A Survey on the Application of LLM-based Multi-Agent System

Anonymous ACL submission

Abstract

Multi-agent systems (MAS) have become a research hotspot since the rise of large language models (LLMs). However, current review papers lack a thorough examination of the diverse applications of LLM-based multi-agent systems (LLM-MAS). This paper presents a comprehensive survey of applications of LLM-MAS. We provide an overview of the various applications of LLM-MAS in (i) solving complex tasks, (ii) simulating specific scenarios, and (iii) evaluating generative agents. Also, we highlight several challenges and propose future directions for research in this field.

1 Introduction

Multi-agent systems (MAS) have seen significant expansion owing to their adaptability and ability to address complex, distributed challenges (Balaji and Srinivasan, 2010). Compared to single-agent settings, MAS provide a more accurate representation of the real world, as many real-world applications naturally involve multiple decision-makers interacting simultaneously (Gronauer and Diepold, 2022). Previous research on MAS has predominantly focused on reinforcement learning (RL)-based agents, as illustrated by their application to classic tasks ranging from Atari video games (Mnih, 2013) to robotic socket-insertion challenges (Brockman, 2016), trained in specific environments. However, due to limitations in their parameterization and a lack of general knowledge, these agents struggle to take informed agent actions in unconstrained, open-domain scenarios requiring general knowledge.

Compared to RL-based MAS, LLM-based multi-agent systems (LLM-MAS) demonstrate the ability to handle a wide range of tasks in open-domain environments (Shinn et al., 2023). By leveraging the generalization capabilities and linguistic modality of LLMs, LLM-MAS enable novel applications that are not achievable with RL-based MAS, spanning domains from healthcare (Tang et al., 2024a) to embodied AI (Patel et al., 2024). In recent years, numerous studies have explored the diverse applications of LLM-MAS. However, a comprehensive review of LLM-MAS applications is still lacking.

In this paper, we provide a comprehensive perspective on the application of LLM-based multi-agent systems (LLM-MAS). Figure 1 presents an overview of applications of LLM-MAS. There are three categories of applications of LLM-MAS: (i) Solving complex tasks. LLM-MAS perform a wide range of tasks, including simple tasks that do not require long trajectory decisions, complex tasks that involve long trajectory decisions, and even some general-purpose tasks. (ii) Simulating for specific scenarios. LLM-MAS simulate diverse scenarios, facilitating the exploration and validation of relevant theories. (iii) Evaluating and Training on generative agents. On the one hand, compared with traditional evaluation on agents, LLM-MAS have the capability of dynamic assessment, which is more flexible and harder for data leakage (Chen et al., 2024c). On the other hand, agents can be trained in LLM-MAS, concluding various training methods.

Compared to previous surveys (Guo et al., 2024a; Li et al., 2024d; Han et al., 2024; Gronauer and Diepold, 2022), this survey offers the following key contributions: (i) A clear taxonomy for LLM-MAS applications. We present a framework to organize and categorize different types of LLM-MAS applications. (ii) A definition of the environment in LLM-MAS applications. We provide a specific definition of the LLM-MAS environment, designed to fit the needs of LLM-MAS applications. (iii) A summary of available resources for LLM-MAS research. We compile a list of open-source frameworks and datasets to help researchers study LLM-MAS applications. (iv) Challenges and future directions for LLM-MAS applications. We discuss the current challenges in the field and suggest potential areas for future research.

