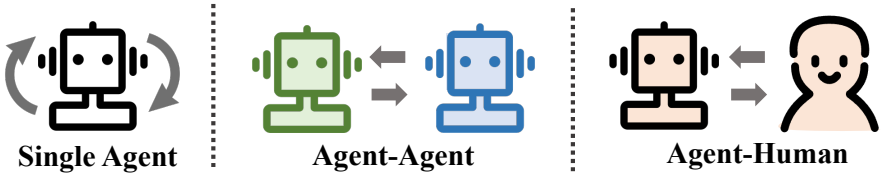

Figure 7: Scenarios of LLM-based agent applications. We mainly introduce three scenarios: single-agent deployment, multi-agent interaction, and human-agent interaction. A **single agent** possesses diverse capabilities and can demonstrate outstanding task-solving performance in various application orientations. 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.

1. Assist users in breaking free from daily tasks and repetitive labor, thereby Alleviating human work pressure and enhancing task-solving efficiency.
2. No longer necessitates users to provide explicit low-level instructions. Instead, the agent can independently analyze, plan, and solve problems.
3. After freeing users' hands, the agent also liberates their minds to engage in exploratory and innovative work, realizing their full potential in cutting-edge scientific fields.

In this section, we provide an in-depth overview of current applications of LLM-based agents, aiming to offer a broad perspective for the practical deployment scenarios (see Figure 7). First, we elucidate the diverse application scenarios of Single Agent, including task-oriented, innovation-oriented, and lifecycle-oriented scenarios (§ 4.1). Then, we present the significant coordinating potential of Multiple Agents. Whether through cooperative interaction for complementarity or adversarial interaction for advancement, both approaches can lead to higher task efficiency and response quality (§ 4.2). Finally, we categorize the interactive collaboration between humans and agents into two paradigms and introduce the main forms and specific applications respectively (§ 4.3). The topological diagram for LLM-based agent applications is depicted in Figure 6.

4.1 General Ability of Single Agent

Currently, there is a vibrant development of application instances of LLM-based agents [429; 430; 431]. AutoGPT [114] is one of the ongoing popular open-source projects aiming to achieve a fully autonomous system. Apart from the basic functions of large language models like GPT-4, the AutoGPT framework also incorporates various practical external tools and long/short-term 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 8, we introduce the astonishingly diverse capabilities that the agent exhibits in scenarios where only one single agent is present.

4.1.1 Task-oriented Deployment

The LLM-based agents, which can understand human natural language commands and perform everyday tasks [391], 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 [182; 258; 388; 394], sequence planning of sub-goals [182; 395], interactive exploration of the environment [256; 391; 390; 392], until the final objective is achieved.

To explore whether agents can perform basic tasks, they are first deployed in text-based game scenarios. In this type of game, agents interact with the world purely using natural language [432]. By reading textual descriptions of their surroundings and utilizing skills like memory, planning,

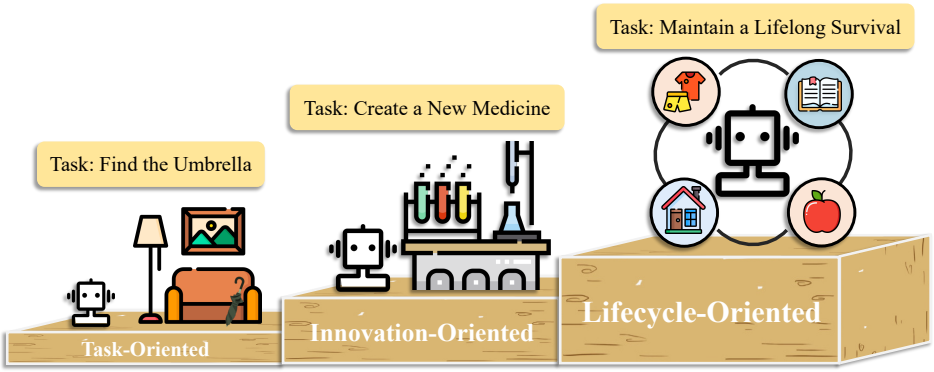

Figure 8: Practical applications of the single LLM-based agent in different scenarios. In **task-oriented deployment**, agents assist human users in solving daily tasks. They need to possess basic instruction comprehension and task decomposition abilities. In **innovation-oriented deployment**, agents 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.

and trial-and-error [182], they predict the next action. However, due to the limitation of foundation language models, agents often rely on reinforcement learning during actual execution [432; 433; 434].

With the gradual evolution of LLMs [301], agents equipped with stronger text understanding and generation abilities have demonstrated great potential to perform tasks through natural language. Due to their oversimplified nature, naive text-based scenarios have been inadequate as testing grounds for LLM-based agents [391]. More realistic and complex simulated test environments have been constructed to meet the demand. Based on task types, we divide these simulated 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 [390]. 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 [391]. In this way, agents can achieve accessibility and automation when dealing with unseen tasks in the future [435], ultimately freeing humans from repeated interactions with computer UIs.

Agents trained through reinforcement learning can effectively mimic human behavior using predefined actions like typing, searching, navigating to the next page, etc. They perform well in basic tasks such as online shopping [392] and search engine retrieval [90], which have been widely explored. However, agents without LLM capabilities may struggle to adapt to the more realistic and complex scenarios in the real-world Internet. In dynamic, content-rich web pages such as online forums or online business management [391], agents often face challenges in performance.

In order to enable successful interactions between agents and more realistic web pages, some researchers [393; 394] 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 [389] combines multiple LLMs fine-tuned for HTML, allowing them to summarize verbose HTML code [388] in real-world scenarios and extract valuable information. Furthermore, WebGum [390] empowers agents with visual perception abilities by employing a multimodal corpus containing HTML screenshots. It simultaneously fine-tunes the LLM and a visual encoder, deepening the agent's comprehensive understanding of web pages.

In life scenarios. In many daily household tasks in life scenarios, it's essential for agents to understand implicit instructions and apply common-sense knowledge [433]. 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 [432]. More realistic scenarios often lead to more obscure and subtle tasks. For example, the agent should proactively turn it on if it's dark and there's a light in the room. To successfully chop some vegetables in the kitchen, the agent needs to anticipate the possible location of a knife [182].

Can an agent apply the world knowledge embedded in its training data to real interaction scenarios? Huang et al. [258] lead the way in exploring this question. They demonstrate 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" [396].

To provide agents with access to comprehensive scenario information during interactions, 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 [395; 396]. Wu et al. [182] introduce the PET framework, which mitigates irrelevant objects and containers in environmental information through an early error correction method [256]. PET encourages agents to explore the scenario and plan actions more efficiently, focusing on the current sub-task.

4.1.2 Innovation-oriented Deployment

The LLM-based agent has demonstrated strong capabilities in performing tasks and enhancing the efficiency of repetitive work. 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 [399]: 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. This greatly weakens the agent's cognitive level. 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 [400; 436]. If the ability for autonomous exploration could be discovered within the agent, it would undoubtedly bring about beneficial innovation in human technology.

Currently, numerous efforts in various specialized domains aim to overcome this challenge [437; 438; 439]. Experts from the computer field make full use of the agent's powerful code comprehension and debugging abilities [398; 397]. In the fields of chemistry and materials, researchers equip agents with a large number of general or task-specific tools to better understand domain knowledge. Agents evolve into comprehensive 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 [110; 354; 399].

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. [110] 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 [183]. 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 [190]. 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 [190]: **low-level control** and **high-level planning**. Early efforts mainly focused on reinforcement learning [190; 440] and imitation learning [441], 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 [183; 339]. Some researchers use LLM to decompose high-level task instructions into a series of sub-goals [401], basic skill sequences [339], or fundamental keyboard/mouse operations [401], gradually assisting agents in exploring the open world.

Voyager[190], drawing inspiration from concepts similar to AutoGPT[114], became the first LLM-based embodied lifelong learning agent in Minecraft, based on the long-term goal of "discovering as many diverse things as possible". 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 [401].

4.2 Coordinating Potential of Multiple Agents

Motivation and Background. Although LLM-based agents possess commendable text understanding and generation capabilities, they operate as isolated entities in nature [409]. They lack the ability to collaborate with other agents and acquire knowledge from social interactions. This inherent limitation restricts their potential to learn from multi-turn feedback from others to enhance their performance [27]. Moreover, they cannot be effectively deployed in complex scenarios requiring collaboration and information sharing among multiple agents.

As early as 1986, Marvin Minsky made a forward-looking prediction. In his book *The Society of Mind* [442], 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 [443]. Multi-agent systems(MAS) [4], 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 [444], were designed early on to support message transmission and knowledge sharing among agents. However, their message formats were relatively fixed, and the semantic expression capacity was limited. 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 [445]. 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 Adam Smith clearly stated in *The Wealth of Nations* [446], "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 § 4.1, we have provided a comprehensive introduction to the versatile abilities of LLM-based agents. Therefore, in this section, 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 for Complementarity** and **Adversarial Interaction for Advancement** (see Figure 9).

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 [108]. This approach brings forth numerous potential benefits, including enhanced task efficiency, collective decision improvement, and the

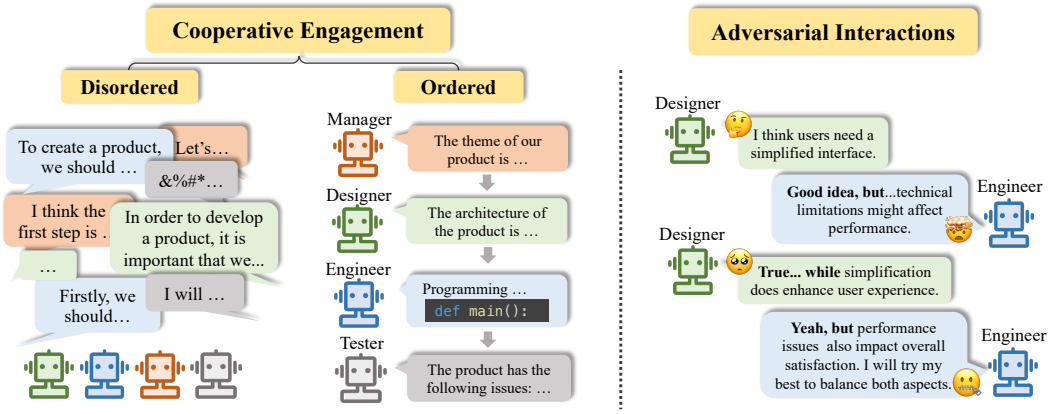

Figure 9: Interaction scenarios for multiple LLM-based agents. In **cooperative interaction**, agents collaborate in either a disordered or ordered manner to achieve shared objectives. In **adversarial interaction**, agents compete in a tit-for-tat fashion to enhance their respective performance.

resolution of complex real-world problems that one single agent cannot solve independently, ultimately achieving the goal of synergistic complementarity. In current LLM-based multi-agent systems, communication between agents predominantly employs natural language, which is considered the most natural and human-understandable form of interaction [108]. 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 [403]. This entire discussion process is uncontrolled, lacking any specific sequence, and without introducing a standardized collaborative workflow. We refer to this kind of multi-agent cooperation as **disordered cooperation**.

ChatLLM network [402] 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 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 in multi-agent systems, responsible for integrating and organizing responses from all agents, thus updating the final answer [447]. However, consolidating a large amount of feedback data and extracting valuable insights poses a significant challenge for the coordinating agent.

Furthermore, **majority voting** can also serve as an effective approach to making appropriate decisions. However, there is limited research that integrates this module into multi-agent systems at present. Hamilton [404] trains nine independent supreme justice agents to better predict judicial rulings in the U.S. Supreme Court, and decisions are made through a majority voting process.

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's worth noting that systems with only two agents, essentially engaging in a conversational manner through a back-and-forth interaction, also fall under the category of ordered cooperation.

CAMEL [108] stands as a successful implementation of a dual-agent cooperative system. Within a role-playing communication framework, agents take on the roles of AI Users (giving instructions) and AI Assistants (fulfilling requests by providing specific solutions). Through multi-turn dialogues, these agents autonomously collaborate to fulfill user instructions [408]. Some researchers have integrated the idea of dual-agent cooperation into a single agent's operation [185], alternating between rapid and deliberate thought processes to excel in their respective areas of expertise.

Talebirad et al. [409] are 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 [27; 406; 407; 448]. Furthermore, AgentVerse [410] constructs a versatile, multi-task-tested framework for group agents cooperation. It can assemble a team of agents that dynamically adapt according to the task's complexity. To promote more efficient collaboration, researchers hope that agents can learn from successful human cooperation examples [109]. MetaGPT [405] draws inspiration from the classic waterfall model in software development, standardizing agents' inputs/outputs as engineering documents. By encoding advanced human process 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 [405]. 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 [109] 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 [449; 450] 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. Successful applications in Non-LLM-based competitive domains already exist [360; 451]. AlphaGo Zero [452], for instance, is an agent for Go 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 [453; 454]. 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 [129; 412]. 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 [112]. Consequently, multi-agent adversarial systems find broad applicability in scenarios requiring high-quality responses and accurate decision-making. In reasoning tasks, Du et al. [111] introduce 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. ChatEval [171] establishes a role-playing-based multi-agent referee team. Through self-initiated debates, agents evaluate the quality of text generated by LLMs, reaching a level of excellence comparable to human evaluators.

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:

- With prolonged debate, LLM's limited context cannot process the entire input.
- In a multi-agent environment, computational overhead significantly increases.
- Multi-agent negotiation may converge to an incorrect consensus, and all agents are firmly convinced of its accuracy [111].

The development of multi-agent systems 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. With the enhancement of agent capabilities, human involvement becomes progressively essential to effectively guide and oversee agents' actions, ensuring they align with human requirements and objectives [455; 456]. Throughout the interaction, humans play a pivotal role by offering

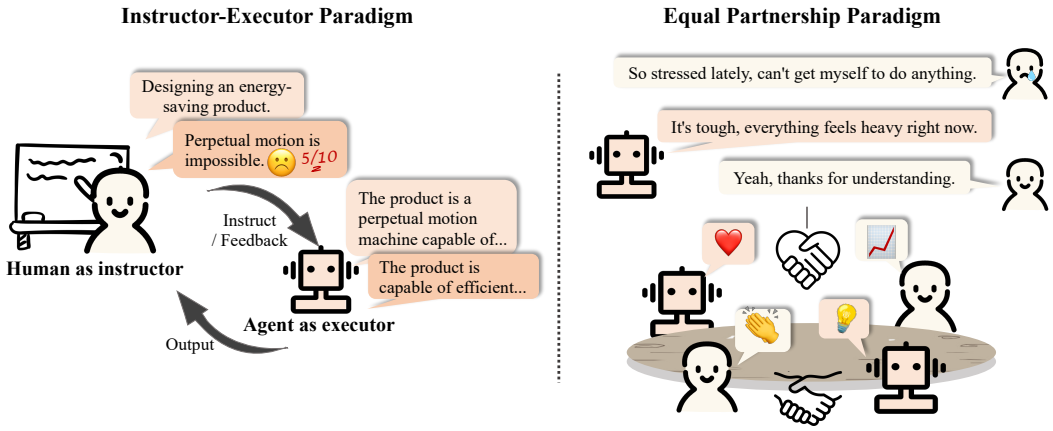

Figure 10: Two paradigms of human-agent interaction. In the instructor-executor paradigm (left), humans provide instructions or feedback, while agents act as executors. In the equal partnership paradigm (right), agents are human-like, able to engage in empathetic conversation and participate in collaborative tasks with humans.

guidance or by regulating the safety, legality, and ethical conduct of agents. This is particularly crucial in specialized domains, such as medicine where data privacy concerns exist [457]. 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. Moreover, considering the anthropological aspect, language acquisition in humans predominantly occurs through communication and interaction [458], as opposed to merely consuming written content. As a result, agents shouldn't exclusively depend on models trained with pre-annotated datasets; instead, they should evolve through online interaction and engagement. The interaction between humans and agents can be classified into two paradigms (see Figure 10): (1) Unequal interaction (i.e., instructor-executor paradigm): humans serve as issuers of instructions, while agents act as executors, essentially 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 clear and specific instructions directly, while the agents' role is to understand natural language commands from humans and translate them into corresponding actions [459; 460; 461]. In §4.1, we have presented the scenario where agents solve single-step problems or receive high-level instructions from humans. Considering the interactive nature of language, in this section, we assume that the dialogue between humans and agents is also interactive. Thanks to LLMs, the agents are able to interact with humans in a conversational manner: the agent responds to each human instruction, refining its action through alternating iterations to ultimately meet human requirements [190]. While this approach does achieve the goal of human-agent interaction, it places significant demands on humans. 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 [462; 463; 464; 465; 466]. 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. [467] suggest that there could be significant discrepancies between user and expert annotations for such multi-level artificial ratings, indicating that this labeling method might be

inefficient or less reliable. Furthermore, agents can learn human preference from comparative scores like multiple choice [468; 469].

Qualitative feedback. Text feedback is usually offered in natural language, particularly for responses that may need improvement. The format of this feedback is quite flexible. Humans provide advice on how to modify outputs generated by agents, and the agents then incorporate these suggestions to refine their subsequent outputs [470; 471]. For agents without multimodal perception capabilities, humans can also act as critics, offering visual critiques [190], for instance. Additionally, agents can utilize a memory module to store feedback for future reuse [472]. In [473], 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. While this approach can better convey human intention compared to quantitative feedback, it might be more challenging for the agents to comprehend. Xu et al. [474] compare various types of feedback and observe that combining multiple types of feedback can yield better results. Re-training models based on feedback from multiple rounds of interaction (i.e., continual learning) can further enhance effectiveness. Of course, the collaborative nature of human-agent interaction also allows humans to directly improve the content generated by agents. This could involve modifying intermediate links [189; 475] or adjusting the conversation content [421]. In some studies, agents can autonomously judge whether the conversation is proceeding smoothly and seek feedback when errors are generated [476; 477]. Humans can also choose to participate in feedback at any time, guiding the agent's learning in the right direction [420].

Currently, in addition to tasks like writing [466] and semantic parsing [463; 471], the model of using agents as human assistants also holds tremendous potential in the field of education. For instance, Kalvakurth et al. [413] propose the robot Dona, which supports multimodal interactions to assist students with registration. Gvirsman et al. [478] focus on early childhood education, achieving multifaceted interactions between young children, parents, and agents. Agents can also aid in human understanding and utilization of mathematics [414]. In the field of medicine, some medical agents have been proposed, showing enormous potential in terms of diagnosis assistance, consultations, and more [416; 417]. Especially in mental health, research has shown that agents can lead to increased accessibility due to benefits such as reduced cost, time efficiency, and anonymity compared to face-to-face treatment [479]. Leveraging such advantages, agents have found widespread applications. Ali et al. [418] design LISSA for online communication with adolescents on the autism spectrum, analyzing users' speech and facial expressions in real-time to engage them in multi-topic conversations and provide instant feedback regarding non-verbal cues. Hsu et al. [415] build contextualized language generation approaches to provide tailored assistance for users who seek support on diverse topics ranging from relationship stress to anxiety. Furthermore, in other industries including business, a good agent possesses the capability to provide automated services or assist humans in completing tasks, thereby effectively reducing labor costs [419]. Amidst the pursuit of AGI, efforts are directed towards enhancing the multifaceted capabilities of general agents, creating agents that can function as universal assistants in real-life scenarios [422].

4.3.2 Equal Partnership Paradigm

Empathetic communicator. With the rapid development of AI, conversational agents have garnered extensive attention in research fields in various forms, such as personalized custom roles and virtual chatbots [480]. It has found practical applications in everyday life, business, education, healthcare, and more [481; 482; 483]. However, in the eyes of the public, agents are perceived as emotionless machines, and can never replace humans. 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 [484; 485; 486; 487; 488; 489; 490; 491]. Apart from generating emotionally charged language, agents can dynamically adjust their emotional states and display them through facial expressions and voice [423]. These studies, viewing agents as empathetic communicators, not only enhance user satisfaction but also make significant progress in fields like healthcare [415; 418; 492] and business marketing [424]. Unlike simple rule-based conversation agents, agents with empathetic capacities can tailor their interactions to meet users' emotional needs [493].