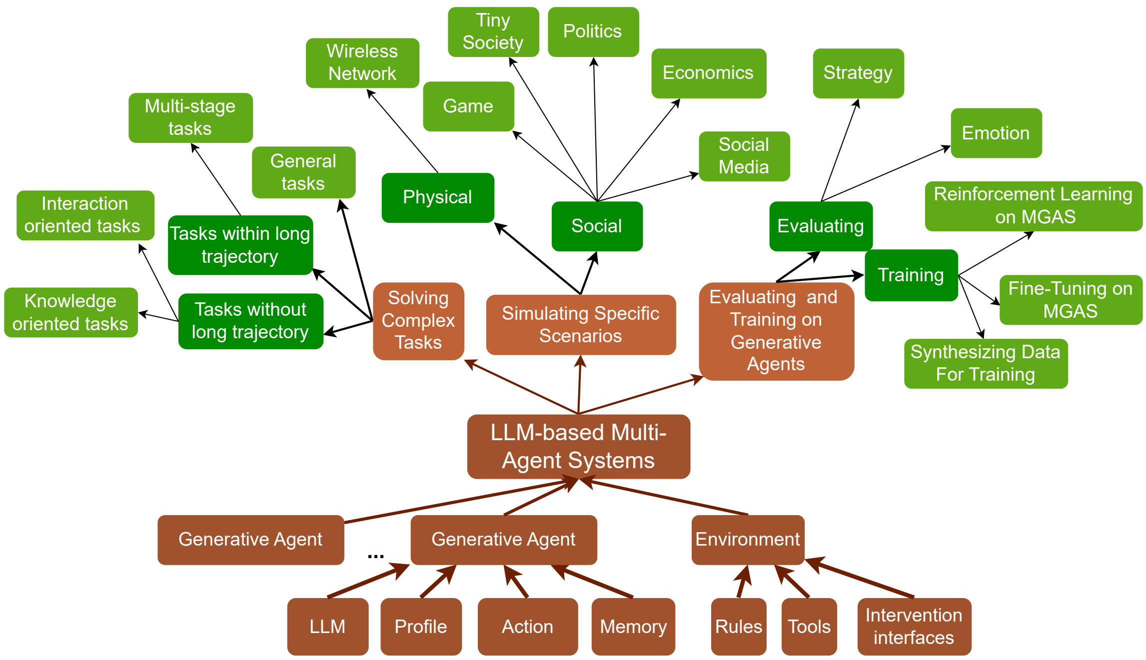

Figure 1: Overview of the application and construction of LLM-MAS.

2 Core Components of LLM-MAS

LLM-MAS refer to systems that include a collection of generative agents capable of interacting and collaborating within a shared environmental setting (Wang et al., 2024c). We will analyze generative agents and the environment in the following.

2.1 Generative Agents

Generative agents refer to the components of LLM-MAS that have role definitions, can perceive the environment, make decisions, and perform complex actions to interact with the environment (Wang et al., 2024a).

Compared to traditional agents, generative agents can be able to perform complex behaviors, such as generating complete personalized blog posts based on historical information (Park et al., 2022). Therefore, in addition to using LLMs as the core, generative agents also require the following characteristics: (i) *Profiling* refers to agents typically assuming distinct roles, each accompanied by detailed descriptions that encompass their characteristics, capabilities, and constraints(Guo et al., 2024a). (ii) *Memory* stores historical trajectories and retrieves relevant memories for subsequent agent actions, enabling the ability to take long-term actions while solving the problem of limited LLM context windows. There usually are three memory layers: long-term, short-term, and sensory memory (Park et al., 2023). (iii) *Planning* is to formulate general behavior for a longer period in the future (Yao et al., 2023). (iv) *Action* executes the interaction between the generative agent and the environment (Wang et al., 2024a). Generative agents are required to choose one of several candidate behaviors to execute, such as voting for whom (Xu et al., 2024a), or generate behaviors without mandatory constraints, such as generating a paragraph of text (Li et al., 2023b).

Generative agents can communicate with each other to achieve cooperation within the system. The communication of generative agents can be roughly divided into two purposes. (i) The first purpose is to achieve collaboration, share the information obtained by themselves with other intelligent agents, and to some extent, aggregate multiple intelligent agents into a complete system, achieving performance beyond independent intelligent agents (Yuan et al., 2023); (ii) The second purpose is to achieve consensus, allowing for greater similarity in behavior or strategy among some agents, thereby enabling faster convergence to Nash equilibrium (Oroojlooy and Hajinezhad, 2023).