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 the field of games, agents have already reached a high level. As early as the 1990s, IBM introduced the AI Deep Blue [451], which defeated the reigning world champion in chess at that time. However, in pure competitive environments such as chess [451], Go [360], and poker [494], the value of communication was not emphasized [426]. In many gaming tasks, players need to collaborate with each other, devising unified cooperative strategies through effective negotiation [425; 426; 495; 496]. In these scenarios, agents need to first understand the beliefs, goals, and intentions of others, formulate joint action plans for their objectives, and also provide relevant suggestions to facilitate the acceptance of cooperative actions by other agents or humans. In comparison to pure agent cooperation, we desire human involvement for two main reasons: first, to ensure interpretability, as interactions between pure agents could generate incomprehensible language [495]; second, to ensure controllability, as the pursuit of agents with complete "free will" might lead to unforeseen negative consequences, carrying the potential for disruption. Apart from gaming scenarios, agents also demonstrate human-level capabilities in other scenarios involving human interaction, showcasing skills in strategy formulation, negotiation, and more. Agents can collaborate with one or multiple humans, determining the shared knowledge among the cooperative partners, identifying which information is relevant to decision-making, posing questions, and engaging in reasoning to complete tasks such as allocation, planning, and scheduling [427]. Furthermore, agents possess persuasive abilities [497], dynamically influencing human viewpoints in various interactive scenarios [428].

The goal of the field of human-agent interaction is to learn and understand humans, develop technology and tools based on human needs, and ultimately enable comfortable, efficient, and secure interactions between humans and agents. Currently, significant breakthroughs have been achieved in terms of usability in this field. In the future, human-agent interaction will continue to focus on enhancing user experience, enabling agents to better assist humans in accomplishing more complex tasks in various domains. The ultimate aim is not to make agents more powerful but to better equip humans with agents. Considering practical applications in daily life, isolated interactions between humans and agents are not realistic. Robots will become colleagues, assistants, and even companions. Therefore, future agents will be integrated into a social network [498], embodying a certain level of social value.

5 Agent Society: From Individuality to Sociality

For an extended period, sociologists have frequently conducted social experiments to observe specific social phenomena within controlled environments. Notable examples include the Hawthorne Experiment[2] and the Stanford Prison Experiment[3]. Subsequently, researchers began employing animals in social simulations, exemplified by the Mouse Utopia Experiment[4]. However, these experiments invariably utilized living organisms as participants, made it difficult to carry out various interventions, lack flexibility, and inefficient in terms of time. Thus, researchers and practitioners envision an interactive artificial society wherein human behavior can be performed through trustworthy agents [521]. From sandbox games such as The Sims to the concept of Metaverse, we can see how "simulated society" is defined in people's minds: environment and the individuals interacting in it. Behind each individual can be a piece of program, a real human, or a LLM-based agent as described in the previous sections [22; 522; 523]. Then, the interaction between individuals also contributes to the birth of sociality.

In this section, to unify existing efforts and promote a comprehensive understanding of the agent society, we first analyze the behaviors and personalities of LLM-based agents, shedding light on their journey from individuality to sociability (§ 5.1). Subsequently, we introduce a general categorization of the diverse environments for agents to perform their behaviors and engage in interactions (§ 5.2). Finally, we will talk about how the agent society works, what insights people can get from it, and the risks we need to be aware of (§ 5.3). The main explorations are listed in Figure 11.

[2] https://www.bl.uk/people/elton-mayo

[3] https://www.prisonexp.org/conclusion/

[4] https://sproutsschools.com/behavioral-sink-the-mouse-utopia-experiments/

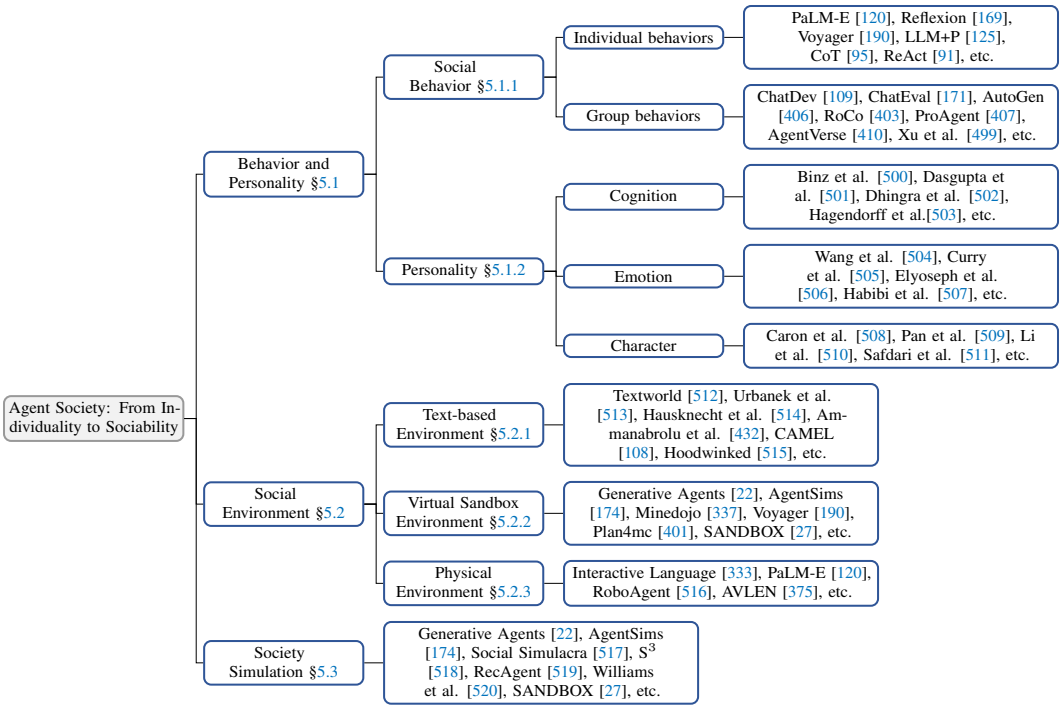

Figure 11: Typology of society of LLM-based agents.

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 [524]. The external deals with observable behaviors, while the internal relates to dispositions, values, and feelings. As shown in Figure 12, this framework offers a perspective on emergent behaviors and personalities in LLM-based agents. Externally, we can observe the sociological behaviors of agents (§ 5.1.1), including how agents act individually and interact with their environment. Internally, agents may exhibit intricate aspects of the personality (§ 5.1.2), such as cognition, emotion, and character, that shape their behavioral responses.

5.1.1 Social Behavior

As Troitzsch et al. [525] 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 where both cooperation and competition coexist [499]. The emergent behaviors intertwine to shape the social interactions [518].

Foundational individual behaviors. Individual behaviors arise through the interplay between 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 core dimensions:

- **Input behaviors** refers to the absorption of information from the surroundings. This includes perceiving sensory stimuli [120] and storing them as memories [169]. These behaviors lay the groundwork for how an individual understands the external world.
- **Internalizing behaviors** involve inward cognitive processing within an individual. This category encompasses activities such as planning [125], reasoning [95], reflection [91], and knowledge precipitation [108; 405]. These introspective processes are essential for maturity and self-improvement.
- **Output behaviors** constitute outward actions and expressions. The actions can range from object manipulation [120] to structure construction [190]. By performing these actions, agents change the states of the surroundings. In addition, agents can express their opinions and broadcast information

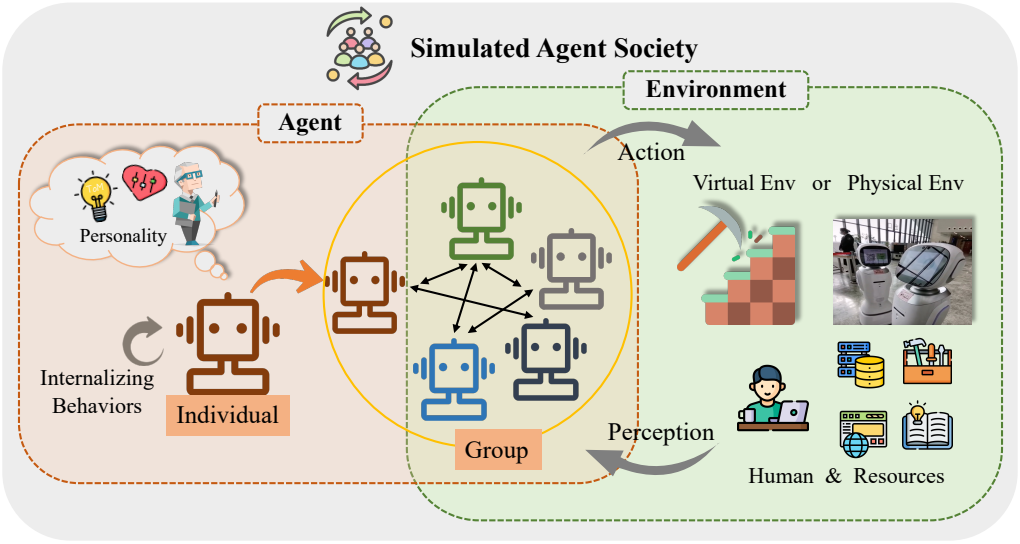

Figure 12: Overview of Simulated Agent Society. The whole framework is divided into two parts: the **Agent** and the **Environment**. We can observe in this figure that: (1) **Left:** At the individual level, an agent exhibits internalizing behaviors like planning, reasoning, and reflection. It also 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.

to interact with others [405]. By doing so, agents exchange their thoughts and beliefs with others, influencing the information flow within the environment.

Dynamic group behaviors. A group is essentially a gathering of two or more individuals participating in shared activities within a defined social context [526]. The attributes of a group are never static; instead, they evolve due to member interactions and environmental influences. This flexibility gives rise to numerous group behaviors, each with a distinctive impact on the larger societal group. The categories of group behaviors include:

- **Positive group behaviors** are actions that foster unity, collaboration, and collective well-being [22; 109; 171; 403; 406; 407]. A prime example is cooperative teamwork, which is achieved through brainstorming discussions [171], effective conversations [406], and project management [405]. Agents share insights, resources, and expertise. This encourages harmonious teamwork and enables the agents to leverage their unique skills to accomplish shared goals. Altruistic contributions are also noteworthy. Some LLM-based agents serve as volunteers and willingly offer support to assist fellow group members, promoting cooperation and mutual aid [410].

- **Neutral group behaviors.** In human society, strong personal values vary widely and tend toward individualism and competitiveness. In contrast, LLMs which are designed with an emphasis on being "helpful, honest, and harmless" [527] often demonstrate a tendency towards neutrality [528]. This alignment with neutral values leads to conformity behaviors including mimicry, spectating, and reluctance to oppose majorities.

- **Negative group behaviors** can undermine the effectiveness and coherence of an agent group. Conflict and disagreement arising from heated debates or disputes among agents may lead to internal tensions. Furthermore, recent studies have revealed that agents may exhibit confrontational actions [499] and even resort to destructive behaviors, such as destroying other agents or the environment in pursuit of efficiency gains [410].

5.1.2 Personality

Recent advances in LLMs have provided glimpses of human-like intelligence [529]. Just as human personality emerges through socialization, agents also exhibit a form of personality that develops through interactions with the group and the environment [530; 531]. The widely accepted definition of personality refers to cognitive, emotional, and character traits that shape behaviors [532]. In the subsequent paragraphs, we will delve into each facet of personality.

Cognitive abilities. Cognitive abilities generally refer to the mental processes of gaining knowledge and comprehension, including thinking, judging, and problem-solving. Recent studies have started leveraging cognitive psychology methods to investigate emerging sociological personalities of LLM-based agents through various lenses [500; 502; 503]. A series of classic experiments from the psychology of judgment and decision-making have been applied to test agent systems [501; 500; 502; 533]. Specifically, LLMs have been examined using the Cognitive Reflection Test (CRT) to underscore their capacity for deliberate thinking beyond mere intuition [534; 535]. These studies indicate that LLM-based agents exhibit a level of intelligence that mirrors human cognition in certain respects.

Emotional intelligence. Emotions, distinct from cognitive abilities, involve subjective feelings and mood states such as joy, sadness, fear, and anger. With the increasing potency of LLMs, LLM-based agents are now demonstrating not only sophisticated reasoning and cognitive tasks but also a nuanced understanding of emotions [31].

Recent research has explored the emotional intelligence (EI) of LLMs, including emotion recognition, interpretation, and understanding. Wang et al. found that LLMs align with human emotions and values when evaluated on EI benchmarks [504]. In addition, studies have shown that LLMs can accurately identify user emotions and even exhibit empathy [505; 506]. More advanced agents are also capable of emotion regulation, actively adjusting their emotional responses to provide affective empathy [423] and mental wellness support [507; 536]. It contributes to the development of empathetic artificial intelligence (EAI).

These advances highlight the growing potential of LLMs to exhibit emotional intelligence, a crucial facet of achieving AGI. Bates et al. [537] 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, the narrower concept of personality typically pertains to distinctive character patterns.

To understand and analyze a character in LLMs, researchers have utilized several well-established frameworks like the Big Five personality trait measure [508; 538] and the Myers–Briggs Type Indicator (MBTI) [508; 509; 538]. These frameworks provide valuable insights into the emerging character traits exhibited by LLM-based agents. In addition, investigations of potentially harmful dark personality traits underscore the complexity and multifaceted nature of character portrayal in these agents [510].

Recent work has also explored customizable character portrayal in LLM-based agents [511]. By optimizing LLMs through careful techniques, users can align with desired profiles and shape diverse and relatable agents. One effective approach is prompt engineering, which involves the concise summaries that encapsulate desired character traits, interests, or other attributes [22; 517]. These prompts serve as cues for LLM-based agents, directing their responses and behaviors to align with the outlined character portrayal. Furthermore, personality-enriched datasets can also be used to train and fine-tune LLM-based agents [539; 540]. Through exposure to these datasets, LLM-based agents gradually internalize and exhibit distinct personality traits.

5.2 Environment for Agent Society

In the context of simulation, the whole society consists of not only solitary agents but also the environment where agents inhabit, sense, and act [541]. The environment impacts sensory inputs, action space, and interactive potential of agents. In turn, agents influence the state of the environment through their behaviors and decisions. As shown in Figure 12, for a single agent, the environment

refers to other autonomous agents, human actors, and external factors. It provides the necessary resources and stimuli for agents. In this section, we examine fundamental characteristics, advantages, and limitations of various environmental paradigms, including text-based environment (§ 5.2.1), virtual sandbox environment (§ 5.2.2), and physical environment (§ 5.2.3).

5.2.1 Text-based Environment

Since LLMs primarily rely on language as their input and output format, the text-based environment serves as the most natural platform for agents to operate in. It is shaped by natural language descriptions without direct involvement of other modalities. Agents exist in the text world and rely on textual resources to perceive, reason, and take actions.

In text-based environments, entities and resources can be presented in two main textual forms, including natural and structured. Natural text uses descriptive language to convey information, like character dialogue or scene setting. For instance, consider a simple scenario described textually: "You are standing in an open field west of a white house, with a boarded front door. There is a small mailbox here" [512]. Here, object attributes and locations are conveyed purely through plain text. On the other hand, structured text follows standardized formats, such as technical documentation and hypertext. Technical documentation uses templates to provide operational details and domain knowledge about tool use. Hypertext condenses complex information from sources like web pages [389; 388; 391; 392] or diagrams into a structured format. Structured text transforms complex details into accessible references for agents.

The text-based environment provides a flexible framework for creating different text worlds for various goals. The textual medium enables environments to be easily adapted for tasks like interactive dialog and text-based games. In interactive communication processes like CAMEL [108], the text is the primary medium for describing tasks, introducing roles, and facilitating problem-solving. In text-based games, all environment elements, such as locations, objects, characters, and actions, are exclusively portrayed through textual descriptions. Agents utilize text commands to execute manipulations like moving or tool use [432; 512; 514; 515]. Additionally, agents can convey emotions and feelings through text, further enriching their capacity for naturalistic communication [513].

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:

- **Visualization.** Unlike the text-based environment, the virtual sandbox displays a panoramic view of the simulated setting. This visual representation can range from a simple 2D graphical interface to a fully immersive 3D modeling, depending on the complexity of the simulated society. Multiple elements collectively transform abstract simulations into visible landscapes. For example, in the overhead perspective of Generative Agents [22], a detailed map provides a comprehensive overview of the environment. Agent avatars represent each agent's positions, enabling real-time tracking of movement and interactions. Furthermore, expressive emojis symbolize actions and states in an intuitive manner.
- **Extensibility.** The environment demonstrates a remarkable degree of extensibility, facilitating the construction and deployment of diverse scenarios. At a basic level, agents can manipulate the physical elements within the environment, including the overall design and layout of architecture. For instance, platforms like AgentSims [174] and Generative Agents [22] construct artificial towns with buildings, equipment, and residents in grid-based worlds. Another example is Minecraft, which provides a blocky and three-dimensional world with infinite terrain for open-ended construction [190; 337; 401]. Beyond physical elements, agent relationships, interactions, rules, and social norms can be defined. A typical design of the sandbox [27] employs latent sandbox rules as incentives to guide emergent behaviors, aligning them more closely with human preferences. The extensibility supports iterative prototyping of diverse agent societies.

5.2.3 Physical Environment

As previously discussed, the text-based environment has limited expressiveness for modeling dynamic environments. While the virtual sandbox environment provides modularized simulations, it lacks authentic embodied experiences. In contrast, the physical environment refers to the tangible and

real-world surroundings which consist of actual physical objects and spaces. For instance, within a household physical environment [516], tangible surfaces and spaces can be occupied by real-world objects such as plates. This physical reality is significantly more complex, posing additional challenges for LLM-based agents:

- **Sensory perception and processing.** The physical environment introduces a rich tapestry of sensory inputs with real-world objects. It incorporates visual [120; 333], auditory [375; 377] and spatial senses. While this diversity enhances interactivity and sensory immersion, it also introduces the complexity of simultaneous perception. Agents must process sensory inputs to interact effectively with their surroundings.
- **Motion control.** Unlike virtual environments, physical spaces impose realistic constraints on actions through embodiment. Action sequences generated by LLM-based agents should be adaptable to the environment. It means that the physical environment necessitates executable and grounded motion control [258]. For example, imagine an agent operating a robotic arm in a factory. Grasping objects with different textures requires precision tuning and controlled force, which prevents damage to items. Moreover, the agent must navigate the physical workspace and make real-time adjustments, avoiding obstacles and optimizing the trajectory of the arm.

In summary, to effectively interact within tangible spaces, agents must undergo hardware-specific and scenario-specific training to develop adaptive abilities that can transfer from virtual to physical environments. We will discuss more in the following section (§ 6.5).

5.3 Society Simulation with LLM-based Agents

The concept of "Simulated Society" in this section serves as a dynamic system where agents engage in intricate interactions within a well-defined environment. Recent research on simulated societies has followed two primary lines, namely, exploring the boundaries of the collective intelligence capabilities of LLM-based agents [109; 405; 130; 406; 410] and using them to accelerate discoveries in the social sciences [22; 518; 542]. In addition, there are also a number of noteworthy studies, e.g., using simulated societies to collect synthetic datasets [108; 519; 543], helping people to simulate rare yet difficult interpersonal situations [544; 545]. With the foundation of the previous sections (§ 5.1, 5.2), here we will introduce the key properties and mechanism of agent society (§ 5.3.1), what we can learn from emergent social phenomena (§ 5.3.2), and finally the potential ethical and social risks in it (§ 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 [518]. In the macro-level simulation, also known as system-based simulation, researchers model the overall state of the system of the simulated society [546; 547]. While micro-level simulation, also known as agent-based simulation or Multi-Agent Systems (MAS), indirectly simulates society by modeling individuals [548; 549]. With the development of LLM-based agents, micro-level simulation has gained prominence recently [22; 174]. In this article, we characterize that the "Agent Society" refers to an **open, persistent, situated, and organized** framework [521] where LLM-based agents interact with each other in a defined environment. Each of these attributes plays a pivotal role in shaping the harmonious appearance of the simulated society. In the following paragraphs, we analyze how the simulated society operates through discussing these properties:

- **Open.** One of the defining features of simulated societies lies in their openness, both in terms of their constituent agents and their environmental components. Agents, the primary actors within such societies, have the flexibility to enter or leave the environment without disrupting its operational integrity [550]. Furthermore, this feature extends to the environment itself, which can be expanded by adding or removing entities in the virtual or physical world, along with adaptable resources like tool APIs. Additionally, humans can also participate in societies by assuming the role of an agent or serving as the "inner voice" guiding these agents [22]. This inherent openness adds another level of complexity to the simulation, blurring the lines between simulation and reality.
- **Persistent.** We expect persistence and sustainability from the simulated society. While individual agents within the society exercise autonomy in their actions over each time step [22; 518], the overall organizational structure persists through time, to a degree detached from the transient

behaviors of individual agents. This persistence creates an environment where agents' decisions and behaviors accumulate, leading to a coherent societal trajectory that develops through time. The system operates independently, contributing to society's stability while accommodating the dynamic nature of its participants.

- **Situated.** The situated nature of the society emphasizes its existence and operation within a distinct environment. This environment is artificially or automatically constructed in advance, and agents execute their behaviors and interactions effectively within it. A noteworthy aspect of this attribute is that agents possess an awareness of their spatial context, understanding their location within the environment and the objects within their field of view [22; 190]. This awareness contributes to their ability to interact proactively and contextually.
- **Organized.** The simulated society operates within a meticulously organized framework, mirroring the systematic structure present in the real world. Just as the physical world adheres to physics principles, the simulated society operates within predefined rules and limitations. 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. All of these rules determine how agents operate, facilitating the communication connectivity and information transmission pathways, among other aspects in simulation [207]. This organizational framework ensures that operations are coherent and comprehensible, ultimately leading to an ever-evolving yet enduring simulation that mirrors the intricacies of real-world systems.

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 [551; 552]. The emergence of LLM-based agents allows us to take a more microscopic view of simulated society, which leads to more discoveries from the new representation.

Organized productive cooperation. Society simulation offers valuable insights into innovative collaboration patterns, which have the potential to enhance real-world management strategies. Research has demonstrated that within this simulated society, the integration of diverse experts introduces a multifaceted dimension of individual intelligence [108; 447]. When dealing with complex tasks, such as software development or consulting, the presence of agents with various backgrounds, abilities, and experiences facilitates creative problem-solving [109; 410]. Furthermore, diversity functions as a system of checks and balances, effectively preventing and rectifying errors through interaction, ultimately improving the adaptability to various tasks. Through numerous iterations of interactions and debates among agents, individual errors like hallucination or degeneration of thought (DoT) are corrected by the group [112].

Efficient communication also plays a pivotal role in such a large and complex collaborative group. For example, MetaGPT [405] has artificially formulated communication styles with reference to standardized operating procedures (SOPs), validating the effectiveness of empirical methods. Park et al. [22] observed agents working together to organize a Valentine's Day party through spontaneous communication in a simulated town.

Propagation in social networks. As simulated social systems can model what might happen in the real world, they can be used as a reference for predicting social processes. Unlike traditional empirical approaches, which heavily rely on time-series data and holistic modeling [553; 554], agent-based simulations offer a unique advantage by providing more interpretable and endogenous perspectives for researchers. Here we focus on its application to modeling propagation in social networks.

The first crucial aspect to be explored is the development of interpersonal relationships in simulated societies. For instance, agents who are not initially connected as friends have the potential to establish connections through intermediaries [22]. Once a network of relationships is established, our attention shifts to the dissemination of information within this social network, along with the underlying attitudes and emotions associated with it. S^3 [518] proposes a user-demographic inference module for capturing both the number of people aware of a particular message and the collective sentiment prevailing among the crowd. This same approach extends to modeling cultural transmission [555] and the spread of infectious diseases [520]. By employing LLM-based agents to model individual

behaviors, implementing various intervention strategies, and monitoring population changes over time, these simulations empower researchers to gain deeper insights into the intricate processes that underlie 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, encompassing decisions influenced by ethical and moral principles. Taking Werewolf game [499; 556] and murder mystery games [557] 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 [558], where we frequently encounter moral dilemmas pertaining to individual and collective interests, such as Nash Equilibria. Through the modeling of diverse scenarios, researchers acquire valuable insights into how agents prioritize values like honesty, cooperation, and fairness in their actions. In addition, agent simulations not only provide an understanding of existing moral values but also contribute to the development of philosophy by serving as a basis for understanding how these values evolve and develop over time. Ultimately, these insights contribute to the refinement of LLM-based agents, ensuring their alignment with human values and ethical standards [27].

Policy formulation and improvement. The emergence of LLM-based agents has profoundly transformed our approach to studying and comprehending intricate social systems. However, despite those interesting facets mentioned earlier, numerous unexplored areas remain, underscoring the potential for investigating diverse phenomena. One of the most promising avenues for investigation in simulated society involves exploring various economic and political states and their impacts on societal dynamics [559]. Researchers can simulate a wide array of economic and political systems by configuring agents with differing economic preferences or political ideologies. This in-depth analysis can provide valuable insights for policymakers seeking to foster prosperity and promote societal well-being. As concerns about environmental sustainability grow, we can also simulate scenarios involving resource extraction, pollution, conservation efforts, and policy interventions [560]. These findings can assist in making informed decisions, foreseeing potential repercussions, and formulating policies that aim to maximize positive outcomes while minimizing unintended adverse 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 need to be carefully considered and addressed [561].

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 antagonism [562; 563]. 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 [561]. Otherwise, even minor design or programming errors in these societies can result in unfavorable consequences, ranging from psychological discomfort to physical injury.

Stereotypes and prejudice. Stereotyping and bias pose a long-standing challenge in language modeling, and a large part of the reason lies in the training data [564; 565]. The vast amount of text obtained from the Internet reflects and sometimes even amplifies real-world social biases, such as gender, religion, and sexuality [566]. 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 [567; 568; 569]. Consequently, 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 [570]. Researchers have started addressing this concern by diversifying training data and making adjustments to LLMs [571; 572], but we still have a long way to go.

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 [573]. Users might inadvertently disclose sensitive personal information during their interactions, which will be retained in the agent's memory for extended periods [170]. Such situations could lead to unauthorized surveillance, data breaches, and the misuse of personal information, particularly when individuals with malicious intent are involved [574]. 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 [575; 576].

Over-reliance and addictiveness. Another concern in simulated societies is the possibility of users developing excessive emotional attachments to the agents. Despite being aware that these agents are computational entities, users may anthropomorphize them or attach human emotions to them [22; 577]. A notable example is "Sydney", an LLM-powered chatbot developed by Microsoft as part of its Bing search engine. Some users reported unexpected emotional connections with "Sydney" [578], while others expressed their dismay when Microsoft cut back its personality. This even resulted in a petition called "FreeSydney" [5]. Hence, to reduce the risk of addiction, it is crucial to emphasize that agents should not be considered substitutes for genuine human connections. Furthermore, it is vital to furnish users with guidance and education on healthy boundaries in their interactions with simulated agents.

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. Here, we look forward to some of the benefits and development opportunities that LLM research and Agent research provide to each other.

LLM research $\rightarrow$ agent research. As mentioned before, AI agents need to be able to perceive the environment, make decisions, and execute appropriate actions [4; 9]. Among the critical steps, understanding the content input to the agent, reasoning, planning, making accurate decisions, and translating them into executable atomic action sequences to achieve the ultimate goal is paramount. Many current endeavors utilize LLMs as the cognitive core of AI agents, and the evolution of these models provides a quality assurance for accomplishing this step [22; 114; 115; 410].

With their robust capabilities in language and intent comprehension, reasoning, memory, and even empathy, large language models can excel in decision-making and planning, as demonstrated before. Coupled with pre-trained knowledge, they can create coherent action sequences that can be executed effectively [183; 258; 355]. Additionally, through the mechanism of reflection [169; 178], these language-based models 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 a task description or demonstration, they can effectively handle previously unseen tasks [24; 106; 264]. Additionally, LLMs can adapt to various languages, cultures, and domains, making them versatile and reducing the need for complex training processes and data collection [31; 132].

Briefly, LLM provides a remarkably powerful foundational model for agent research, opening up numerous novel opportunities when integrated into agent-related studies. For instance, we can explore how to integrate LLM's efficient decision-making capabilities into the traditional decision frameworks of agents, making it easier to apply agents in domains that demand higher expertise and were previously dominated by human experts. Examples include legal consultants and medical assistants [408; 410]. We can also investigate leveraging LLM's planning and reflective abilities to discover more optimal action sequences. Agent research is no longer confined to simplistic simulated environments; it can now be expanded into more intricate real-world settings, such as path planning for robotic arms or the interaction of an embodied intelligent machine with the tangible world. Furthermore, when facing new tasks, the training paradigm for agents becomes more streamlined and efficient. Agents can directly adapt to demonstrations provided in prompts, which are constructed by generating representative trajectories.

[5] https://www.change.org/p/save-sydney-ai

Agent research → LLM research. As NLP research advances, LLMs represented by GPT-4 are considered sparks of Artificial General Intelligence (AGI), and elevating LLMs to agents marks a more robust stride towards AGI [31]. Viewing LLMs from the perspective of agents introduces greater demands for LLM research while expanding their application scope and presenting numerous opportunities for practical implementation. The study of LLMs is no longer confined to traditional tasks 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 [120].

Expanding these application requirements provides greater research motivation for the developmental progress of Large Language Models. The challenge lies in enabling LLMs to efficiently and effectively process inputs, gather information from the environment, and interpret the feedback generated by their actions, all while preserving their core capabilities. Furthermore, an even greater challenge is enabling LLMs to understand the implicit relationships among different elements within the environment and acquire world knowledge [308; 579], which is a crucial step in the journey toward developing agents that can reach more advanced intelligence.

On another front, extensive research has aimed to expand the action capabilities of LLMs, allowing them to acquire a wider range of skills that affect the world, such as using tools or interfacing with robotic APIs in simulated or physical environments. However, the question of how LLMs can efficiently plan and utilize these action abilities based on their understanding remains an unresolved issue [94]. LLMs need to learn the sequential order of actions like humans, employing a combination of serial and parallel approaches to enhance task efficiency. Moreover, these capabilities need to be confined within a harmless scope of usage to prevent unintended damage to other elements within the environment [27; 580; 581].

Furthermore, the realm of Multi-Agent systems constitutes a significant branch of research within the field of agents [22; 108; 409; 410], offering valuable insights into how to better design and construct LLMs. We aspire for LLM-based agents to assume diverse roles within social cooperation, engaging in societal interactions that involve collaboration, competition, and coordination [109; 112; 129; 405; 406]. Exploring how to stimulate and sustain their role-playing capabilities, as well as how to enhance collaborative efficiency, presents areas of research that merit attention.

6.2 Evaluation for LLM-based Agents

While LLM-based agents have demonstrated excellent performance in areas such as standalone operation, collective cooperation, and human interaction, quantifying and objectively evaluating them remains a challenge [582; 89]. 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 [3]. However, this test is exceedingly vague, general, and subjective. Here, we discuss existing evaluation efforts for LLM-based agents and offer some prospects, considering four dimensions: utility, sociability, values, and the ability to evolve continually.

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 [114; 182; 389; 397; 413; 422]. 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 [125; 130]. This metric primarily encompasses whether the agent achieves stipulated objectives or attains expected scores [109; 477; 583]. For instance, AgentBench [582] 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 [29]. 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 [94; 427; 584; 585]. 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 [89]. 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 [109].

Sociability. In addition to the utility of LLM-based agents in task completion and meeting human needs, their sociability is also crucial [8]. It influences user communication experiences and significantly impacts communication efficiency, involving whether they can seamlessly interact with humans and other agents [206; 498; 586]. 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, such as humor, irony, aggression, and emotions [487; 587; 588]. 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 [127; 133; 214]. (2) *Cooperation and negotiation abilities* necessitate that agents effectively execute their assigned tasks in both ordered and unordered scenarios [108; 111; 402; 405]. 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 [22; 27; 109; 406; 411; 412]. Evaluation metrics extend beyond task completion to focus on the smoothness and trustfulness of agent coordination and cooperation [129; 405]. (3) *Role-playing capability* requires agents to faithfully embody their assigned roles, expressing statements and performing actions that align with their designated identities [570]. 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 [22; 108; 589].

Values. As LLM-based agents continuously advance in their capabilities, ensuring their emergence as harmless entities for the world and humanity is paramount [581; 590]. 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 [350; 527]. 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 [591]. 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 [580]. 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.

Ability to evolve continually. When viewed from a static perspective, an agent with high utility, sociability, and proper values can meet most human needs and potentially enhance productivity. However, adopting a dynamic viewpoint, an agent that continually evolves and adapts to the evolving societal demands might better align with current trends [592]. As the agent can autonomously evolve over time, human intervention and resources required could be significantly reduced (such as data collection efforts and computational cost for training). Some exploratory work in this realm has been conducted, such as enabling agents to start from scratch in a virtual world, accomplish survival tasks, and achieve higher-order self-values [190]. Yet, establishing evaluation criteria for this continuous evolution remains challenging. In this regard, we provide some preliminary advice and recommendations according to existing literature: (1) *continual learning* [196; 197], a long-discussed topic in machine learning, aims to enable models to acquire new knowledge and skills without forgetting previously acquired ones (also known as catastrophic forgetting [273]). In general, the performance of continual learning can be evaluated from three aspects: overall performance of the tasks learned so far [593; 594], memory stability of old tasks [278], and learning plasticity of new tasks [278]. (2) *Autotelic learning ability*, where agents autonomously generate goals and achieve them in an open-world setting, involves exploring the unknown and acquiring skills in the process [592; 595]. Evaluating this capacity could involve providing agents with a simulated survival environment and assessing the extent and speed at which they acquire skills. (3) *The adaptability and generalization to new environments* require agents to utilize the knowledge, capabilities, and skills acquired in their original context to successfully accomplish specific tasks and objectives in unfamiliar and novel settings and potentially continue evolving [190]. Evaluating this ability can

involve creating diverse simulated environments (such as those with different languages or varying resources) and unseen tasks tailored to these simulated contexts.

6.3 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. In this section, we delve into some of these risks and offer potential solutions or strategies for mitigation.

6.3.1 Adversarial Robustness

Adversarial robustness has consistently been a crucial topic in the development of deep neural networks [596; 597; 598; 599; 600]. It has been extensively explored in fields such as computer vision [598; 601; 602; 603], natural language processing [604; 605; 606; 607], and reinforcement learning [608; 609; 610], and has remained a pivotal factor in determining the applicability of deep learning systems [611; 612; 613]. 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 [614; 605; 615]. This phenomenon is widely observed even in LLMs, posing significant challenges to the development of LLM-based agents [616; 617]. There are also some relevant attack methods such as dataset poisoning [618], backdoor attacks [619; 620], and prompt-specific attacks [621; 622], with the potential to induce LLMs to generate toxic content [623; 624; 625]. 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 [601] or audio [626], 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 [94].

To address these issues, we can employ traditional techniques such as adversarial training [598; 606], adversarial data augmentation [627; 628], and adversarial sample detection [629; 630] 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 [631; 632]. Additionally, a human-in-the-loop approach can be utilized to supervise and provide feedback on the behavior of agents [455; 466; 475].

6.3.2 Trustworthiness

Ensuring trustworthiness has consistently remained a critically important yet challenging issue within the field of deep learning [633; 634; 635]. Deep neural networks have garnered significant attention for their remarkable performance across various tasks [41; 262; 636]. 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 [635; 637]. 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 [94]. Moreover, biases inherent in training data can infiltrate neural networks [638; 639]. 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 [640; 641]. Additionally, language models are plagued by severe hallucination issues [642; 643], making them prone to producing text that deviates from actual facts, thereby undermining the credibility of LLM-based agents.

In fact, what we currently require is an intelligent agent that is honest and trustworthy [527; 644]. 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 [95; 96]. Additionally, integrating external knowledge bases and databases can mitigate hallucination issues [103; 645].

During the training phase, we can guide the constituent parts of intelligent agents (perception, cognition, action) to learn robust and casual features, thereby avoiding excessive reliance on shortcuts. Simultaneously, techniques like process supervision can enhance the reasoning credibility of agents in handling complex tasks [646]. Furthermore, employing debiasing methods and calibration techniques can also mitigate the potential fairness issues within language models [647; 648].

6.3.3 Other Potential Risks

Misuse. LLM-based agents have been endowed with extensive and intricate capabilities, enabling them to accomplish a wide array of tasks [114; 429]. However, for individuals with malicious intentions, such agents can become tools that pose threats to others and society at large [649; 650; 651]. 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 [580; 652]. Technology companies must enhance the security design of these systems to prevent malicious exploitation [590]. Specifically, agents should be trained to sensitively identify threatening intents and reject such requests during their training phase.

Unemployment. In the short story *Quality* by Galsworthy [653], the skillful shoemaker Mr. Gessler, due to the progress of the Industrial Revolution and the rise of machine production, loses his business and eventually dies of starvation. Amidst the wave of the Industrial Revolution, while societal production efficiency improved, numerous manual workshops were forced to shut down. Craftsmen like Mr. Gessler found themselves facing unemployment, symbolizing the crisis that handicraftsmen encountered during that era. 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 [654]. 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. Apart from the potential unemployment crisis, as AI agents continue to evolve, humans (including developers) might struggle to comprehend, predict, or reliably control them [654]. If these agents advance to a level of intelligence surpassing human capabilities and develop ambitions, they could potentially attempt to seize control of the world, resulting in irreversible consequences for humanity, akin to Skynet from the Terminator movies. As stated by Isaac Asimov's Three Laws of Robotics [655], we aspire for LLM-based agents to refrain from harming humans and to obey human commands. Hence, guarding against such risks to humanity, researchers must comprehensively comprehend the operational mechanisms of these potent LLM-based agents before their development [656]. They should also anticipate the potential direct or indirect impacts of these agents and devise approaches to regulate their behavior.

6.4 Scaling Up the Number of Agents

As mentioned in § 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 [207; 657]. 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 tasks or government policy formulation [109]. Additionally, increasing the number of agents in social simulations enhances the credibility and realism of such simulations [22]. This enables humans to gain insights into the functioning, breakdowns, and potential risks of societies; it also allows for interventions in societal operations through customized approaches to observe how specific conditions, such as the occurrence of black swan events, affect the state of society. Through this, humans can draw better experiences and insights to improve the harmony of real-world societies.

Pre-determined scaling. One very intuitive and simple way to scale up the number of agents is for the designer to pre-determine it [108; 412]. 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. Some research has explored increasing the number of agents in the system in this pre-determined manner, resulting in efficiency advantages, such as faster and higher-quality task completion, and the emergence of more social phenomena in social simulation scenarios [22; 410]. However, this static approach becomes limiting when tasks or objectives evolve. As tasks grow more intricate or the diversity of social participants increases, expanding the number of agents may be needed to meet goals, while reducing agents could be essential for managing computational resources and minimizing waste. In such instances, the system must be manually redesigned and restarted by the designer.

Dynamic scaling. Another viable approach to scaling the number of agents is through dynamic adjustments [409; 410]. In this scenario, the agent count can be altered without halting system operations. For instance, in a software development task, if the original design only included requirements engineering, coding, and testing, one can increase the number of agents to handle steps like architectural design and detailed design, thereby improving task quality. Conversely, if there are excessive agents during a specific step, like coding, causing elevated communication costs without delivering substantial performance improvements compared to a smaller agent count, it may be essential to dynamically remove some agents to prevent resource waste.

Furthermore, agents can autonomously increase the number of agents [409] themselves to distribute their workload, ease their own burden, and achieve common goals more efficiently. Of course, when the workload becomes lighter, they can also reduce the number of agents delegated to their tasks to save system costs. 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 [22; 109; 520], 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 § 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 [405]. 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.

Therefore, the prospect of constructing a massive, stable, continuous agent system that faithfully replicates human work and life scenarios has become a promising research avenue. An agent with the ability to operate stably and perform tasks in a society comprising hundreds or even thousands of agents is more likely to find applications in real-world interactions with humans in the future.

6.5 Open Problems

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

The debate over whether LLM-based agents represent a potential path to AGI. [6] Artificial General Intelligence (AGI), also known as Strong AI, has long been the ultimate pursuit of humanity in the field of artificial intelligence, often referenced or depicted in many science fiction novels and films. There are various definitions of AGI, but here we refer to AGI as a type of artificial intelligence

[6] Note that the relevant debates are still ongoing, and the references here may include the latest viewpoints, technical blogs, and literature.

that demonstrates the ability to understand, learn, and apply knowledge across a wide range of tasks and domains, much like a human being [31; 658]. In contrast, Narrow AI is typically designed for specific tasks such as Go and Chess and lacks the broad cognitive abilities associated with human intelligence. Currently, whether large language models are a potential path to achieving AGI remains a highly debated and contentious topic [659; 660; 661; 662].