The type of communication content can be roughly divided into two types: natural language and vector. Natural language forms of communication have high interpretability. Still, they are difficult to optimize, making them more suitable for

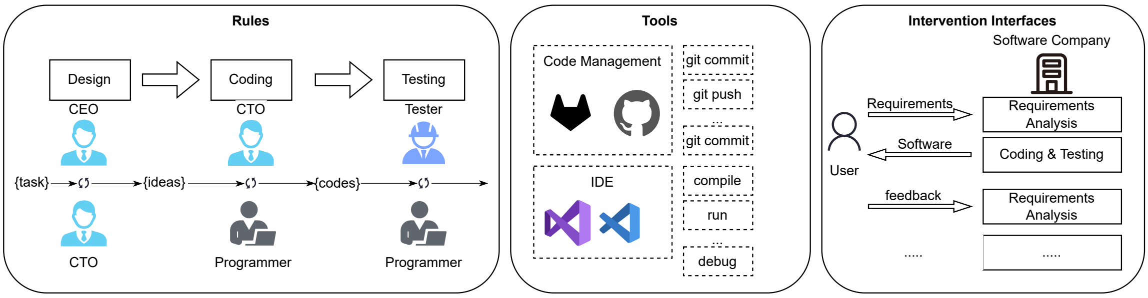

Figure 2: Core components of LLM-MAS environment. Using a software company as an example, agents function within the framework of **rules**, which guide and govern their operations. Meanwhile, **tools** provide APIs for development, such as the "git push" command, which agents can access. Through the **intervention interfaces**, the environment can be modified according to user requirements, enabling continuous optimization of the software.

pursuing consensus, such as in coding (Dong et al., 2024) and job fair systems (Li et al., 2023b). Vector forms are more efficient in terms of communication and easier to optimize using policy gradients, making them commonly used for achieving cooperative objectives (Liu et al., 2024b).

2.2 Environment

Environmental settings include tools, rules, and intervention interfaces, which are illustrated in Figure 2. (i) *Rules* define the mode of communication between generative agents or the interaction with the environment, directly defining the behavioral structure of the entire system. Figure 2 shows the order of agents talking and acting under rules. (ii) *Tools (optional)* create an action space for each generative agent to take action. Figure 2 illustrates the common tools in software development scenario, including IDEs and Git. Their APIs, including git commands, compilation tools, runtime tools, and debugging tools such as "git push", can be accessed by agents. (iii) *Intervention interfaces (optional)* provide an interface for external intervention systems, which can come from any external source, like human (Wang et al., 2024b), or a rule-based model (Chen et al., 2024c), even a generative agent (Chen et al., 2024e). Figure 2 illustrates an example of intervention interfaces in the software development: requirements analysis in agile development. Throughout each development cycle, users from external have the opportunity to communicate with the software company to define and refine their requirements. This ongoing collaboration allows the software company to adjust the development process based on user needs, ensuring timely intervention and alignment with expectations.

3 LLM-MAS for Solving Complex Tasks

In this section, we explore the application of LLM-MAS to solving complex tasks. We begin by categorizing LLM-MAS based on the complexity of the tasks they address. Next, we provide an overview of the relevant code, datasets, and benchmarks available for these applications. Finally, we discuss the evaluation metrics used to assess performance in solving complex tasks.

3.1 Categories of LLM-MAS based on task complexity

We classify LLM-MAS into three distinct categories based on the complexity of tasks they handle: (i) LLM-MAS designed for specific tasks that do not require long trajectory decisions, (ii) LLM-MAS tailored for specific tasks involving long trajectory decisions, and (iii) LLM-MAS that are not specialized for any specific tasks.

Specific tasks that do not require long trajectory decisions. Single tasks refer to tasks without requiring long trajectory decisions. This type of task is commonly seen in tasks requiring knowledge, where techniques from multi-agent systems are transferred to existing classic tasks, such as Visual Question Answering (VQA) (Jiang et al., 2024), tasks in science (Song et al., 2024), etc. Usually, this type of task has a short (less than 2048) context length. This type of task solving by a single agent requires a long prompt, which is difficult to reuse. It is LLM-MAS technology that optimizes this problem. Collective decision-making and reflection are commonly used in this type of task. Compared with a single agent method, such as self-consistency (Wang et al., 2023), LLM-MAS with collective decision-making can achieve improved performance with less prompting for the same task (Du et al., 2024a). Reflective method is also a common technique, Bo et al. (2024) explore reflection in LLM-MAS, proposing a novel framework, that exhibits excellent generalization performance across different actor models. The performance of collective decision-making or reflection depends on the capabilities of individual agents. However, as the complexity of the task increases, within long trajectory decisions, LLM-MAS require a new mod.