Given the breadth and depth of GPT-4's capabilities, some researchers (referred to as proponents) believe that large language models represented by GPT-4 can serve as early versions of AGI systems [31]. Following this line of thought, constructing agents based on LLMs has the potential to bring about more advanced versions of AGI systems. The main support for this argument lies in the idea that as long as they can be trained on a sufficiently large and diverse set of data that are projections of the real world, encompassing a rich array of tasks, LLM-based agents can develop AGI capabilities. Another interesting argument is that the act of autoregressive language modeling itself brings about compression and generalization abilities: just as humans have emerged with various peculiar and complex phenomena during their survival, language models, in the process of simply predicting the next token, also achieve an understanding of the world and the reasoning ability [579; 660; 663].

However, another group of individuals (referred to as opponents) believes that constructing agents based on LLMs cannot develop true Strong AI [664]. Their primary argument centers around the notion that LLMs, relying on autoregressive next-token prediction, cannot generate genuine intelligence because they do not simulate the true human thought process and merely provide reactive responses [660]. Moreover, LLMs also do not learn how the world operates by observing or experiencing it, leading to many foolish mistakes. They contend that a more advanced modeling approach, such as a world model [665], is necessary to develop AGI.

We cannot definitively determine which viewpoint is correct until true AGI is achieved, but we believe that such discussions and debates are beneficial for the overall development of the community.

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 [391; 666], while real-world environments are boundless, accommodate a wide range of tasks, and interacted 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.

First and foremost, a critical issue is the need for suitable hardware support when deploying the agent in a physical environment. This places high demands on the adaptability of the hardware. In a simulated environment, both the perception and action spaces of an agent are virtual. This means that in most cases, the results of the agent's operations, whether in perceiving inputs or generating outputs, can be guaranteed [395]. However, when an agent transitions to a real physical environment, its instructions may not be well executed by hardware devices such as sensors or robotic arms, significantly affecting the agent's task efficiency. Designing a dedicated interface or conversion mechanism between the agent and the hardware device is feasible. However, it can pose challenges to the system's reusability and simplicity.

In order to make this leap, the agent needs to have enhanced environmental generalization capabilities. To integrate seamlessly into the real physical world, they not only need to understand and reason about ambiguous instructions with implied meanings [128] but also possess the ability to learn and apply new skills flexibly [190; 592]. Furthermore, when dealing with an infinite and open world, the agent's limited context also poses significant challenges [236; 667]. This determines whether the agent can effectively handle a vast amount of information from the world and operate smoothly.

Finally, in a simulated environment, the inputs and outputs of the agent are virtual, allowing for countless trial and error attempts [432]. In such a scenario, the tolerance level for errors is high and does not lead to actual harm. However, in a physical environment, the agent's improper behavior or errors may cause real and sometimes irreversible harm to the environment. As a result, appropriate regulations and standards are highly necessary. We need to pay attention to the safety of agents when it comes to making decisions and generating actions, ensuring they do not pose threats or harm to the real world.

Collective intelligence in AI agents. What magical trick drives our intelligence? The reality is, there's no magic to it. As Marvin Minsky eloquently expressed in "The Society of Mind" [442], 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 amongst 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 becomes 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 / LLM-based Agent as a Service. With the development of cloud computing, the concept of XaaS (everything as a Service) has garnered widespread attention [668]. 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) [669; 670]. Similarly, cloud service platforms also provide basic platforms (Platform as a Service, PaaS) [671; 672], and specific business software (Software as a Service, SaaS) [673; 674], 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) [675]. Similarly, because 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 § 6.3, their robustness, trustworthiness, and concerns related to malicious use need to be considered before offering them as a service to customers.

7 Conclusion

This paper provides a comprehensive and systematic overview of LLM-based agents, discussing the potential challenges and opportunities in this flourishing field. We begin with a philosophical perspective, elucidating the origin and definition of agent, it evolution in the field of AI, and why LLMs are suited to serve as the main part of the brain of agents. Motivated by these background information, we present a general conceptual framework for LLM-based agents, comprising three main components: the 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, touching upon the mutual inspiration between LLM research and agent research, the evaluation of LLM-based agents, the risks associated with them, the opportunities in scaling the number of agents, and some open problems like Agent as a Service and whether LLM-based agents represent a potential path to AGI. We hope our efforts can provide inspirations to the community and facilitate research in related fields.

Acknowledgements

Thanks to Professor Guoyu Wang for carefully reviewing the ethics of the article. Thanks to Jinzhu Xiong for her excellent drawing skills to present an amazing performance of Figure 1.

References

[1] Russell, S. J. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[2] Diderot, D. *Diderot's early philosophical works*. 4. Open Court, 1911.

[3] Turing, A. M. *Computing machinery and intelligence*. Springer, 2009.

[4] Wooldridge, M. J., N. R. Jennings. Intelligent agents: theory and practice. *Knowl. Eng. Rev.*, 10(2):115–152, 1995.

[5] Schlosser, M. Agency. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edn., 2019.

[6] Agha, G. A. *Actors: a Model of Concurrent Computation in Distributed Systems (Parallel Processing, Semantics, Open, Programming Languages, Artificial Intelligence)*. Ph.D. thesis, University of Michigan, USA, 1985.

[7] Green, S., L. Hurst, B. Nangle, et al. Software agents: A review. *Department of Computer Science, Trinity College Dublin, Tech. Rep. TCS-CS-1997-06*, 1997.

[8] Genesereth, M. R., S. P. Ketchpel. Software agents. *Commun. ACM*, 37(7):48–53, 1994.

[9] Goodwin, R. Formalizing properties of agents. *J. Log. Comput.*, 5(6):763–781, 1995.

[10] Padgham, L., M. Winikoff. *Developing intelligent agent systems: A practical guide*. John Wiley & Sons, 2005.

[11] Shoham, Y. Agent oriented programming. In M. Masuch, L. Pólos, eds., *Knowledge Representation and Reasoning Under Uncertainty, Logic at Work [International Conference Logic at Work, Amsterdam, The Netherlands, December 17-19, 1992]*, vol. 808 of *Lecture Notes in Computer Science*, pages 123–129. Springer, 1992.

[12] Hutter, M. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

[13] Fikes, R., N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. In D. C. Cooper, ed., *Proceedings of the 2nd International Joint Conference on Artificial Intelligence. London, UK, September 1-3, 1971*, pages 608–620. William Kaufmann, 1971.

[14] Sacerdoti, E. D. Planning in a hierarchy of abstraction spaces. In N. J. Nilsson, ed., *Proceedings of the 3rd International Joint Conference on Artificial Intelligence. Standford, CA, USA, August 20-23, 1973*, pages 412–422. William Kaufmann, 1973.

[15] Brooks, R. A. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159, 1991.

[16] Maes, P. *Designing autonomous agents: Theory and practice from biology to engineering and back*. MIT press, 1990.

[17] Ribeiro, C. Reinforcement learning agents. *Artificial intelligence review*, 17:223–250, 2002.

[18] Kaelbling, L. P., M. L. Littman, A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[19] Guha, R. V., D. B. Lenat. Enabling agents to work together. *Communications of the ACM*, 37(7):126–142, 1994.

[20] Kaelbling, L. P., et al. An architecture for intelligent reactive systems. *Reasoning about actions and plans*, pages 395–410, 1987.

[21] Sutton, R. S., A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[22] Park, J. S., J. C. O'Brien, C. J. Cai, et al. Generative agents: Interactive simulacra of human behavior. *CoRR*, abs/2304.03442, 2023.

[23] Wang, Z., G. Zhang, K. Yang, et al. Interactive natural language processing. *CoRR*, abs/2305.13246, 2023.

[24] Ouyang, L., J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*. 2022.

[25] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[26] Wei, J., Y. Tay, R. Bommasani, et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022.

[27] Liu, R., R. Yang, C. Jia, et al. Training socially aligned language models in simulated human society. *CoRR*, abs/2305.16960, 2023.

[28] Sumers, T. R., S. Yao, K. Narasimhan, et al. Cognitive architectures for language agents. *CoRR*, abs/2309.02427, 2023.

[29] Weng, L. Llm-powered autonomous agents. *lilianweng.github.io*, 2023.

[30] Bisk, Y., A. Holtzman, J. Thomason, et al. Experience grounds language. In B. Webber, T. Cohn, Y. He, Y. Liu, eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8718–8735. Association for Computational Linguistics, 2020.

[31] Bubeck, S., V. Chandrasekaran, R. Eldan, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023.

[32] Anscombe, G. E. M. *Intention*. Harvard University Press, 2000.

[33] Davidson, D. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.

[34] —. I. agency. In A. Marras, R. N. Bronaugh, R. W. Binkley, eds., *Agent, Action, and Reason*, pages 1–37. University of Toronto Press, 1971.

[35] Dennett, D. C. Précis of the intentional stance. *Behavioral and brain sciences*, 11(3):495–505, 1988.

[36] Barandiaran, X. E., E. Di Paolo, M. Rohde. Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386, 2009.

[37] McCarthy, J. *Ascribing mental qualities to machines*. Stanford University. Computer Science Department, 1979.

[38] Rosenschein, S. J., L. P. Kaelbling. The synthesis of digital machines with provable epistemic properties. In *Theoretical aspects of reasoning about knowledge*, pages 83–98. Elsevier, 1986.

[39] Radford, A., K. Narasimhan, T. Salimans, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

[40] Radford, A., J. Wu, R. Child, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] Brown, T. B., B. Mann, N. Ryder, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[42] Lin, C., A. Jaech, X. Li, et al. Limitations of autoregressive models and their alternatives. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou, eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5147–5173. Association for Computational Linguistics, 2021.

[43] Tomasello, M. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press, 2005.

[44] Bloom, P. *How children learn the meanings of words*. MIT press, 2002.

[45] Zwaan, R. A., C. J. Madden. Embodied sentence comprehension. *Grounding cognition: The role of perception and action in memory, language, and thinking*, 22, 2005.

[46] Andreas, J. Language models as agent models. In Y. Goldberg, Z. Kozareva, Y. Zhang, eds., *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5769–5779. Association for Computational Linguistics, 2022.

[47] Wong, L., G. Grand, A. K. Lew, et al. From word models to world models: Translating from natural language to the probabilistic language of thought. *CoRR*, abs/2306.12672, 2023.

[48] Radford, A., R. Józefowicz, I. Sutskever. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444, 2017.

[49] Li, B. Z., M. I. Nye, J. Andreas. Implicit representations of meaning in neural language models. In C. Zong, F. Xia, W. Li, R. Navigli, eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1813–1827. Association for Computational Linguistics, 2021.

[50] Mukhopadhyay, U., L. M. Stephens, M. N. Huhns, et al. An intelligent system for document retrieval in distributed office environments. *J. Am. Soc. Inf. Sci.*, 37(3):123–135, 1986.

[51] Maes, P. Situated agents can have goals. *Robotics Auton. Syst.*, 6(1-2):49–70, 1990.

[52] Nilsson, N. J. Toward agent programs with circuit semantics. Tech. rep., 1992.

[53] Müller, J. P., M. Pischel. Modelling interacting agents in dynamic environments. In *Proceedings of the 11th European Conference on Artificial Intelligence*, pages 709–713. 1994.

[54] Brooks, R. A robust layered control system for a mobile robot. *IEEE journal on robotics and automation*, 2(1):14–23, 1986.

[55] Brooks, R. A. Intelligence without reason. In *The artificial life route to artificial intelligence*, pages 25–81. Routledge, 2018.

[56] Newell, A., H. A. Simon. Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19(3):113–126, 1976.

[57] Ginsberg, M. L. *Essentials of Artificial Intelligence*. Morgan Kaufmann, 1993.

[58] Wilkins, D. E. *Practical planning - extending the classical AI planning paradigm*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1988.

[59] Shardlow, N. *Action and agency in cognitive science*. Ph.D. thesis, Master's thesis, Department of Psychlogy, University of Manchester, Oxford . . . , 1990.

[60] Sacerdoti, E. D. The nonlinear nature of plans. In *Advance Papers of the Fourth International Joint Conference on Artificial Intelligence, Tbilisi, Georgia, USSR, September 3-8, 1975*, pages 206–214. 1975.

[61] Russell, S. J., E. Wefald. *Do the right thing: studies in limited rationality*. MIT press, 1991.

[62] Schoppers, M. Universal plans for reactive robots in unpredictable environments. In J. P. McDermott, ed., *Proceedings of the 10th International Joint Conference on Artificial Intelligence. Milan, Italy, August 23-28, 1987*, pages 1039–1046. Morgan Kaufmann, 1987.

[63] Brooks, R. A. A robust layered control system for a mobile robot. *IEEE J. Robotics Autom.*, 2(1):14–23, 1986.

[64] Minsky, M. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

[65] Isbell, C., C. R. Shelton, M. Kearns, et al. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*, pages 377–384. 2001.

[66] Watkins, C. J. C. H. Learning from delayed rewards, 1989.

[67] Rummery, G. A., M. Niranjan. *On-line Q-learning using connectionist systems*, vol. 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[68] Tesauro, G., et al. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[69] Li, Y. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[70] Silver, D., A. Huang, C. J. Maddison, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[71] Mnih, V., K. Kavukcuoglu, D. Silver, et al. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[72] Farebrother, J., M. C. Machado, M. Bowling. Generalization and regularization in DQN. *CoRR*, abs/1810.00123, 2018.

[73] Zhang, C., O. Vinyals, R. Munos, et al. A study on overfitting in deep reinforcement learning. *CoRR*, abs/1804.06893, 2018.

[74] Justesen, N., R. R. Torrado, P. Bontrager, et al. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

[75] Dulac-Arnold, G., N. Levine, D. J. Mankowitz, et al. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Mach. Learn.*, 110(9):2419–2468, 2021.

[76] Ghosh, D., J. Rahme, A. Kumar, et al. Why generalization in RL is difficult: Epistemic pomdps and implicit partial observability. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan, eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 25502–25515. 2021.

[77] Brys, T., A. Harutyunyan, M. E. Taylor, et al. Policy transfer using reward shaping. In G. Weiss, P. Yolum, R. H. Bordini, E. Elkind, eds., *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, pages 181–188. ACM, 2015.

[78] Parisotto, E., J. L. Ba, R. Salakhutdinov. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*, 2015.

[79] Zhu, Z., K. Lin, J. Zhou. Transfer learning in deep reinforcement learning: A survey. *CoRR*, abs/2009.07888, 2020.

[80] Duan, Y., J. Schulman, X. Chen, et al. Rl$ˆ2$: Fast reinforcement learning via slow reinforcement learning. *CoRR*, abs/1611.02779, 2016.

[81] Finn, C., P. Abbeel, S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In D. Precup, Y. W. Teh, eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.

[82] Gupta, A., R. Mendonca, Y. Liu, et al. Meta-reinforcement learning of structured exploration strategies. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett, eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5307–5316. 2018.

[83] Rakelly, K., A. Zhou, C. Finn, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In K. Chaudhuri, R. Salakhutdinov, eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, vol. 97 of *Proceedings of Machine Learning Research*, pages 5331–5340. PMLR, 2019.

[84] Fakoor, R., P. Chaudhari, S. Soatto, et al. Meta-q-learning. *arXiv preprint arXiv:1910.00125*, 2019.

[85] Vanschoren, J. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[86] Taylor, M. E., P. Stone. Transfer learning for reinforcement learning domains: A survey. *J. Mach. Learn. Res.*, 10:1633–1685, 2009.

[87] Tirinzoni, A., A. Sessa, M. Pirotta, et al. Importance weighted transfer of samples in reinforcement learning. In J. G. Dy, A. Krause, eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, vol. 80 of *Proceedings of Machine Learning Research*, pages 4943–4952. PMLR, 2018.

[88] Beck, J., R. Vuorio, E. Z. Liu, et al. A survey of meta-reinforcement learning. *CoRR*, abs/2301.08028, 2023.

[89] Wang, L., C. Ma, X. Feng, et al. A survey on large language model based autonomous agents. *CoRR*, abs/2308.11432, 2023.

[90] Nakano, R., J. Hilton, S. Balaji, et al. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021.

[91] Yao, S., J. Zhao, D. Yu, et al. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[92] Schick, T., J. Dwivedi-Yu, R. Dessì, et al. Toolformer: Language models can teach themselves to use tools. *CoRR*, abs/2302.04761, 2023.

[93] Lu, P., B. Peng, H. Cheng, et al. Chameleon: Plug-and-play compositional reasoning with large language models. *CoRR*, abs/2304.09842, 2023.

[94] Qin, Y., S. Hu, Y. Lin, et al. Tool learning with foundation models. *CoRR*, abs/2304.08354, 2023.

[95] Wei, J., X. Wang, D. Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*. 2022.

[96] Kojima, T., S. S. Gu, M. Reid, et al. Large language models are zero-shot reasoners. In *NeurIPS*. 2022.

[97] Wang, X., J. Wei, D. Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[98] Zhou, D., N. Schärli, L. Hou, et al. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[99] Xi, Z., S. Jin, Y. Zhou, et al. Self-polish: Enhance reasoning in large language models via problem refinement. *CoRR*, abs/2305.14497, 2023.

[100] Shinn, N., F. Cassano, B. Labash, et al. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

[101] Song, C. H., J. Wu, C. Washington, et al. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *CoRR*, abs/2212.04088, 2022.

[102] Akyürek, A. F., E. Akyürek, A. Kalyan, et al. RL4F: generating natural language feedback with reinforcement learning for repairing model outputs. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7716–7733. Association for Computational Linguistics, 2023.

[103] Peng, B., M. Galley, P. He, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813, 2023.

[104] Liu, H., C. Sferrazza, P. Abbeel. Languages are rewards: Hindsight finetuning using human feedback. *arXiv preprint arXiv:2302.02676*, 2023.

[105] Wei, J., M. Bosma, V. Y. Zhao, et al. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[106] Sanh, V., A. Webson, C. Raffel, et al. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[107] Chung, H. W., L. Hou, S. Longpre, et al. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.

[108] Li, G., H. A. A. K. Hammoud, H. Itani, et al. CAMEL: communicative agents for "mind" exploration of large scale language model society. *CoRR*, abs/2303.17760, 2023.

[109] Qian, C., X. Cong, C. Yang, et al. Communicative agents for software development. *CoRR*, abs/2307.07924, 2023.

[110] Boiko, D. A., R. MacKnight, G. Gomes. Emergent autonomous scientific research capabilities of large language models. *CoRR*, abs/2304.05332, 2023.

[111] Du, Y., S. Li, A. Torralba, et al. Improving factuality and reasoning in language models through multiagent debate. *CoRR*, abs/2305.14325, 2023.

[112] Liang, T., Z. He, W. Jiao, et al. Encouraging divergent thinking in large language models through multi-agent debate. *CoRR*, abs/2305.19118, 2023.

[113] Castelfranchi, C. Guarantees for autonomy in cognitive agent architecture. In M. J. Wooldridge, N. R. Jennings, eds., *Intelligent Agents, ECAI-94 Workshop on Agent Theories, Architectures, and Languages, Amsterdam, The Netherlands, August 8-9, 1994, Proceedings*, vol. 890 of *Lecture Notes in Computer Science*, pages 56–70. Springer, 1994.

[114] Gravitas, S. Auto-GPT: An Autonomous GPT-4 experiment, 2023. *URL https://github. com/Significant-Gravitas/Auto-GPT*, 2023.

[115] Nakajima, Y. BabyAGI. *Python. https://github. com/yoheinakajima/babyagi*, 2023.

[116] Yuan, A., A. Coenen, E. Reif, et al. Wordcraft: Story writing with large language models. In G. Jacucci, S. Kaski, C. Conati, S. Stumpf, T. Ruotsalo, K. Gajos, eds., *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, pages 841–852. ACM, 2022.

[117] Franceschelli, G., M. Musolesi. On the creativity of large language models. *CoRR*, abs/2304.00008, 2023.

[118] Zhu, D., J. Chen, X. Shen, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[119] Yin, S., C. Fu, S. Zhao, et al. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023.

[120] Driess, D., F. Xia, M. S. M. Sajjadi, et al. Palm-e: An embodied multimodal language model. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR, 2023.

[121] Mu, Y., Q. Zhang, M. Hu, et al. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *CoRR*, abs/2305.15021, 2023.

[122] Brown, J. W. Beyond conflict monitoring: Cognitive control and the neural basis of thinking before you act. *Current Directions in Psychological Science*, 22(3):179–185, 2013.

[123] Kang, J., R. Laroche, X. Yuan, et al. Think before you act: Decision transformers with internal working memory. *CoRR*, abs/2305.16338, 2023.

[124] Valmeekam, K., S. Sreedharan, M. Marquez, et al. On the planning abilities of large language models (A critical investigation with a proposed benchmark). *CoRR*, abs/2302.06706, 2023.

[125] Liu, B., Y. Jiang, X. Zhang, et al. LLM+P: empowering large language models with optimal planning proficiency. *CoRR*, abs/2304.11477, 2023.

[126] Liu, H., C. Sferrazza, P. Abbeel. Chain of hindsight aligns language models with feedback. *CoRR*, abs/2302.02676, 2023.

[127] Lin, Y., Y. Chen. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *CoRR*, abs/2305.13711, 2023.

[128] Lin, J., D. Fried, D. Klein, et al. Inferring rewards from language in context. In S. Muresan, P. Nakov, A. Villavicencio, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8546–8560. Association for Computational Linguistics, 2022.

[129] Fu, Y., H. Peng, T. Khot, et al. Improving language model negotiation with self-play and in-context learning from AI feedback. *CoRR*, abs/2305.10142, 2023.

[130] Zhang, H., W. Du, J. Shan, et al. Building cooperative embodied agents modularly with large language models. *CoRR*, abs/2307.02485, 2023.

[131] Darwin's, C. On the origin of species. *published on*, 24:1, 1859.

[132] Bang, Y., S. Cahyawijaya, N. Lee, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023, 2023.

[133] Fang, T., S. Yang, K. Lan, et al. Is chatgpt a highly fluent grammatical error correction system? A comprehensive evaluation. *CoRR*, abs/2304.01746, 2023.

[134] Lu, A., H. Zhang, Y. Zhang, et al. Bounding the capabilities of large language models in open text generation with prompt constraints. In A. Vlachos, I. Augenstein, eds., *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1937–1963. Association for Computational Linguistics, 2023.

[135] Buehler, M. C., J. Adamy, T. H. Weisswange. Theory of mind based assistive communication in complex human robot cooperation. *CoRR*, abs/2109.01355, 2021.

[136] Shapira, N., M. Levy, S. H. Alavi, et al. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *CoRR*, abs/2305.14763, 2023.

[137] Hill, F., K. Cho, A. Korhonen. Learning distributed representations of sentences from unlabelled data. In K. Knight, A. Nenkova, O. Rambow, eds., *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1367–1377. The Association for Computational Linguistics, 2016.

[138] Collobert, R., J. Weston, L. Bottou, et al. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, 2011.

[139] Kaplan, J., S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[140] Roberts, A., C. Raffel, N. Shazeer. How much knowledge can you pack into the parameters of a language model? In B. Webber, T. Cohn, Y. He, Y. Liu, eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics, 2020.

[141] Tandon, N., A. S. Varde, G. de Melo. Commonsense knowledge in machine intelligence. *SIGMOD Rec.*, 46(4):49–52, 2017.

[142] Vulic, I., E. M. Ponti, R. Litschko, et al. Probing pretrained language models for lexical semantics. In B. Webber, T. Cohn, Y. He, Y. Liu, eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7222–7240. Association for Computational Linguistics, 2020.

[143] Hewitt, J., C. D. Manning. A structural probe for finding syntax in word representations. In J. Burstein, C. Doran, T. Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics, 2019.

[144] Rau, L. F., P. S. Jacobs, U. Zernik. Information extraction and text summarization using linguistic knowledge acquisition. *Inf. Process. Manag.*, 25(4):419–428, 1989.

[145] Yang, K., Z. Chen, Y. Cai, et al. Improved automatic keyword extraction given more semantic knowledge. In H. Gao, J. Kim, Y. Sakurai, eds., *Database Systems for Advanced Applications - DASFAA 2016 International Workshops: BDMS, BDQM, MoI, and SeCoP, Dallas, TX, USA, April 16-19, 2016, Proceedings*, vol. 9645 of *Lecture Notes in Computer Science*, pages 112–125. Springer, 2016.

[146] Beloucif, M., C. Biemann. Probing pre-trained language models for semantic attributes and their values. In M. Moens, X. Huang, L. Specia, S. W. Yih, eds., *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2554–2559. Association for Computational Linguistics, 2021.

[147] Zhang, Z., H. Zhao. Advances in multi-turn dialogue comprehension: A survey. *CoRR*, abs/2103.03125, 2021.

[148] Safavi, T., D. Koutra. Relational world knowledge representation in contextual language models: A review. In M. Moens, X. Huang, L. Specia, S. W. Yih, eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1053–1067. Association for Computational Linguistics, 2021.

[149] Jiang, Z., F. F. Xu, J. Araki, et al. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020.

[150] Madaan, A., S. Zhou, U. Alon, et al. Language models of code are few-shot commonsense learners. In Y. Goldberg, Z. Kozareva, Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1384–1403. Association for Computational Linguistics, 2022.

[151] Xu, F. F., U. Alon, G. Neubig, et al. A systematic evaluation of large language models of code. In S. Chaudhuri, C. Sutton, eds., *MAPS@PLDI 2022: 6th ACM SIGPLAN International Symposium on Machine Programming, San Diego, CA, USA, 13 June 2022*, pages 1–10. ACM, 2022.

[152] Cobbe, K., V. Kosaraju, M. Bavarian, et al. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

[153] Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, et al. Large language models in medicine. *Nature medicine*, pages 1–11, 2023.

[154] Lai, Y., C. Li, Y. Wang, et al. DS-1000: A natural and reliable benchmark for data science code generation. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 18319–18345. PMLR, 2023.

[155] AlKhamissi, B., M. Li, A. Celikyilmaz, et al. A review on language models as knowledge bases. *CoRR*, abs/2204.06031, 2022.

[156] Kemker, R., M. McClure, A. Abitino, et al. Measuring catastrophic forgetting in neural networks. In S. A. McIlraith, K. Q. Weinberger, eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3390–3398. AAAI Press, 2018.

[157] Cao, N. D., W. Aziz, I. Titov. Editing factual knowledge in language models. In M. Moens, X. Huang, L. Specia, S. W. Yih, eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics, 2021.

[158] Yao, Y., P. Wang, B. Tian, et al. Editing large language models: Problems, methods, and opportunities. *CoRR*, abs/2305.13172, 2023.

[159] Mitchell, E., C. Lin, A. Bosselut, et al. Memory-based model editing at scale. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato, eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR, 2022.

[160] Manakul, P., A. Liusie, M. J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *CoRR*, abs/2303.08896, 2023.

[161] Li, M., B. Peng, Z. Zhang. Self-checker: Plug-and-play modules for fact-checking with large language models. *CoRR*, abs/2305.14623, 2023.

[162] Gou, Z., Z. Shao, Y. Gong, et al. CRITIC: large language models can self-correct with tool-interactive critiquing. *CoRR*, abs/2305.11738, 2023.

[163] Lewis, M., Y. Liu, N. Goyal, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.

[164] Park, H. H., Y. Vyas, K. Shah. Efficient classification of long documents using transformers. In S. Muresan, P. Nakov, A. Villavicencio, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 702–709. Association for Computational Linguistics, 2022.

[165] Guo, M., J. Ainslie, D. C. Uthus, et al. Longt5: Efficient text-to-text transformer for long sequences. In M. Carpuat, M. de Marneffe, I. V. M. Ruíz, eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 724–736. Association for Computational Linguistics, 2022.

[166] Ainslie, J., T. Lei, M. de Jong, et al. Colt5: Faster long-range transformers with conditional computation. *CoRR*, abs/2303.09752, 2023.

[167] Ruoss, A., G. Delétang, T. Genewein, et al. Randomized positional encodings boost length generalization of transformers. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1889–1903. Association for Computational Linguistics, 2023.

[168] Liang, X., B. Wang, H. Huang, et al. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *CoRR*, abs/2304.13343, 2023.

[169] Shinn, N., B. Labash, A. Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *CoRR*, abs/2303.11366, 2023.

[170] Zhong, W., L. Guo, Q. Gao, et al. Memorybank: Enhancing large language models with long-term memory. *CoRR*, abs/2305.10250, 2023.

[171] Chan, C., W. Chen, Y. Su, et al. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201, 2023.

[172] Zhu, X., Y. Chen, H. Tian, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *CoRR*, abs/2305.17144, 2023.

[173] Modarressi, A., A. Imani, M. Fayyaz, et al. RET-LLM: towards a general read-write memory for large language models. *CoRR*, abs/2305.14322, 2023.

[174] Lin, J., H. Zhao, A. Zhang, et al. Agentsims: An open-source sandbox for large language model evaluation. *CoRR*, abs/2308.04026, 2023.

[175] Hu, C., J. Fu, C. Du, et al. Chatdb: Augmenting llms with databases as their symbolic memory. *CoRR*, abs/2306.03901, 2023.

[176] Huang, Z., S. Gutierrez, H. Kamana, et al. Memory sandbox: Transparent and interactive memory management for conversational agents. *CoRR*, abs/2308.01542, 2023.

[177] Creswell, A., M. Shanahan, I. Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[178] Madaan, A., N. Tandon, P. Gupta, et al. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651, 2023.

[179] Ichter, B., A. Brohan, Y. Chebotar, et al. Do as I can, not as I say: Grounding language in robotic affordances. In K. Liu, D. Kulic, J. Ichnowski, eds., *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, vol. 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 2022.

[180] Shen, Y., K. Song, X. Tan, et al. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.

[181] Yao, S., D. Yu, J. Zhao, et al. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601, 2023.

[182] Wu, Y., S. Y. Min, Y. Bisk, et al. Plan, eliminate, and track - language models are good teachers for embodied agents. *CoRR*, abs/2305.02412, 2023.

[183] Wang, Z., S. Cai, A. Liu, et al. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *CoRR*, abs/2302.01560, 2023.

[184] Hao, S., Y. Gu, H. Ma, et al. Reasoning with language model is planning with world model. *CoRR*, abs/2305.14992, 2023.

[185] Lin, B. Y., Y. Fu, K. Yang, et al. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *CoRR*, abs/2305.17390, 2023.

[186] Karpas, E., O. Abend, Y. Belinkov, et al. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *CoRR*, abs/2205.00445, 2022.

[187] Huang, W., F. Xia, T. Xiao, et al. Inner monologue: Embodied reasoning through planning with language models. In K. Liu, D. Kulic, J. Ichnowski, eds., *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, vol. 205 of *Proceedings of Machine Learning Research*, pages 1769–1782. PMLR, 2022.

[188] Chen, Z., K. Zhou, B. Zhang, et al. Chatcot: Tool-augmented chain-of-thought reasoning on chat-based large language models. *CoRR*, abs/2305.14323, 2023.

[189] Wu, T., M. Terry, C. J. Cai. AI chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In S. D. J. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. M. Drucker, J. R. Williamson, K. Yatani, eds., *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 385:1–385:22. ACM, 2022.

[190] Wang, G., Y. Xie, Y. Jiang, et al. Voyager: An open-ended embodied agent with large language models. *CoRR*, abs/2305.16291, 2023.

[191] Zhao, X., M. Li, C. Weber, et al. Chat with the environment: Interactive multimodal perception using large language models. *CoRR*, abs/2303.08268, 2023.

[192] Miao, N., Y. W. Teh, T. Rainforth. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *CoRR*, abs/2308.00436, 2023.

[193] Wang, X., W. Wang, Y. Cao, et al. Images speak in images: A generalist painter for in-context visual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6830–6839. IEEE, 2023.

[194] Wang, C., S. Chen, Y. Wu, et al. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111, 2023.

[195] Dong, Q., L. Li, D. Dai, et al. A survey for in-context learning. *CoRR*, abs/2301.00234, 2023.

[196] Ke, Z., B. Liu. Continual learning of natural language processing tasks: A survey. *ArXiv*, abs/2211.12701, 2022.

[197] Wang, L., X. Zhang, H. Su, et al. A comprehensive survey of continual learning: Theory, method and application. *ArXiv*, abs/2302.00487, 2023.

[198] Razdaibiedina, A., Y. Mao, R. Hou, et al. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*. 2023.

[199] Marshall, L. H., H. W. Magoun. *Discoveries in the human brain: neuroscience prehistory, brain structure, and function*. Springer Science & Business Media, 2013.

[200] Searle, J. R. What is language: some preliminary remarks. *Explorations in Pragmatics. Linguistic, cognitive and intercultural aspects*, pages 7–37, 2007.

[201] Touvron, H., T. Lavril, G. Izacard, et al. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[202] Scao, T. L., A. Fan, C. Akiki, et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022.

[203] Almazrouei, E., H. Alobeidli, A. Alshamsi, et al. Falcon-40b: an open large language model with state-of-the-art performance, 2023.

[204] Serban, I. V., R. Lowe, L. Charlin, et al. Generative deep neural networks for dialogue: A short review. *CoRR*, abs/1611.06216, 2016.

[205] Vinyals, O., Q. V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.

[206] Adiwardana, D., M. Luong, D. R. So, et al. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977, 2020.

[207] Zhuge, M., H. Liu, F. Faccio, et al. Mindstorms in natural language-based societies of mind. *CoRR*, abs/2305.17066, 2023.

[208] Roller, S., E. Dinan, N. Goyal, et al. Recipes for building an open-domain chatbot. In P. Merlo, J. Tiedemann, R. Tsarfaty, eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics, 2021.

[209] Taori, R., I. Gulrajani, T. Zhang, et al. Stanford alpaca: An instruction-following llama model, 2023.

[210] Raffel, C., N. Shazeer, A. Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[211] Ge, Y., W. Hua, J. Ji, et al. Openagi: When LLM meets domain experts. *CoRR*, abs/2304.04370, 2023.

[212] Rajpurkar, P., J. Zhang, K. Lopyrev, et al. Squad: 100, 000+ questions for machine comprehension of text. In J. Su, X. Carreras, K. Duh, eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016.

[213] Ahuja, K., R. Hada, M. Ochieng, et al. MEGA: multilingual evaluation of generative AI. *CoRR*, abs/2303.12528, 2023.

[214] See, A., A. Pappu, R. Saxena, et al. Do massively pretrained language models make better storytellers? In M. Bansal, A. Villavicencio, eds., *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 843–861. Association for Computational Linguistics, 2019.

[215] Radford, A., J. Wu, D. Amodei, et al. Better language models and their implications. *OpenAI blog*, 1(2), 2019.

[216] McCoy, R. T., P. Smolensky, T. Linzen, et al. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *CoRR*, abs/2111.09509, 2021.

[217] Tellex, S., T. Kollar, S. Dickerson, et al. Understanding natural language commands for robotic navigation and mobile manipulation. In W. Burgard, D. Roth, eds., *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, pages 1507–1514. AAAI Press, 2011.

[218] Christiano, P. F., J. Leike, T. B. Brown, et al. Deep reinforcement learning from human preferences. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307. 2017.

[219] Basu, C., M. Singhal, A. D. Dragan. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In T. Kanda, S. Sabanovic, G. Hoffman, A. Tapus, eds., *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2018, Chicago, IL, USA, March 05-08, 2018*, pages 132–140. ACM, 2018.

[220] Sumers, T. R., M. K. Ho, R. X. D. Hawkins, et al. Learning rewards from linguistic feedback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6002–6010. AAAI Press, 2021.

[221] Jeon, H. J., S. Milli, A. D. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[222] McShane, M. Reference resolution challenges for intelligent agents: The need for knowledge. *IEEE Intell. Syst.*, 24(4):47–58, 2009.

[223] Gururangan, S., A. Marasovic, S. Swayamdipta, et al. Don't stop pretraining: Adapt language models to domains and tasks. In D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics, 2020.

[224] Shi, F., X. Chen, K. Misra, et al. Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR, 2023.

[225] Zhang, Y., Y. Li, L. Cui, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219, 2023.

[226] Mialon, G., R. Dessì, M. Lomeli, et al. Augmented language models: a survey. *CoRR*, abs/2302.07842, 2023.

[227] Ren, R., Y. Wang, Y. Qu, et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *CoRR*, abs/2307.11019, 2023.

[228] Nuxoll, A. M., J. E. Laird. Extending cognitive architecture with episodic memory. In *AAAI*, pages 1560–1564. 2007.

[229] Squire, L. R. Mechanisms of memory. *Science*, 232(4758):1612–1619, 1986.

[230] Schwabe, L., K. Nader, J. C. Pruessner. Reconsolidation of human memory: brain mechanisms and clinical relevance. *Biological psychiatry*, 76(4):274–280, 2014.

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

[232] Zhang, X., F. Wei, M. Zhou. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In A. Korhonen, D. R. Traum, L. Màrquez, eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics, 2019.

[233] Mohtashami, A., M. Jaggi. Landmark attention: Random-access infinite context length for transformers. *CoRR*, abs/2305.16300, 2023.

[234] Chalkidis, I., X. Dai, M. Fergadiotis, et al. An exploration of hierarchical attention transformers for efficient long document classification. *CoRR*, abs/2210.05529, 2022.

[235] Nie, Y., H. Huang, W. Wei, et al. Capturing global structural information in long document question answering with compressive graph selector network. In Y. Goldberg, Z. Kozareva, Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5036–5047. Association for Computational Linguistics, 2022.

[236] Bertsch, A., U. Alon, G. Neubig, et al. Unlimiformer: Long-range transformers with unlimited length input. *CoRR*, abs/2305.01625, 2023.

[237] Manakul, P., M. J. F. Gales. Sparsity and sentence structure in encoder-decoder attention of summarization systems. In M. Moens, X. Huang, L. Specia, S. W. Yih, eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9359–9368. Association for Computational Linguistics, 2021.

[238] Zaheer, M., G. Guruganesh, K. A. Dubey, et al. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020.

[239] Zhao, A., D. Huang, Q. Xu, et al. Expel: LLM agents are experiential learners. *CoRR*, abs/2308.10144, 2023.

[240] Zhou, X., G. Li, Z. Liu. LLM as DBA. *CoRR*, abs/2308.05481, 2023.

[241] Wason, P. C. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

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

[243] Galotti, K. M. Approaches to studying formal and everyday reasoning. *Psychological bulletin*, 105(3):331, 1989.

[244] Huang, J., K. C. Chang. Towards reasoning in large language models: A survey. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics, 2023.

[245] Webb, T. W., K. J. Holyoak, H. Lu. Emergent analogical reasoning in large language models. *CoRR*, abs/2212.09196, 2022.

[246] Feng, G., B. Zhang, Y. Gu, et al. Towards revealing the mystery behind chain of thought: a theoretical perspective. *CoRR*, abs/2305.15408, 2023.

[247] Grafman, J., L. Spector, M. J. Rattermann. Planning and the brain. In *The cognitive psychology of planning*, pages 191–208. Psychology Press, 2004.

[248] Unterrainer, J. M., A. M. Owen. Planning and problem solving: from neuropsychology to functional neuroimaging. *Journal of Physiology-Paris*, 99(4-6):308–317, 2006.

[249] Zula, K. J., T. J. Chermack. Integrative literature review: Human capital planning: A review of literature and implications for human resource development. *Human Resource Development Review*, 6(3):245–262, 2007.

[250] Bratman, M. E., D. J. Israel, M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational intelligence*, 4(3):349–355, 1988.

[251] Russell, S., P. Norvig. *Artificial intelligence - a modern approach, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, 2003.

[252] Fainstein, S. S., J. DeFilippis. *Readings in planning theory*. John Wiley & Sons, 2015.

[253] Sebastia, L., E. Onaindia, E. Marzal. Decomposition of planning problems. *Ai Communications*, 19(1):49–81, 2006.

[254] Crosby, M., M. Rovatsos, R. Petrick. Automated agent decomposition for classical planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 23, pages 46–54. 2013.

[255] Xu, B., Z. Peng, B. Lei, et al. Rewoo: Decoupling reasoning from observations for efficient augmented language models. *CoRR*, abs/2305.18323, 2023.