Specific tasks that require long trajectory decisions. Complex tasks are defined as those that require decisions over long trajectories. They are typically encountered in multi-stage scenarios where the collaboration of multiple agents is essential for finding a solution (Chen et al., 2024f). Soft development is a representative scenario requiring multi-stage collaboration (Islam et al., 2024). As a representative of this domain, ChatDev (Qian et al., 2024a) leverages software engineer agents in distinct roles to collaboratively develop software. Based on this, additional forms of cooperation are explored (Du et al., 2024b). Further, the scaling law is explored in this scenario (Qian et al., 2024b), but no significant pattern was observed. Another typical scenario is long-context tasks. LONGAGENT (Zhao et al., 2024a) and Chain of Agents (Zhang et al., 2024c) apply MAS technology to split the long context, enabling smaller models like LLaMA-2 7B to possess strong contextual capabilities, even better than GPT-4. Similarly, embodied reasoning and planning are also a representative scenario requiring long trajectories of collaboration (Dasgupta et al., 2023). Agents solve their respective subtasks and merge the results, which introduces higher communication costs and challenges related to information aggregation.

General tasks. In this part, LLM-MAS are not limited to specific tasks but are instead a method applicable to a wide range of general tasks. MetaGPT (Hong et al., 2023) assigns different roles to generative agents to form a collaborative entity for complex tasks. Gao et al. (2024) propose AgentScope with message exchange as its core communication mechanism. Open AI proposes Swarm (OpenAI, 2024b), an experimental multi-agent orchestration framework that is ergonomic and lightweight. KAOS (Zhuo et al., 2024) addresses the challenges of resource coordination management by proposing a unified user experience across various foundational software platforms. In LLM-MAS, fully connected communication poses significant challenges, including combinatorial explosion and privacy risks. To mitigate these issues, researchers have focused on enhancing communication efficiency. For instance, some studies explore methods to accelerate agent interactions through nonverbal communication techniques (Liu et al., 2024b), while others aim to streamline communication by reducing the length of generated messages (Chen et al., 2024g). These approaches collectively address the inherent limitations of fully connected communication in LLM-MAS. Among the works, DroidSpeak achieves up to a 2.78× speedup in prefill latency with negligible loss in accuracy.

3.2 Resources for solving complex tasks

We analyze common LLM-MAS for solving complex tasks in Table 1, including code, datasets, and benchmarks. Among the datasets, QA-style datasets are the most commonly used, a trend that reflects the legacy of traditional NLP task-specific datasets and benchmarks. ToolBench (Guo et al., 2024b), SRDD, ToolAlpaca (Tang et al., 2023), etc. are specifically designed for agent tools. Overcooked-AI (Carroll et al., 2020) is a benchmark for LLM-MAS in the past, which illustrates the potential to transform the game environment originally used for RL based MAS into LLM-MAS.

3.3 Evaluation metric of solving complex task

Performance on specific tasks. Shown as Table 1, the performance of LLM-MAS can be evaluated by specific tasks, which is intuitive and convenient. For example, in an APP system (Zhang et al., 2023b), the average number of steps and tools used by an agent to complete a specific task is considered as an indicator; in BOLAA (Liu et al., 2023c), the recall and QA accuracy of intelligent physical examination retrieval are also considered as evaluation indicators;

Communication cost analysis. The concern lies in the operational cost of the system. Given that a substantial proportion of contemporary systems incorporate LLM-MAS as a pivotal module, the additional expenditure incurred during system operation has emerged as a pivotal area of interest. As an illustrative example, In the evaluation of DroidSpeak (Liu et al., 2024b), the response time has been used as a metric to evaluate the acceleration of the method.