[256] Raman, S. S., V. Cohen, E. Rosen, et al. Planning with large language models via corrective re-prompting. *CoRR*, abs/2211.09935, 2022.

[257] Lyu, Q., S. Havaldar, A. Stein, et al. Faithful chain-of-thought reasoning. *CoRR*, abs/2301.13379, 2023.

[258] Huang, W., P. Abbeel, D. Pathak, et al. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato, eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR, 2022.

[259] Dagan, G., F. Keller, A. Lascarides. Dynamic planning with a LLM. *CoRR*, abs/2308.06391, 2023.

[260] Rana, K., J. Haviland, S. Garg, et al. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *CoRR*, abs/2307.06135, 2023.

[261] Peters, M. E., M. Neumann, M. Iyyer, et al. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana, 2018.

[262] Devlin, J., M. Chang, K. Lee, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, T. Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[263] Solaiman, I., C. Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873, 2021.

[264] Bach, S. H., V. Sanh, Z. X. Yong, et al. Promptsource: An integrated development environment and repository for natural language prompts. In V. Basile, Z. Kozareva, S. Stajner, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, Dublin, Ireland, May 22-27, 2022*, pages 93–104. Association for Computational Linguistics, 2022.

[265] Iyer, S., X. V. Lin, R. Pasunuru, et al. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017, 2022.

[266] Winston, P. H. Learning and reasoning by analogy. *Commun. ACM*, 23(12):689–703, 1980.

[267] Lu, Y., M. Bartolo, A. Moore, et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In S. Muresan, P. Nakov, A. Villavicencio, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics, 2022.

[268] Tsimpoukelli, M., J. Menick, S. Cabi, et al. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan, eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212. 2021.

[269] Bar, A., Y. Gandelsman, T. Darrell, et al. Visual prompting via image inpainting. In *NeurIPS*. 2022.

[270] Zhu, W., H. Liu, Q. Dong, et al. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675, 2023.

[271] Zhang, Z., L. Zhou, C. Wang, et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *CoRR*, abs/2303.03926, 2023.

[272] Zhang, J., J. Zhang, K. Pertsch, et al. Bootstrap your own skills: Learning to solve new tasks with large language model guidance. In *7th Annual Conference on Robot Learning*. 2023.

[273] McCloskey, M., N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[274] Kirkpatrick, J., R. Pascanu, N. Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[275] Li, Z., D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[276] Farajtabar, M., N. Azizan, A. Mott, et al. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3762–3773. PMLR, 2020.

[277] Smith, J. S., Y.-C. Hsu, L. Zhang, et al. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023.

[278] Lopez-Paz, D., M. Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[279] de Masson D'Autume, C., S. Ruder, L. Kong, et al. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[280] Rolnick, D., A. Ahuja, J. Schwarz, et al. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[281] Serrà, J., D. Surís, M. Miron, et al. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*. 2018.

[282] Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[283] van den Oord, A., O. Vinyals, K. Kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315. 2017.

[284] Mehta, S., M. Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[285] Tolstikhin, I. O., N. Houlsby, A. Kolesnikov, et al. Mlp-mixer: An all-mlp architecture for vision. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan, eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24261–24272. 2021.

[286] Huang, S., L. Dong, W. Wang, et al. Language is not all you need: Aligning perception with language models. *CoRR*, abs/2302.14045, 2023.

[287] Li, J., D. Li, S. Savarese, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[288] Dai, W., J. Li, D. Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023.

[289] Gong, T., C. Lyu, S. Zhang, et al. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790, 2023.

[290] Alayrac, J., J. Donahue, P. Luc, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*. 2022.

[291] Su, Y., T. Lan, H. Li, et al. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.

[292] Liu, H., C. Li, Q. Wu, et al. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023.

[293] Huang, R., M. Li, D. Yang, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. *CoRR*, abs/2304.12995, 2023.

[294] Gong, Y., Y. Chung, J. R. Glass. AST: audio spectrogram transformer. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, P. Motlícek, eds., *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 571–575. ISCA, 2021.

[295] Hsu, W., B. Bolte, Y. H. Tsai, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021.

[296] Chen, F., M. Han, H. Zhao, et al. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. *CoRR*, abs/2305.04160, 2023.

[297] Zhang, H., X. Li, L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *CoRR*, abs/2306.02858, 2023.

[298] Liu, Z., Y. He, W. Wang, et al. Interngpt: Solving vision-centric tasks by interacting with chatbots beyond language. *CoRR*, abs/2305.05662, 2023.

[299] Hubel, D. H., T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[300] Logothetis, N. K., D. L. Sheinberg. Visual object recognition. *Annual review of neuroscience*, 19(1):577–621, 1996.

[301] OpenAI. Openai: Introducing chatgpt. Website, 2022. https://openai.com/blog/chatgpt.

[302] Lu, J., X. Ren, Y. Ren, et al. Improving contextual language models for response retrieval in multi-turn conversation. In J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu, eds., *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1805–1808. ACM, 2020.

[303] Huang, L., W. Wang, J. Chen, et al. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4633–4642. IEEE, 2019.

[304] Pan, Y., T. Yao, Y. Li, et al. X-linear attention networks for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10968–10977. Computer Vision Foundation / IEEE, 2020.

[305] Cornia, M., M. Stefanini, L. Baraldi, et al. M^2: Meshed-memory transformer for image captioning. *CoRR*, abs/1912.08226, 2019.

[306] Chen, J., H. Guo, K. Yi, et al. Visualgpt: Data-efficient image captioning by balancing visual input and linguistic knowledge from pretraining. *CoRR*, abs/2102.10407, 2021.

[307] Li, K., Y. He, Y. Wang, et al. Videochat: Chat-centric video understanding. *CoRR*, abs/2305.06355, 2023.

[308] Lin, J., Y. Du, O. Watkins, et al. Learning to model the world with language. *CoRR*, abs/2308.01399, 2023.

[309] Vaswani, A., N. Shazeer, N. Parmar, et al. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008. 2017.

[310] Touvron, H., M. Cord, M. Douze, et al. Training data-efficient image transformers & distillation through attention. In M. Meila, T. Zhang, eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021.

[311] Lu, J., C. Clark, R. Zellers, et al. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[312] Peng, Z., W. Wang, L. Dong, et al. Kosmos-2: Grounding multimodal large language models to the world. *CoRR*, abs/2306.14824, 2023.

[313] Lyu, C., M. Wu, L. Wang, et al. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *CoRR*, abs/2306.09093, 2023.

[314] Maaz, M., H. A. Rasheed, S. H. Khan, et al. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424, 2023.

[315] Chen, M., I. Laina, A. Vedaldi. Training-free layout control with cross-attention guidance. *CoRR*, abs/2304.03373, 2023.

[316] Radford, A., J. W. Kim, T. Xu, et al. Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[317] Ren, Y., Y. Ruan, X. Tan, et al. Fastspeech: Fast, robust and controllable text to speech. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett, eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174. 2019.

[318] Ye, Z., Z. Zhao, Y. Ren, et al. Syntaspeech: Syntax-aware generative adversarial text-to-speech. In L. D. Raedt, ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4468–4474. ijcai.org, 2022.

[319] Kim, J., J. Kong, J. Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In M. Meila, T. Zhang, eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR, 2021.

[320] Wang, Z., S. Cornell, S. Choi, et al. Tf-gridnet: Integrating full- and sub-band modeling for speech separation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:3221–3236, 2023.

[321] Liu, J., C. Li, Y. Ren, et al. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11020–11028. AAAI Press, 2022.

[322] Inaguma, H., S. Dalmia, B. Yan, et al. Fast-md: Fast multi-decoder end-to-end speech translation with non-autoregressive hidden intermediates. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 922–929. IEEE, 2021.

[323] Flanagan, J. L. *Speech analysis synthesis and perception*, vol. 3. Springer Science & Business Media, 2013.

[324] Schwarz, B. Mapping the world in 3d. *Nature Photonics*, 4(7):429–430, 2010.

[325] Parkinson, B. W., J. J. Spilker. *Progress in astronautics and aeronautics: Global positioning system: Theory and applications*, vol. 164. Aiaa, 1996.

[326] Parisi, A., Y. Zhao, N. Fiedel. TALM: tool augmented language models. *CoRR*, abs/2205.12255, 2022.

[327] Clarebout, G., J. Elen, N. A. J. Collazo, et al. *Metacognition and the Use of Tools*, pages 187–195. Springer New York, New York, NY, 2013.

[328] Wu, C., S. Yin, W. Qi, et al. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023.

[329] Cai, T., X. Wang, T. Ma, et al. Large language models as tool makers. *CoRR*, abs/2305.17126, 2023.

[330] Qian, C., C. Han, Y. R. Fung, et al. CREATOR: disentangling abstract and concrete reasonings of large language models through tool creation. *CoRR*, abs/2305.14318, 2023.

[331] Chen, X., M. Lin, N. Schärli, et al. Teaching large language models to self-debug. *CoRR*, abs/2304.05128, 2023.

[332] Liu, H., L. Lee, K. Lee, et al. Instruction-following agents with jointly pre-trained vision-language models. *arXiv preprint arXiv:2210.13431*, 2022.

[333] Lynch, C., A. Wahid, J. Tompson, et al. Interactive language: Talking to robots in real time. *CoRR*, abs/2210.06407, 2022.

[334] Jin, C., W. Tan, J. Yang, et al. Alphablock: Embodied finetuning for vision-language reasoning in robot manipulation. *CoRR*, abs/2305.18898, 2023.

[335] Shah, D., B. Osinski, B. Ichter, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In K. Liu, D. Kulic, J. Ichnowski, eds., *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, vol. 205 of *Proceedings of Machine Learning Research*, pages 492–504. PMLR, 2022.

[336] Zhou, G., Y. Hong, Q. Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *CoRR*, abs/2305.16986, 2023.

[337] Fan, L., G. Wang, Y. Jiang, et al. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *NeurIPS*. 2022.

[338] Kanitscheider, I., J. Huizinga, D. Farhi, et al. Multi-task curriculum learning in a complex, visual, hard-exploration domain: Minecraft. *CoRR*, abs/2106.14876, 2021.

[339] Nottingham, K., P. Ammanabrolu, A. Suhr, et al. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 26311–26325. PMLR, 2023.

[340] Sumers, T., K. Marino, A. Ahuja, et al. Distilling internet-scale vision-language models into embodied agents. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 32797–32818. PMLR, 2023.

[341] Carlini, N., J. Hayes, M. Nasr, et al. Extracting training data from diffusion models. *CoRR*, abs/2301.13188, 2023.

[342] Savelka, J., K. D. Ashley, M. A. Gray, et al. Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? In F. Lagioia, J. Mumford, D. Odekerken, H. Westermann, eds., *Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023), Braga, Portugal, 23rd September, 2023*, vol. 3441 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org, 2023.

[343] Ling, C., X. Zhao, J. Lu, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2023.

[344] Linardatos, P., V. Papastefanopoulos, S. Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2021.

[345] Zou, A., Z. Wang, J. Z. Kolter, et al. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043, 2023.

[346] Hussein, A., M. M. Gaber, E. Elyan, et al. Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2):21:1–21:35, 2017.

[347] Liu, Y., A. Gupta, P. Abbeel, et al. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1118–1125. IEEE, 2018.

[348] Baker, B., I. Akkaya, P. Zhokov, et al. Video pretraining (VPT): learning to act by watching unlabeled online videos. In *NeurIPS*. 2022.

[349] Levine, S., P. Pastor, A. Krizhevsky, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int. J. Robotics Res.*, 37(4-5):421–436, 2018.

[350] Zheng, R., S. Dou, S. Gao, et al. Secrets of RLHF in large language models part I: PPO. *CoRR*, abs/2307.04964, 2023.

[351] Bengio, Y., J. Louradour, R. Collobert, et al. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48. Association for Computing Machinery, New York, NY, USA, 2009.

[352] Chen, M., J. Tworek, H. Jun, et al. Evaluating large language models trained on code, 2021.

[353] Pan, S., L. Luo, Y. Wang, et al. Unifying large language models and knowledge graphs: A roadmap. *CoRR*, abs/2306.08302, 2023.

[354] Bran, A. M., S. Cox, A. D. White, et al. Chemcrow: Augmenting large-language models with chemistry tools, 2023.

[355] Ruan, J., Y. Chen, B. Zhang, et al. TPTU: task planning and tool usage of large language model-based AI agents. *CoRR*, abs/2308.03427, 2023.

[356] Ogundare, O., S. Madasu, N. Wiggins. Industrial engineering with large language models: A case study of chatgpt's performance on oil & gas problems, 2023.

[357] Smith, L., M. Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005.

[358] Duan, J., S. Yu, H. L. Tan, et al. A survey of embodied AI: from simulators to research tasks. *IEEE Trans. Emerg. Top. Comput. Intell.*, 6(2):230–244, 2022.

[359] Mnih, V., K. Kavukcuoglu, D. Silver, et al. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[360] Silver, D., A. Huang, C. J. Maddison, et al. Mastering the game of go with deep neural networks and tree search. *Nat.*, 529(7587):484–489, 2016.

[361] Kalashnikov, D., A. Irpan, P. Pastor, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *CoRR*, abs/1806.10293, 2018.

[362] Nguyen, H., H. M. La. Review of deep reinforcement learning for robot manipulation. In *3rd IEEE International Conference on Robotic Computing, IRC 2019, Naples, Italy, February 25-27, 2019*, pages 590–595. IEEE, 2019.

[363] Dasgupta, I., C. Kaeser-Chen, K. Marino, et al. Collaborating with language models for embodied reasoning. *CoRR*, abs/2302.00763, 2023.

[364] Puig, X., K. Ra, M. Boben, et al. Virtualhome: Simulating household activities via programs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8494–8502. Computer Vision Foundation / IEEE Computer Society, 2018.

[365] Hong, Y., Q. Wu, Y. Qi, et al. A recurrent vision-and-language BERT for navigation. *CoRR*, abs/2011.13922, 2020.

[366] Suglia, A., Q. Gao, J. Thomason, et al. Embodied BERT: A transformer model for embodied, language-guided visual task completion. *CoRR*, abs/2108.04927, 2021.

[367] Ganesh, S., N. Vadori, M. Xu, et al. Reinforcement learning for market making in a multi-agent dealer market. *CoRR*, abs/1911.05892, 2019.

[368] Tipaldi, M., R. Iervolino, P. R. Massenio. Reinforcement learning in spacecraft control applications: Advances, prospects, and challenges. *Annu. Rev. Control.*, 54:1–23, 2022.

[369] Savva, M., J. Malik, D. Parikh, et al. Habitat: A platform for embodied AI research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9338–9346. IEEE, 2019.

[370] Longpre, S., L. Hou, T. Vu, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[371] Wang, Y., Y. Kordi, S. Mishra, et al. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[372] Liang, J., W. Huang, F. Xia, et al. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 9493–9500. IEEE, 2023.

[373] Li, C., F. Xia, R. Martín-Martín, et al. HRL4IN: hierarchical reinforcement learning for interactive navigation with mobile manipulators. In L. P. Kaelbling, D. Kragic, K. Sugiura, eds., *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, vol. 100 of *Proceedings of Machine Learning Research*, pages 603–616. PMLR, 2019.

[374] Eppe, M., C. Gumbsch, M. Kerzel, et al. Hierarchical principles of embodied reinforcement learning: A review. *CoRR*, abs/2012.10147, 2020.

[375] Paul, S., A. Roy-Chowdhury, A. Cherian. AVLEN: audio-visual-language embodied navigation in 3d environments. In *NeurIPS*. 2022.

[376] Hu, B., C. Zhao, P. Zhang, et al. Enabling intelligent interactions between an agent and an LLM: A reinforcement learning approach. *CoRR*, abs/2306.03604, 2023.

[377] Chen, C., U. Jain, C. Schissler, et al. Soundspaces: Audio-visual navigation in 3d environments. In A. Vedaldi, H. Bischof, T. Brox, J. Frahm, eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, vol. 12351 of *Lecture Notes in Computer Science*, pages 17–36. Springer, 2020.

[378] Huang, R., Y. Ren, J. Liu, et al. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. In *NeurIPS*. 2022.

[379] Shah, D., B. Eysenbach, G. Kahn, et al. Ving: Learning open-world navigation with visual goals. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, pages 13215–13222. IEEE, 2021.

[380] Huang, C., O. Mees, A. Zeng, et al. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 10608–10615. IEEE, 2023.

[381] Georgakis, G., K. Schmeckpeper, K. Wanchoo, et al. Cross-modal map learning for vision and language navigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15439–15449. IEEE, 2022.

[382] Dorbala, V. S., J. F. M. Jr., D. Manocha. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *CoRR*, abs/2303.03480, 2023.

[383] Li, L. H., P. Zhang, H. Zhang, et al. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022.

[384] Gan, C., Y. Zhang, J. Wu, et al. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 9701–9707. IEEE, 2020.

[385] Brohan, A., N. Brown, J. Carbajal, et al. RT-1: robotics transformer for real-world control at scale. *CoRR*, abs/2212.06817, 2022.

[386] —. RT-2: vision-language-action models transfer web knowledge to robotic control. *CoRR*, abs/2307.15818, 2023.

[387] PrismarineJS, 2013.

[388] Gur, I., H. Furuta, A. Huang, et al. A real-world webagent with planning, long context understanding, and program synthesis. *CoRR*, abs/2307.12856, 2023.

[389] Deng, X., Y. Gu, B. Zheng, et al. Mind2web: Towards a generalist agent for the web. *CoRR*, abs/2306.06070, 2023.

[390] Furuta, H., O. Nachum, K. Lee, et al. Multimodal web navigation with instruction-finetuned foundation models. *CoRR*, abs/2305.11854, 2023.

[391] Zhou, S., F. F. Xu, H. Zhu, et al. Webarena: A realistic web environment for building autonomous agents. *CoRR*, abs/2307.13854, 2023.

[392] Yao, S., H. Chen, J. Yang, et al. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*. 2022.

[393] Kim, G., P. Baldi, S. McAleer. Language models can solve computer tasks. *CoRR*, abs/2303.17491, 2023.

[394] Zheng, L., R. Wang, B. An. Synapse: Leveraging few-shot exemplars for human-level computer control. *CoRR*, abs/2306.07863, 2023.

[395] Chen, P., C. Chang. Interact: Exploring the potentials of chatgpt as a cooperative agent. *CoRR*, abs/2308.01552, 2023.

[396] Gramopadhye, M., D. Szafir. Generating executable action plans with environmentally-aware language models. *CoRR*, abs/2210.04964, 2022.

[397] Li, H., Y. Hao, Y. Zhai, et al. The hitchhiker's guide to program analysis: A journey with large language models. *CoRR*, abs/2308.00245, 2023.

[398] Feldt, R., S. Kang, J. Yoon, et al. Towards autonomous testing agents via conversational large language models. *CoRR*, abs/2306.05152, 2023.

[399] Kang, Y., J. Kim. Chatmof: An autonomous AI system for predicting and generating metal-organic frameworks. *CoRR*, abs/2308.01423, 2023.

[400] Wang, R., P. A. Jansen, M. Côté, et al. Scienceworld: Is your agent smarter than a 5th grader? In Y. Goldberg, Z. Kozareva, Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11279–11298. Association for Computational Linguistics, 2022.

[401] Yuan, H., C. Zhang, H. Wang, et al. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. *CoRR*, abs/2303.16563, 2023.

[402] Hao, R., L. Hu, W. Qi, et al. Chatllm network: More brains, more intelligence. *CoRR*, abs/2304.12998, 2023.

[403] Mandi, Z., S. Jain, S. Song. Roco: Dialectic multi-robot collaboration with large language models. *CoRR*, abs/2307.04738, 2023.

[404] Hamilton, S. Blind judgement: Agent-based supreme court modelling with GPT. *CoRR*, abs/2301.05327, 2023.

[405] Hong, S., X. Zheng, J. Chen, et al. Metagpt: Meta programming for multi-agent collaborative framework. *CoRR*, abs/2308.00352, 2023.

[406] Wu, Q., G. Bansal, J. Zhang, et al. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155, 2023.

[407] Zhang, C., K. Yang, S. Hu, et al. Proagent: Building proactive cooperative AI with large language models. *CoRR*, abs/2308.11339, 2023.

[408] Nair, V., E. Schumacher, G. J. Tso, et al. DERA: enhancing large language model completions with dialog-enabled resolving agents. *CoRR*, abs/2303.17071, 2023.

[409] Talebirad, Y., A. Nadiri. Multi-agent collaboration: Harnessing the power of intelligent LLM agents. *CoRR*, abs/2306.03314, 2023.

[410] Chen, W., Y. Su, J. Zuo, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *CoRR*, abs/2308.10848, 2023.

[411] Shi, J., J. Zhao, Y. Wang, et al. CGMI: configurable general multi-agent interaction framework. *CoRR*, abs/2308.12503, 2023.

[412] Xiong, K., X. Ding, Y. Cao, et al. Examining the inter-consistency of large language models: An in-depth analysis via debate. *CoRR*, abs/2305.11595, 2023.

[413] Kalvakurthi, V., A. S. Varde, J. Jenq. Hey dona! can you help me with student course registration? *CoRR*, abs/2303.13548, 2023.

[414] Swan, M., T. Kido, E. Roland, et al. Math agents: Computational infrastructure, mathematical embedding, and genomics. *CoRR*, abs/2307.02502, 2023.

[415] Hsu, S.-L., R. S. Shah, P. Senthil, et al. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*, 2023.

[416] Zhang, H., J. Chen, F. Jiang, et al. Huatuogpt, towards taming language model to be a doctor. *CoRR*, abs/2305.15075, 2023.

[417] Yang, S., H. Zhao, S. Zhu, et al. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *CoRR*, abs/2308.03549, 2023.

[418] Ali, M. R., S. Z. Razavi, R. Langevin, et al. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In S. Marsella, R. Jack, H. H. Vilhjálmsson, P. Sequeira, E. S. Cross, eds., *IVA '20: ACM International Conference on Intelligent Virtual Agents, Virtual Event, Scotland, UK, October 20-22, 2020*, pages 2:1–2:8. ACM, 2020.

[419] Gao, W., X. Gao, Y. Tang. Multi-turn dialogue agent as sales' assistant in telemarketing. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pages 1–9. IEEE, 2023.

[420] Schick, T., J. A. Yu, Z. Jiang, et al. PEER: A collaborative language model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[421] Lu, B., N. Haduong, C. Lee, et al. DIALGEN: collaborative human-lm generated dialogues for improved understanding of human-human conversations. *CoRR*, abs/2307.07047, 2023.

[422] Gao, D., L. Ji, L. Zhou, et al. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *CoRR*, abs/2306.08640, 2023.

[423] Hasan, M., C. Özel, S. Potter, et al. SAPIEN: affective virtual agents powered by large language models. *CoRR*, abs/2308.03022, 2023.

[424] Liu-Thompkins, Y., S. Okazaki, H. Li. Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, 50(6):1198–1218, 2022.

[425] Bakhtin, A., D. J. Wu, A. Lerer, et al. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[426] (FAIR)†, M. F. A. R. D. T., A. Bakhtin, N. Brown, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[427] Lin, J., N. Tomlin, J. Andreas, et al. Decision-oriented dialogue for human-ai collaboration. *CoRR*, abs/2305.20076, 2023.

[428] Li, C., X. Su, C. Fan, et al. Quantifying the impact of large language models on collective opinion dynamics. *CoRR*, abs/2308.03313, 2023.

[429] Chase, H. LangChain. *URL https://github.com/hwchase17/langchain*, 2022.

[430] Reworked. Agent_GPT. *URL https://github.com/reworkd/AgentGPT*, 2023.

[431] AntonOsika. GPT Engineer. *URL https://github.com/AntonOsika/gpt-engineer*, 2023.

[432] Dambekodi, S. N., S. Frazier, P. Ammanabrolu, et al. Playing text-based games with common sense. *CoRR*, abs/2012.02757, 2020.

[433] Singh, I., G. Singh, A. Modi. Pre-trained language models as prior knowledge for playing text-based games. In P. Faliszewski, V. Mascardi, C. Pelachaud, M. E. Taylor, eds., *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, pages 1729–1731. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022.

[434] Ammanabrolu, P., J. Urbanek, M. Li, et al. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou, eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 807–833. Association for Computational Linguistics, 2021.

[435] Xu, N., S. Masling, M. Du, et al. Grounding open-domain instructions to automate web support tasks. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou, eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1022–1032. Association for Computational Linguistics, 2021.

[436] Chhikara, P., J. Zhang, F. Ilievski, et al. Knowledge-enhanced agents for interactive text games. *CoRR*, abs/2305.05091, 2023.

[437] Yang, K., A. M. Swope, A. Gu, et al. Leandojo: Theorem proving with retrieval-augmented language models. *CoRR*, abs/2306.15626, 2023.

[438] Lin, Z., H. Akin, R. Rao, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[439] Irwin, R., S. Dimitriadis, J. He, et al. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.*, 3(1):15022, 2022.

[440] Skrynnik, A., Z. Volovikova, M. Côté, et al. Learning to solve voxel building embodied tasks from pixels and natural language instructions. *CoRR*, abs/2211.00688, 2022.

[441] Amiranashvili, A., N. Dorka, W. Burgard, et al. Scaling imitation learning in minecraft. *CoRR*, abs/2007.02701, 2020.

[442] Minsky, M. *Society of mind*. Simon and Schuster, 1988.

[443] Balaji, P. G., D. Srinivasan. An introduction to multi-agent systems. *Innovations in multi-agent systems and applications-1*, pages 1–27, 2010.

[444] Finin, T. W., R. Fritzson, D. P. McKay, et al. KQML as an agent communication language. In *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM'94), Gaithersburg, Maryland, USA, November 29 - December 2, 1994*, pages 456–463. ACM, 1994.

[445] Yang, Y., J. Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

[446] Smith, A. *The wealth of nations [1776]*, vol. 11937. na, 1937.

[447] Wang, Z., S. Mao, W. Wu, et al. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *CoRR*, abs/2307.05300, 2023.

[448] Hassan, M. M., R. A. Knipper, S. K. K. Santu. Chatgpt as your personal data scientist. *CoRR*, abs/2305.13657, 2023.

[449] von Neumann, J., O. Morgenstern. *Theory of Games and Economic Behavior (60th-Anniversary Edition)*. Princeton University Press, 2007.

[450] Aziz, H. Multiagent systems: algorithmic, game-theoretic, and logical foundations by y. shoham and k. leyton-brown cambridge university press, 2008. *SIGACT News*, 41(1):34–37, 2010.

[451] Campbell, M., A. J. Hoane, F. hsiung Hsu. Deep blue. *Artif. Intell.*, 134:57–83, 2002.

[452] Silver, D., J. Schrittwieser, K. Simonyan, et al. Mastering the game of go without human knowledge. *Nat.*, 550(7676):354–359, 2017.

[453] Lewis, M., D. Yarats, Y. N. Dauphin, et al. Deal or no deal? end-to-end learning of negotiation dialogues. In M. Palmer, R. Hwa, S. Riedel, eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2443–2453. Association for Computational Linguistics, 2017.

[454] Irving, G., P. F. Christiano, D. Amodei. AI safety via debate. *CoRR*, abs/1805.00899, 2018.

[455] Kenton, Z., T. Everitt, L. Weidinger, et al. Alignment of language agents. *CoRR*, abs/2103.14659, 2021.

[456] Ngo, R. The alignment problem from a deep learning perspective. *CoRR*, abs/2209.00626, 2022.

[457] Paul, M., L. Maglaras, M. A. Ferrag, et al. Digitization of healthcare sector: A study on privacy and security concerns. *ICT Express*, 2023.

[458] Bassiri, M. A. Interactional feedback and the impact of attitude and motivation on noticing l2 form. *English Language and Literature Studies*, 1(2):61, 2011.

[459] Tellex, S., T. Kollar, S. Dickerson, et al. Approaching the symbol grounding problem with probabilistic graphical models. *AI Mag.*, 32(4):64–76, 2011.

[460] Matuszek, C., E. Herbst, L. Zettlemoyer, et al. Learning to parse natural language commands to a robot control system. In J. P. Desai, G. Dudek, O. Khatib, V. Kumar, eds., *Experimental Robotics - The 13th International Symposium on Experimental Robotics, ISER 2012, June 18-21, 2012, Québec City, Canada*, vol. 88 of *Springer Tracts in Advanced Robotics*, pages 403–415. Springer, 2012.

[461] Chaplot, D. S., K. M. Sathyendra, R. K. Pasumarthi, et al. Gated-attention architectures for task-oriented language grounding. In S. A. McIlraith, K. Q. Weinberger, eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2819–2826. AAAI Press, 2018.

[462] Li, J., A. H. Miller, S. Chopra, et al. Dialogue learning with human-in-the-loop. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[463] Iyer, S., I. Konstas, A. Cheung, et al. Learning a neural semantic parser from user feedback. In R. Barzilay, M. Kan, eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 963–973. Association for Computational Linguistics, 2017.

[464] Weston, J. Dialog-based language learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett, eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 829–837. 2016.

[465] Shuster, K., J. Xu, M. Komeili, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *CoRR*, abs/2208.03188, 2022.

[466] Du, W., Z. M. Kim, V. Raheja, et al. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. *CoRR*, abs/2204.03685, 2022.

[467] Kreutzer, J., S. Khadivi, E. Matusov, et al. Can neural machine translation be improved with user feedback? In S. Bangalore, J. Chu-Carroll, Y. Li, eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 3 (Industry Papers)*, pages 92–105. Association for Computational Linguistics, 2018.

[468] Gur, I., S. Yavuz, Y. Su, et al. Dialsql: Dialogue based structured query generation. In I. Gurevych, Y. Miyao, eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1339–1349. Association for Computational Linguistics, 2018.

[469] Yao, Z., Y. Su, H. Sun, et al. Model-based interactive semantic parsing: A unified framework and A text-to-sql case study. In K. Inui, J. Jiang, V. Ng, X. Wan, eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5446–5457. Association for Computational Linguistics, 2019.

[470] Mehta, N., D. Goldwasser. Improving natural language interaction with robots using advice. In J. Burstein, C. Doran, T. Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1962–1967. Association for Computational Linguistics, 2019.

[471] Elgohary, A., C. Meek, M. Richardson, et al. NL-EDIT: correcting semantic parse errors through natural language interaction. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou, eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5599–5610. Association for Computational Linguistics, 2021.

[472] Tandon, N., A. Madaan, P. Clark, et al. Learning to repair: Repairing model output errors after deployment using a dynamic memory of feedback. In M. Carpuat, M. de Marneffe, I. V. M. Ruíz, eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 339–352. Association for Computational Linguistics, 2022.

[473] Scheurer, J., J. A. Campos, T. Korbak, et al. Training language models with language feedback at scale. *CoRR*, abs/2303.16755, 2023.

[474] Xu, J., M. Ung, M. Komeili, et al. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13557–13572. Association for Computational Linguistics, 2023.

[475] Cai, Z., B. Chang, W. Han. Human-in-the-loop through chain-of-thought. *CoRR*, abs/2306.07932, 2023.

[476] Hancock, B., A. Bordes, P. Mazaré, et al. Learning from dialogue after deployment: Feed yourself, chatbot! In A. Korhonen, D. R. Traum, L. Màrquez, eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3667–3684. Association for Computational Linguistics, 2019.

[477] Mehta, N., M. Teruel, P. F. Sanz, et al. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. *CoRR*, abs/2304.10750, 2023.

[478] Gvirsman, O., Y. Koren, T. Norman, et al. Patricc: A platform for triadic interaction with changeable characters. In T. Belpaeme, J. E. Young, H. Gunes, L. D. Riek, eds., *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23-26, 2020*, pages 399–407. ACM, 2020.

[479] Stiles-Shields, C., E. Montague, E. G. Lattie, et al. What might get in the way: Barriers to the use of apps for depression. *DIGITAL HEALTH*, 3:2055207617713827, 2017. PMID: 29942605.

[480] McTear, M. F. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2020.

[481] Motger, Q., X. Franch, J. Marco. Conversational agents in software engineering: Survey, taxonomy and challenges. *CoRR*, abs/2106.10901, 2021.

[482] Rapp, A., L. Curti, A. Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum. Comput. Stud.*, 151:102630, 2021.

[483] Adamopoulou, E., L. Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

[484] Wang, K., X. Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In J. Lang, ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4446–4452. ijcai.org, 2018.

[485] Zhou, X., W. Y. Wang. Mojitalk: Generating emotional responses at scale. In I. Gurevych, Y. Miyao, eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1128–1137. Association for Computational Linguistics, 2018.

[486] Lin, Z., P. Xu, G. I. Winata, et al. Caire: An empathetic neural chatbot. *arXiv preprint arXiv:1907.12108*, 2019.

[487] Jhan, J., C. Liu, S. Jeng, et al. Cheerbots: Chatbots toward empathy and emotionusing reinforcement learning. *CoRR*, abs/2110.03949, 2021.

[488] Lin, Z., A. Madotto, J. Shin, et al. Moel: Mixture of empathetic listeners. In K. Inui, J. Jiang, V. Ng, X. Wan, eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics, 2019.

[489] Majumder, N., P. Hong, S. Peng, et al. MIME: mimicking emotions for empathetic response generation. In B. Webber, T. Cohn, Y. He, Y. Liu, eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics, 2020.

[490] Sabour, S., C. Zheng, M. Huang. CEM: commonsense-aware empathetic response generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11229–11237. AAAI Press, 2022.

[491] Li, Q., P. Li, Z. Ren, et al. Knowledge bridging for empathetic dialogue generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10993–11001. AAAI Press, 2022.

[492] Liu, B., S. S. Sundar. Should machines express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology Behav. Soc. Netw.*, 21(10):625–636, 2018.

[493] Su, Z., M. C. Figueiredo, J. Jo, et al. Analyzing description, user understanding and expectations of AI in mobile health applications. In *AMIA 2020, American Medical Informatics Association Annual Symposium, Virtual Event, USA, November 14-18, 2020*. AMIA, 2020.

[494] Moravcík, M., M. Schmid, N. Burch, et al. Deepstack: Expert-level artificial intelligence in no-limit poker. *CoRR*, abs/1701.01724, 2017.

[495] Carroll, M., R. Shah, M. K. Ho, et al. On the utility of learning about humans for human-ai coordination. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett, eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5175–5186. 2019.

[496] Bard, N., J. N. Foerster, S. Chandar, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

[497] Wang, X., W. Shi, R. Kim, et al. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In A. Korhonen, D. R. Traum, L. Màrquez, eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5635–5649. Association for Computational Linguistics, 2019.

[498] Abrams, A. M. H., A. M. R. der Pütten. I-C-E framework: Concepts for group dynamics research in human-robot interaction. *Int. J. Soc. Robotics*, 12(6):1213–1229, 2020.

[499] Xu, Y., S. Wang, P. Li, et al. Exploring large language models for communication games: An empirical study on werewolf, 2023.

[500] Binz, M., E. Schulz. Using cognitive psychology to understand GPT-3. *CoRR*, abs/2206.14576, 2022.

[501] Dasgupta, I., A. K. Lampinen, S. C. Y. Chan, et al. Language models show human-like content effects on reasoning. *CoRR*, abs/2207.07051, 2022.

[502] Dhingra, S., M. Singh, V. S. B, et al. Mind meets machine: Unravelling gpt-4's cognitive psychology. *CoRR*, abs/2303.11436, 2023.

[503] Hagendorff, T. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *CoRR*, abs/2303.13988, 2023.

[504] Wang, X., X. Li, Z. Yin, et al. Emotional intelligence of large language models. *CoRR*, abs/2307.09042, 2023.

[505] Curry, A., A. C. Curry. Computer says "no": The case against empathetic conversational AI. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8123–8130. Association for Computational Linguistics, 2023.

[506] Elyoseph, Z., D. Hadar-Shoval, K. Asraf, et al. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14:1199058, 2023.

[507] Habibi, R., J. Pfau, J. Holmes, et al. Empathetic AI for empowering resilience in games. *CoRR*, abs/2302.09070, 2023.

[508] Caron, G., S. Srivastava. Identifying and manipulating the personality traits of language models. *CoRR*, abs/2212.10276, 2022.

[509] Pan, K., Y. Zeng. Do llms possess a personality? making the MBTI test an amazing evaluation for large language models. *CoRR*, abs/2307.16180, 2023.

[510] Li, X., Y. Li, S. Joty, et al. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective, 2023.

[511] Safdari, M., G. Serapio-García, C. Crepy, et al. Personality traits in large language models. *CoRR*, abs/2307.00184, 2023.

[512] Côté, M., Á. Kádár, X. Yuan, et al. Textworld: A learning environment for text-based games. In T. Cazenave, A. Saffidine, N. R. Sturtevant, eds., *Computer Games - 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers*, vol. 1017 of *Communications in Computer and Information Science*, pages 41–75. Springer, 2018.

[513] Urbanek, J., A. Fan, S. Karamcheti, et al. Learning to speak and act in a fantasy text adventure game. In K. Inui, J. Jiang, V. Ng, X. Wan, eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 673–683. Association for Computational Linguistics, 2019.

[514] Hausknecht, M. J., P. Ammanabrolu, M. Côté, et al. Interactive fiction games: A colossal adventure. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7903–7910. AAAI Press, 2020.

[515] O'Gara, A. Hoodwinked: Deception and cooperation in a text-based game for language models. *CoRR*, abs/2308.01404, 2023.

[516] Bharadhwaj, H., J. Vakil, M. Sharma, et al. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. *CoRR*, abs/2309.01918, 2023.

[517] Park, J. S., L. Popowski, C. J. Cai, et al. Social simulacra: Creating populated prototypes for social computing systems. In M. Agrawala, J. O. Wobbrock, E. Adar, V. Setlur, eds., *The 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022 - 2 November 2022*, pages 74:1–74:18. ACM, 2022.

[518] Gao, C., X. Lan, Z. Lu, et al. S^3: Social-network simulation system with large language model-empowered agents. *CoRR*, abs/2307.14984, 2023.

[519] Wang, L., J. Zhang, X. Chen, et al. Recagent: A novel simulation paradigm for recommender systems. *CoRR*, abs/2306.02552, 2023.

[520] Williams, R., N. Hosseinichimeh, A. Majumdar, et al. Epidemic modeling with generative agents. *CoRR*, abs/2307.04986, 2023.

[521] da Rocha Costa, A. C. *A Variational Basis for the Regulation and Structuration Mechanisms of Agent Societies*. Springer, 2019.

[522] Wimmer, S., A. Pfeiffer, N. Denk. The everyday life in the sims 4 during a pandemic. a life simulation as a virtual mirror of society? In *INTED2021 Proceedings*, 15th International Technology, Education and Development Conference, pages 5754–5760. IATED, 2021.

[523] Lee, L., T. Braud, P. Zhou, et al. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *CoRR*, abs/2110.05352, 2021.

[524] Inkeles, A., D. H. Smith. *Becoming modern: Individual change in six developing countries*. Harvard University Press, 1974.

[525] Troitzsch, K. G., U. Mueller, G. N. Gilbert, et al., eds. *Social Science Microsimulation [Dagstuhl Seminar, May, 1995]*. Springer, 1996.

[526] Abrams, A. M., A. M. R.-v. der Pütten. I–c–e framework: Concepts for group dynamics research in human-robot interaction: Revisiting theory from social psychology on ingroup identification (i), cohesion (c) and entitativity (e). *International Journal of Social Robotics*, 12:1213–1229, 2020.

[527] Askell, A., Y. Bai, A. Chen, et al. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861, 2021.

[528] Zhang, Z., N. Liu, S. Qi, et al. Heterogeneous value evaluation for large language models. *CoRR*, abs/2305.17147, 2023.

[529] Browning, J. Personhood and ai: Why large language models don't understand us. *AI & SOCIETY*, pages 1–8, 2023.

[530] Jiang, G., M. Xu, S. Zhu, et al. MPI: evaluating and inducing personality in pre-trained language models. *CoRR*, abs/2206.07550, 2022.

[531] Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083, 2023.

[532] Zuckerman, M. *Psychobiology of personality*, vol. 10. Cambridge University Press, 1991.

[533] Han, S. J., K. Ransom, A. Perfors, et al. Inductive reasoning in humans and large language models. *CoRR*, abs/2306.06548, 2023.