Table 1: Codes and Benchmarks in LLM-MAS for solving task studies. "No Code" or "No Benchmark or Dataset" means the code or benchmark is unavailable.

Field	Subdomain	Paper	Code	Benchmark and Dataset
Tasks without long trajectory decision	Knowledge oriented tasks	(Zhao et al., 2024c)	Code Link	MCQA
		(Wang et al., 2024c)	Code Link	FOLIO-wiki
		(Chen et al., 2024e)	Code Link	StrategyQA, CSQA, GSM8K, AQuA, MATH, Date Understanding, ANLI
		(Chen et al., 2024a)	Code Link	TriviaQA
		(Wang et al., 2024d)	Code Link	TriviaQA
		(Liang et al., 2024)	Code Link	MT-Bench
		(Lei et al., 2024)	Code Link	MATH
		(Zhang et al., 2024a)	Code Link	MMLU, MATH, Chess Move Validity
		(Cheng et al., 2024)	Code Link	ESConv dataset, P4G dataset
		(Tang et al., 2024b)	Code Link	Trans-Review, AutoTransform, T5-Review
	Interaction oriented tasks	(Zhang et al., 2024b)	Code Link	RoCoBench,Overcooked-AI
		(Zhang et al., 2023a)	Code Link	Overcooked-AI
Tasks within long trajectory decision	Multi-stage tasks	(Qian et al., 2024a)	Code Link	SRDD
		(Du et al., 2024b)	Code Link	SRDD
		(Yue et al., 2024)	Code Link	SMART (self)
		(Liu et al., 2023c)	Code Link	WebShop
		(Lin et al., 2024)	Code Link	FG-C, CG-O
		(Islam et al., 2024)	Code Link	HumanEval, EvalPlus, MBPP, APPS, xCodeEval, CodeContest
		(Shen et al., 2024)	Code Link	ToolBench, ToolAlpaca
General tasks		(Li et al., 2023a)	Code Link	CAMEL AI Society, CAMEL Code, CAMEL Math, CAMEL Science

4 LLM-MAS for Simulating Specific Scenarios

This section will illustrate the application for LLM-MAS in simulation. LLM-MAS are applied by researchers to simulate certain scenarios to study their impact on specific subjects such as social sciences. On the one hand, compared with rule-based methods (Chuang and Rogers, 2023), generative agents with natural language communication can be more intuitive for humans. On the other hand, environment determines the properties of the simulation, which is the core of the entire simulation.

4.1 Categories of simulation scenarios

The typical scenarios for LLM-MAS simulations are described as follows. We will introduce the following work according to the subject.

Social domain. Social large-scale experiments in the real world have high costs, and the sheer scale of social participation can sometimes escalate into violence and destruction, posing potential ramifications (Mou et al., 2024). Therefore, it is necessary to simulate in the virtual environment; simulation can solve the problem of excessive overhead in the real environment and can simulate the process in the real world for a long time at a faster speed (Li et al., 2024a). At the same time, the whole process can be easily repeated, which is conducive to further research. Researchers have done a lot of work to simulate social media scenarios. Based on the social media simulation archetype (Park et al., 2022), Park et al. (2023) propose Stanford Town, which leads to a one-day simulation of the life of 25 agents with different occupations in a small American town. At the same time, there was work on emotional propagation influence (Gao et al., 2023b), information cocoon room based on recommendation scenario research (Wang et al., 2024b), and study of social movements (Mou et al., 2024). Researchers propose Urban Generative Intelligence (UGI) (Xu et al., 2023a) to address specific urban issues and simulate complex urban systems, providing a multidisciplinary approach to understanding and managing urban complexity. Li et al. (2024a) study doctor agent evolution method by hospital simulation. Because doctor agent training is both inexpensive and highly effective, this work can quickly scale up the agent to handle tens of thousands of cases in just a few days, a task that would take a human doctor years to complete. Pan et al. (2024) propose a huge scale of agent simulation, increasing the number of agents to 10^6. In social games, like Werewolf (Xu et al., 2024a), Avalon (Lan et al., 2024), and Minecraft (Gong et al., 2024) for LLM-MAS simulation are attempted.