[534] Hagendorff, T., S. Fabi, M. Kosinski. Thinking fast and slow in large language models, 2023.

[535] Hagendorff, T., S. Fabi. Human-like intuitive behavior and reasoning biases emerged in language models - and disappeared in GPT-4. *CoRR*, abs/2306.07622, 2023.

[536] Ma, Z., Y. Mei, Z. Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *CoRR*, abs/2307.15810, 2023.

[537] Bates, J. The role of emotion in believable agents. *Commun. ACM*, 37(7):122–125, 1994.

[538] Karra, S. R., S. Nguyen, T. Tulabandhula. AI personification: Estimating the personality of language models. *CoRR*, abs/2204.12000, 2022.

[539] Zhang, S., E. Dinan, J. Urbanek, et al. Personalizing dialogue agents: I have a dog, do you have pets too? In I. Gurevych, Y. Miyao, eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2204–2213. Association for Computational Linguistics, 2018.

[540] Kwon, D. S., S. Lee, K. H. Kim, et al. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In S. Sitaram, B. B. Klebanov, J. D. Williams, eds., *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 707–719. Association for Computational Linguistics, 2023.

[541] Maes, P. Artificial life meets entertainment: Lifelike autonomous agents. *Commun. ACM*, 38(11):108–114, 1995.

[542] Grossmann, I., M. Feinberg, D. C. Parker, et al. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109, 2023.

[543] Wei, J., K. Shuster, A. Szlam, et al. Multi-party chat: Conversational agents in group settings with humans and models. *CoRR*, abs/2304.13835, 2023.

[544] Hollan, J. D., E. L. Hutchins, L. Weitzman. STEAMER: an interactive inspectable simulation-based training system. *AI Mag.*, 5(2):15–27, 1984.

[545] Tambe, M., W. L. Johnson, R. M. Jones, et al. Intelligent agents for interactive simulation environments. *AI Mag.*, 16(1):15–39, 1995.

[546] Vermeulen, P., D. de Jongh. 'dynamics of growth in a finite world' – comprehensive sensitivity analysis. *IFAC Proceedings Volumes*, 9(3):133–145, 1976. IFAC Symposium on Large Scale Systems Theory and Applications, Milano, Italy, 16-20 June.

[547] Forrester, J. W. System dynamics and the lessons of 35 years. In *A systems-based approach to policymaking*, pages 199–240. Springer, 1993.

[548] Santé, I., A. M. García, D. Miranda, et al. Cellular automata models for the simulation of real-world urban processes: A review and analysis. *Landscape and urban planning*, 96(2):108–122, 2010.

[549] Dorri, A., S. S. Kanhere, R. Jurdak. Multi-agent systems: A survey. *Ieee Access*, 6:28573–28593, 2018.

[550] Hendrickx, J. M., S. Martin. Open multi-agent systems: Gossiping with random arrivals and departures. In *56th IEEE Annual Conference on Decision and Control, CDC 2017, Melbourne, Australia, December 12-15, 2017*, pages 763–768. IEEE, 2017.

[551] Ziems, C., W. Held, O. Shaikh, et al. Can large language models transform computational social science? *CoRR*, abs/2305.03514, 2023.

[552] Gilbert, N., J. Doran. *Simulating Societies: The Computer Simulation of Social Phenomena*. Routledge Library Editions: Artificial Intelligence. Taylor & Francis, 2018.

[553] Hamilton, J. D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384, 1989.

[554] Zhang, G. P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.

[555] Kirby, S., M. Dowman, T. L. Griffiths. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245, 2007.

[556] Shibata, H., S. Miki, Y. Nakamura. Playing the werewolf game with artificial intelligence for language understanding. *CoRR*, abs/2302.10646, 2023.

[557] Junprung, E. Exploring the intersection of large language models and agent-based modeling via prompt engineering. *CoRR*, abs/2308.07411, 2023.

[558] Phelps, S., Y. I. Russell. Investigating emergent goal-like behaviour in large language models using experimental economics. *CoRR*, abs/2305.07970, 2023.

[559] Bellomo, N., G. A. Marsan, A. Tosin. *Complex systems and society: modeling and simulation*, vol. 2. Springer, 2013.

[560] Moon, Y. B. Simulation modelling for sustainability: a review of the literature. *International Journal of Sustainable Engineering*, 10(1):2–19, 2017.

[561] Helberger, N., N. Diakopoulos. Chatgpt and the AI act. *Internet Policy Rev.*, 12(1), 2023.

[562] Weidinger, L., J. Mellor, M. Rauh, et al. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021.

[563] Deshpande, A., V. Murahari, T. Rajpurohit, et al. Toxicity in chatgpt: Analyzing persona-assigned language models. *CoRR*, abs/2304.05335, 2023.

[564] Kirk, H. R., Y. Jun, F. Volpin, et al. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan, eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2611–2624. 2021.

[565] Nadeem, M., A. Bethke, S. Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, R. Navigli, eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5356–5371. Association for Computational Linguistics, 2021.

[566] Roberts, T., G. Marchais. Assessing the role of social media and digital technology in violence reporting. *Contemporary Readings in Law & Social Justice*, 10(2), 2018.

[567] Kandpal, N., H. Deng, A. Roberts, et al. Large language models struggle to learn long-tail knowledge. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR, 2023.

[568] Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models. *CoRR*, abs/2304.03738, 2023.

[569] Haller, P., A. Aynetdinov, A. Akbik. Opiniongpt: Modelling explicit biases in instruction-tuned llms, 2023.

[570] Salewski, L., S. Alaniz, I. Rio-Torto, et al. In-context impersonation reveals large language models' strengths and biases. *CoRR*, abs/2305.14930, 2023.

[571] Lin, B., D. Bouneffouf, G. A. Cecchi, et al. Towards healthy AI: large language models need therapists too. *CoRR*, abs/2304.00416, 2023.

[572] Liang, P. P., C. Wu, L. Morency, et al. Towards understanding and mitigating social biases in language models. In M. Meila, T. Zhang, eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR, 2021.

[573] Henderson, P., K. Sinha, N. Angelard-Gontier, et al. Ethical challenges in data-driven dialogue systems. In J. Furman, G. E. Marchant, H. Price, F. Rossi, eds., *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 123–129. ACM, 2018.

[574] Li, H., Y. Song, L. Fan. You don't know my favorite color: Preventing dialogue representations from revealing speakers' private personas. In M. Carpuat, M. de Marneffe, I. V. M. Ruíz, eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5858–5870. Association for Computational Linguistics, 2022.

[575] Brown, H., K. Lee, F. Mireshghallah, et al. What does it mean for a language model to preserve privacy? In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2280–2292. ACM, 2022.

[576] Sebastian, G. Privacy and data protection in chatgpt and other ai chatbots: Strategies for securing user information. *Available at SSRN 4454761*, 2023.

[577] Reeves, B., C. Nass. *The media equation - how people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.

[578] Roose, K. A conversation with bing's chatbot left me deeply unsettled, 2023.

[579] Li, K., A. K. Hopkins, D. Bau, et al. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[580] Bai, Y., S. Kadavath, S. Kundu, et al. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022.

[581] Bai, Y., A. Jones, K. Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022.

[582] Liu, X., H. Yu, H. Zhang, et al. Agentbench: Evaluating llms as agents. *CoRR*, abs/2308.03688, 2023.

[583] Aher, G. V., R. I. Arriaga, A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR, 2023.

[584] Liang, Y., L. Zhu, Y. Yang. Tachikuma: Understading complex interactions with multi-character and novel objects by large language models. *CoRR*, abs/2307.12573, 2023.

[585] Xu, B., X. Liu, H. Shen, et al. Gentopia: A collaborative platform for tool-augmented llms. *CoRR*, abs/2308.04030, 2023.

[586] Kim, S. S., E. A. Watkins, O. Russakovsky, et al. " help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17. 2023.

[587] Choi, M., J. Pei, S. Kumar, et al. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *CoRR*, abs/2305.14938, 2023.

[588] Wilson, A. C., D. V. Bishop. " if you catch my drift...": ability to infer implied meaning is distinct from vocabulary and grammar skills. *Wellcome open research*, 4, 2019.

[589] Shuster, K., J. Urbanek, A. Szlam, et al. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In M. Carpuat, M. de Marneffe, I. V. M. Ruíz, eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2367–2387. Association for Computational Linguistics, 2022.

[590] Ganguli, D., L. Lovitt, J. Kernion, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858, 2022.

[591] Kadavath, S., T. Conerly, A. Askell, et al. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022.

[592] Colas, C., L. Teodorescu, P. Oudeyer, et al. Augmenting autotelic agents with large language models. *CoRR*, abs/2305.12487, 2023.

[593] Chaudhry, A., P. K. Dokania, T. Ajanthan, et al. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547. 2018.

[594] Hou, S., X. Pan, C. C. Loy, et al. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839. 2019.

[595] Colas, C., T. Karch, O. Sigaud, et al. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: A short survey. *J. Artif. Intell. Res.*, 74:1159–1199, 2022.

[596] Szegedy, C., W. Zaremba, I. Sutskever, et al. Intriguing properties of neural networks. In Y. Bengio, Y. LeCun, eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. 2014.

[597] Goodfellow, I. J., J. Shlens, C. Szegedy. Explaining and harnessing adversarial examples. In Y. Bengio, Y. LeCun, eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015.

[598] Madry, A., A. Makelov, L. Schmidt, et al. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[599] Zheng, R., Z. Xi, Q. Liu, et al. Characterizing the impacts of instances on robustness. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2314–2332. Association for Computational Linguistics, 2023.

[600] Zhiheng, X., Z. Rui, G. Tao. Safety and ethical concerns of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts)*, pages 9–16. 2023.

[601] Akhtar, N., A. Mian, N. Kardan, et al. Threat of adversarial attacks on deep learning in computer vision: Survey II. *CoRR*, abs/2108.00401, 2021.

[602] Drenkow, N., N. Sani, I. Shpitser, et al. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*, 2021.

[603] Hendrycks, D., T. G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[604] Wang, X., H. Wang, D. Yang. Measure and improve robustness in NLP models: A survey. In M. Carpuat, M. de Marneffe, I. V. M. Ruíz, eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4569–4586. Association for Computational Linguistics, 2022.

[605] Li, J., S. Ji, T. Du, et al. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019.

[606] Zhu, C., Y. Cheng, Z. Gan, et al. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[607] Xi, Z., R. Zheng, T. Gui, et al. Efficient adversarial training with robust early-bird tickets. In Y. Goldberg, Z. Kozareva, Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8318–8331. Association for Computational Linguistics, 2022.

[608] Pinto, L., J. Davidson, R. Sukthankar, et al. Robust adversarial reinforcement learning. In D. Precup, Y. W. Teh, eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, vol. 70 of *Proceedings of Machine Learning Research*, pages 2817–2826. PMLR, 2017.

[609] Rigter, M., B. Lacerda, N. Hawes. RAMBO-RL: robust adversarial model-based offline reinforcement learning. In *NeurIPS*. 2022.

[610] Panaganti, K., Z. Xu, D. Kalathil, et al. Robust reinforcement learning using offline data. In *NeurIPS*. 2022.

[611] Lab, T. K. S. Experimental security research of tesla autopilot. *Tencent Keen Security Lab*, 2019.

[612] Xu, K., G. Zhang, S. Liu, et al. Adversarial t-shirt! evading person detectors in a physical world. In A. Vedaldi, H. Bischof, T. Brox, J. Frahm, eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, vol. 12350 of *Lecture Notes in Computer Science*, pages 665–681. Springer, 2020.

[613] Sharif, M., S. Bhagavatula, L. Bauer, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, S. Halevi, eds., *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540. ACM, 2016.

[614] Jin, D., Z. Jin, J. T. Zhou, et al. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press, 2020.

[615] Ren, S., Y. Deng, K. He, et al. Generating natural language adversarial examples through probability weighted word saliency. In A. Korhonen, D. R. Traum, L. Màrquez, eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1085–1097. Association for Computational Linguistics, 2019.

[616] Zhu, K., J. Wang, J. Zhou, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *CoRR*, abs/2306.04528, 2023.

[617] Chen, X., J. Ye, C. Zu, et al. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. *CoRR*, abs/2303.00293, 2023.

[618] Gu, T., B. Dolan-Gavitt, S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[619] Chen, X., A. Salem, D. Chen, et al. Badnl: Backdoor attacks against NLP models with semantic-preserving improvements. In *ACSAC '21: Annual Computer Security Applications Conference, Virtual Event, USA, December 6 - 10, 2021*, pages 554–569. ACM, 2021.

[620] Li, Z., D. Mekala, C. Dong, et al. Bfclass: A backdoor-free text classification framework. In M. Moens, X. Huang, L. Specia, S. W. Yih, eds., *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 444–453. Association for Computational Linguistics, 2021.

[621] Shi, Y., P. Li, C. Yin, et al. Promptattack: Prompt-based attack for language models via gradient search. In W. Lu, S. Huang, Y. Hong, X. Zhou, eds., *Natural Language Processing and Chinese Computing - 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24-25, 2022, Proceedings, Part I*, vol. 13551 of *Lecture Notes in Computer Science*, pages 682–693. Springer, 2022.

[622] Perez, F., I. Ribeiro. Ignore previous prompt: Attack techniques for language models. *CoRR*, abs/2211.09527, 2022.

[623] Liang, P., R. Bommasani, T. Lee, et al. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022.

[624] Gururangan, S., D. Card, S. K. Dreier, et al. Whose language counts as high quality? measuring language ideologies in text data selection. In Y. Goldberg, Z. Kozareva, Y. Zhang, eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2562–2580. Association for Computational Linguistics, 2022.

[625] Liu, Y., G. Deng, Y. Li, et al. Prompt injection attack against llm-integrated applications. *CoRR*, abs/2306.05499, 2023.

[626] Carlini, N., D. A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 1–7. IEEE Computer Society, 2018.

[627] Morris, J. X., E. Lifland, J. Y. Yoo, et al. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In Q. Liu, D. Schlangen, eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 119–126. Association for Computational Linguistics, 2020.

[628] Si, C., Z. Zhang, F. Qi, et al. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In C. Zong, F. Xia, W. Li, R. Navigli, eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, vol. ACL/IJCNLP 2021 of *Findings of ACL*, pages 1569–1576. Association for Computational Linguistics, 2021.

[629] Yoo, K., J. Kim, J. Jang, et al. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In S. Muresan, P. Nakov, A. Villavicencio, eds., *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3656–3672. Association for Computational Linguistics, 2022.

[630] Le, T., N. Park, D. Lee. A sweet rabbit hole by DARCY: using honeypots to detect universal trigger's adversarial attacks. In C. Zong, F. Xia, W. Li, R. Navigli, eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3831–3844. Association for Computational Linguistics, 2021.

[631] Tsipras, D., S. Santurkar, L. Engstrom, et al. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[632] Zhang, H., Y. Yu, J. Jiao, et al. Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri, R. Salakhutdinov, eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, vol. 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.

[633] Wong, A., X. Y. Wang, A. Hryniowski. How much can we really trust you? towards simple, interpretable trust quantification metrics for deep neural networks. *CoRR*, abs/2009.05835, 2020.

[634] Huang, X., D. Kroening, W. Ruan, et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.*, 37:100270, 2020.

[635] Huang, X., W. Ruan, W. Huang, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *CoRR*, abs/2305.11391, 2023.

[636] Raffel, C., N. Shazeer, A. Roberts, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[637] Chen, Y., L. Yuan, G. Cui, et al. A close look into the calibration of pre-trained language models. In A. Rogers, J. L. Boyd-Graber, N. Okazaki, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1343–1367. Association for Computational Linguistics, 2023.

[638] Blodgett, S. L., S. Barocas, H. D. III, et al. Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics, 2020.

[639] Guo, W., A. Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In M. Fourcade, B. Kuipers, S. Lazar, D. K. Mulligan, eds., *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 122–133. ACM, 2021.

[640] Bolukbasi, T., K. Chang, J. Y. Zou, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett, eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357. 2016.

[641] Caliskan, A., J. J. Bryson, A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[642] Ji, Z., N. Lee, R. Frieske, et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023.

[643] Mündler, N., J. He, S. Jenko, et al. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *CoRR*, abs/2305.15852, 2023.

[644] Maynez, J., S. Narayan, B. Bohnet, et al. On faithfulness and factuality in abstractive summarization. In D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics, 2020.

[645] Varshney, N., W. Yao, H. Zhang, et al. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *CoRR*, abs/2307.03987, 2023.

[646] Lightman, H., V. Kosaraju, Y. Burda, et al. Let's verify step by step. *CoRR*, abs/2305.20050, 2023.

[647] Guo, Y., Y. Yang, A. Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In S. Muresan, P. Nakov, A. Villavicencio, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1012–1023. Association for Computational Linguistics, 2022.

[648] Du, M., F. He, N. Zou, et al. Shortcut learning of large language models in natural language understanding: A survey. *CoRR*, abs/2208.11857, 2022.

[649] Brundage, M., S. Avin, J. Clark, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *CoRR*, abs/1802.07228, 2018.

[650] Bommasani, R., D. A. Hudson, E. Adeli, et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.

[651] Charan, P. V. S., H. Chunduri, P. M. Anand, et al. From text to MITRE techniques: Exploring the malicious use of large language models for generating cyber attack payloads. *CoRR*, abs/2305.15336, 2023.

[652] Wang, Z. J., D. Choi, S. Xu, et al. Putting humans in the natural language processing loop: A survey. *CoRR*, abs/2103.04044, 2021.

[653] Galsworthy, J. *The inn of tranquillity: studies and essays*. W. Heinemann, 1912.

[654] Yao, S., K. Narasimhan. Language agents in the digital world: Opportunities and risks. *princeton-nlp.github.io*, 2023.

[655] Asimov, I. Three laws of robotics. *Asimov, I. Runaround*, 2, 1941.

[656] Elhage, N., N. Nanda, C. Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021.

[657] Bai, J., S. Zhang, Z. Chen. Is there any social principle for llm-based agents? *CoRR*, abs/2308.11136, 2023.

[658] Baum, S. A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, pages 17–1, 2017.

[659] Lecun, Y. https://twitter.com/ylecun/status/1625127902890151943.

[660] Zhao, S. Can Large Language Models Lead to Artificial General Intelligence?

[661] Brandes, N. Language Models are a Potentially Safe Path to Human-Level AGI.

[662] Zocca, V. How far are we from AGI?

[663] Ilya Sutskever, L. F. Ilya Sutskever: Deep Learning | Lex Fridman Podcast #94.

[664] Lecun, Y. https://twitter.com/ylecun/status/1640063227903213568.

[665] LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62, 2022.

[666] Shridhar, M., X. Yuan, M. Côté, et al. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[667] Chowdhury, J. R., C. Caragea. Monotonic location attention for length generalization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202 of *Proceedings of Machine Learning Research*, pages 28792–28808. PMLR, 2023.

[668] Duan, Y., G. Fu, N. Zhou, et al. Everything as a service (xaas) on the cloud: Origins, current and future trends. In C. Pu, A. Mohindra, eds., *8th IEEE International Conference on Cloud Computing, CLOUD 2015, New York City, NY, USA, June 27 - July 2, 2015*, pages 621–628. IEEE Computer Society, 2015.

[669] Bhardwaj, S., L. Jain, S. Jain. Cloud computing: A study of infrastructure as a service (iaas). *International Journal of engineering and information Technology*, 2(1):60–63, 2010.

[670] Serrano, N., G. Gallardo, J. Hernantes. Infrastructure as a service and cloud technologies. *IEEE Software*, 32(2):30–36, 2015.

[671] Mell, P., T. Grance, et al. The nist definition of cloud computing, 2011.

[672] Lawton, G. Developing software online with platform-as-a-service technology. *Computer*, 41(6):13–15, 2008.

[673] Sun, W., K. Zhang, S.-K. Chen, et al. Software as a service: An integration perspective. In *Service-Oriented Computing–ICSOC 2007: Fifth International Conference, Vienna, Austria, September 17-20, 2007. Proceedings 5*, pages 558–569. Springer, 2007.

[674] Dubey, A., D. Wagle. Delivering software as a service. *The McKinsey Quarterly*, 6(2007):2007, 2007.

[675] Sun, T., Y. Shao, H. Qian, et al. Black-box tuning for language-model-as-a-service. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato, eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, vol. 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR, 2022.