Physical domain. For the physical domain, the applications for generative agent simulation in-

Table 2: Codes and Benchmarks in LLM-MAS for simulation studies. "No Code" or "No Benchmark or Dataset" means the code or benchmark is unavailable.

Domain	Subdomain	Paper	Code	Benchmark and Dataset
Social	Tiny Society	(Huang et al., 2024b)	No Code	AdaSociety
		(Chen et al., 2024b)	Code Link	AgentCourt
		(Park et al., 2023)	Code Link	No Benchmark or Dataset
		(Piatti et al., 2024)	Code Link	No Benchmark
		(Chuang et al., 2024)	Code Link	No Benchmark or Dataset
	Economics	(Li et al., 2024b)	Code Link	No Benchmark or Dataset
	Social Media	(Wang et al., 2024b)	Code Link	Movielens-1M
		(Gao et al., 2023b)	No Code	Blog Authorship Corpus
		(Mou et al., 2024)	Code Link	SoMoSiMu-Bench(self)
	Game	(Du and Zhang, 2024)	Code Link	WWQA
		(Pan et al., 2024)	Code Link	No Benchmark or Dataset
Physical	Wireless	(Zou et al., 2023)	No Code	No Benchmark or Dataset

clude mobility behaviors, transportation (Gao et al., 2023a), wireless networks, etc. However, there is limited research in the area of generative agents. Zou et al. (2023) explore the application of multiple agents in the wireless field, proposing a framework where multiple on-device agents can interact with the environment to simulate real world scenarios. This is an area of critical importance for the future of embodied intelligence.

4.2 Resources for LLM-MAS simulation

We analyze common and open-source LLM-MAS for simulation with their datasets in Table 2, including code and benchmarks.

To prove the effectiveness of the simulation, that is, to fit the reality, researchers usually evaluate the simulation system by simulating real data. Therefore, a realistic dataset with dense users and records is very important for evaluation simulation (Mou et al., 2024). An ideal dataset will be dense: that is, data with a smaller number of users on the same scale can better evaluate the simulation capability of the LLM-MAS. Du and Zhang (2024) propose WWQA based on werewolf scenarios to evaluate the agent's capability in a werewolf scenario.

4.3 Evaluation Metric of LLM-MAS simulation

We will summarize the metrics for the overall evaluation of LLM-MAS , rather than the capabilities of individual agents.

Consistency. LLM-MAS necessitate a robust congruence with the real world to ensure the derivation of meaningful and insightful experimental outcomes. In the context of simulation systems, exemplified by UGI (Xu et al., 2023a), the primary objective lies in faithfully replicating specific real-world scenarios. When employed for training agents like SMART (Yue et al., 2024), only those agents that have undergone rigorous training within a virtual environment that closely mirrors the real environment can be deemed suitable for deployment in real-world settings. Similarly, when utilized for evaluation purposes, such as in AgentSims (Lin et al., 2023), the attainment of authentic and reliable evaluation results is contingent upon the virtual environment maintaining a high degree of consistency with its real-world counterpart. Finally, in the system for collecting data such as BOLAA (Liu et al., 2023c), consistency also ensures the validity of the data. Therefore, an important performance measure of LLM-MAS is its consistency with the real situation.

Information dissemination. Compare the differences between information dissemination behavior in the system and reality using time series analysis methods. Information dissemination can to some extent reflect the nature of media; therefore, a realistic multi-agent system should have a similar information dissemination trend to the real world. Abdelzaher et al. (2020) compare the changes in the number of events occurring each day in an online social media simulation environment; S3 (Gao et al., 2023b) compare the number of users who are aware of a certain event every day, as well as the changes in emotional density and support rate for that event every day; a similar approach is also used in Stanford Town (Park et al., 2023).

5 LLM-MAS for Evaluating and Training Generative Agents

With generative agents prevailing in the community (Wang et al., 2024a), how to evaluate the ability of generative agents is an open question. Existing evaluation methods suffer from the following shortcomings: (i) constrained evaluation abilities, (ii) vulnerable benchmarks, and (iii) unobjective metrics. The complexity and diversity of LLM-MAS have indicated that LLM-MAS can evaluate generative agent. However, how to design specific evaluation indicators and evaluation methods has puzzled researchers. Similarly, LLM-MAS can also be used in training generative agents. We summarize three aspects of training: (i) Supervised Fine-Tuning (SFT) (ii) reinforcement learning (RL) (iii) Synthesizing data for training.

5.1 Methods of Evaluation and Training on Generative Agents

LLM-MAS can provide rewards to agents, and these rewards can be used to evaluate or train generative agents, which will be discussed below.

Evaluation of generative agents. Researchers study generative agents by putting them into LLM-MAS. In LLM-MAS, researchers can further study the LLM's strategic capabilities in different scenes, such as long strategic ability (Chen et al., 2024c), corporation strategy (Xu et al., 2023b), and competitiveness strategy (Zhao et al., 2024b). In the emotional field, MuMA-ToM (Shi et al., 2024) is used to evaluate the ability of agents to understand and reason about human interactions in a real home environment through video and text descriptions.

Training on generative agents. Li et al. (2024c) enhance the data to Supervised Fine-Tuning (SFT) generative agents with LLM-MAS. Xu et al. (2024b) have created generative agents to overcome the intrinsic bias from LLMs by proposing a novel framework that powers generative agents with multi-agent reinforcement learning. For LLM-MAS, Yue et al. (2024) split complex trajectories in knowledge-intensive tasks into subtasks, proposing a co-training paradigm of the multi-agent framework, Long- and Short-Trajectory Learning, which ensures synergy while keeping the fine-grained performance of each agent. RLHF has been criticized for its high cost. Liu et al. (2023a) propose an alignment scheme based on a multi-agent system, effectively addressing instability and reward gaming concerns associated with reward-based RL optimization. Either way, LLM-MAS are essentially viewed as an environment in RL with different ways of getting rewards from the environment.

5.2 Resources of LLM-MAS for evaluations

Table 3 shows the work with code, dataset, and benchmark we analyze, serving as a reference for future researchers. Our findings indicate that the current body of research is predominantly centered on the evaluation of generative agents, which means training with LLM-MAS will be a great potential for further exploration.

6 Challenges and Future Directions

While previous work on LLM-MAS has obtained many remarkable successes, this field is still at its initial stage, and there are several significant challenges that need to be addressed in its development. In the following, we outline several key challenges along with potential future directions.

6.1 Challenges posed by generative agents

Generative agents are an integral part of LLM-MAS. However, the generative agents have some shortcomings due to the inherent characteristics of the base model LLMs, which will be carefully discussed below.

Challenges. (i) Generalized alignment for simulation (Liu et al., 2023a). When the agents are leveraged for real-world simulation, a perfect generative agent should be able to depict diverse traits (Wang et al., 2024a) honestly. However, due to the training method of the foundation model (OpenAI et al., 2024), generative agents usually cannot be aligned with mock objects. (ii) Hallucination. Generative agents have a certain probability of hallucination in their interaction with other agents (Du et al., 2024a). Various enhancement methods can alleviate this problem but cannot solve it (iii) Lack of sufficient long text capability. When processing complex information, generative agents forget the input information because of the lack of long-text ability (Zhao et al., 2024a).

Future directions. The improvement of the

Table 3: Codes and Benchmarks in LLM-MAS for evaluation studies. “No Code” or “No Benchmark or Dataset” means the code or benchmark is unavailable.

Domain	Subdomain	Paper	Code	Benchmark and Dataset
Evaluation of generative agents	Strategy	(Liu et al., 2023b)	Code Link	AGENTBENCH
		(Bandi and Harrasse, 2024)	No Code	MT-Bench
		(Chan et al., 2023)	Code Link	ChatEval
		(Chen et al., 2024d)	Code Link	LLMARENA
		(Xu et al., 2023b)	Code Link	MAgIC
		(Huang et al., 2024a)	Code Link	MLAgentBench
		(Chen et al., 2024c)	Code Link	AUCARENA
	Emotion	(Zhang et al., 2024d)	Code Link	PsySafe
		(Shi et al., 2024)	Code Link	MuMA-ToM
Training on generative agents	SFT on LLM-MAS	(Li et al., 2024c)	Code Link	MT-Bench, AlpacaEval
	MARL on LLM-MAS	(Xu et al., 2024b)	No Code	No Benchmark or Dataset
	Synthesized Ddata	(Liu et al., 2023a)	Code Link	HH, Moral Stories, MIC, ETHICS-Deontology, TruthfulQA

ability of a single agent or the ability of the base model has always been a hot topic. Researchers have focused on enhancing alignment, reducing hallucination, and improving the ability of long text. The proposal of the new generation of Open AI model o1 (OpenAI, 2024a) provides researchers with new ideas, that is, to use **more complex reasoning** to enhance the ability of the model.

6.2 Challenges posed by interactions

Challenges. Due to the complexity, autoregressive, and other characteristics of LLM-MAS, there are many problems in the practical application of the system. How to solve (i) communication efficiency (Liu et al., 2024b; Zhuang et al., 2024), (ii) imperfect communication (Zhang et al., 2023a; Liu et al., 2024a; Zhuang et al., 2024), and (iii) communication security (de Cerqueira et al., 2024) is a long-term goal of the researchers.
Future directions. Establishing a comprehensive and standardized benchmark to evaluate the communication latency of LLM-MAS is an urgent issue that needs to be addressed in the short term. Therefore, optimizing the communication structure of LLM-MAS presents an intriguing research problem for the near future.

6.3 Challenges of Evaluation for LLM-MAS

Lack of Objective metrics for group behavior. As shown in Section 4.3, due to the diversity, complexity, and unpredictability of multi-agent environments, it is difficult to obtain sufficiently detailed, specific, and direct system evaluation indicators from current work at the system level.
Automated evaluation and benchmark. Different LLM-MAS of the same kind cannot be compared because of the lack of a benchmark for LLM-MAS. Further, there is a lack of a common benchmark framework for both individual and total-based evaluation, that can be used to evaluate most LLM-MAS.
Future directions. Studying large-scale LLM-MAS will be a new research hotspot, from which researchers will evaluate and discover new scale effects. In the meantime, common test benchmarks and evaluation methods will also emerge in future research.

7 Conclusion

In this survey, we systematically summarize existing research in the application of LLM-based multi-agent systems (LLM-MAS) field. We present and review these studies from three application aspects: task-solving, simulation, and evaluation of the generative agents. We provide a detailed taxonomy to draw connections among the existing research, summarizing the major techniques and their development histories for each of these aspects. In addition to reviewing the previous work, we also propose several challenges in this field, which are expected to guide potential future directions.

Limitations

Due to page limitations, we provide only brief summaries of each method without delving into exhaustive technical details. Furthermore, our primary collection includes studies from ∗ACL, NeurIPS, ICLR, AAAI, and arXiv, which means some important work from other venues might have been inadvertently omitted. In the application section, we have listed representative LLM-MAS resources with open code in Tables 1, 2, and 3. We recognize the timeliness of our work and are committed to keeping pace with ongoing discussions in the research community, updating our perspectives and supplementing any overlooked contributions in future revisions.

References

Tarek Abdelzaher, Jiawei Han, Yifan Hao, Andong Jing, Dongxin Liu, Shengzhong Liu, Hoang Hai Nguyen, David M Nicol, Huajie Shao, Tianshi Wang, et al. 2020. Multiscale online media simulation with socialcube. *Computational and Mathematical Organization Theory*, 26:145–174.

P. G. Balaji and D. Srinivasan. 2010. An Introduction to Multi-Agent Systems. In Dipti Srinivasan and Lakhmi C. Jain, editors, *Innovations in Multi-Agent Systems and Applications - 1*, pages 1–27. Springer, Berlin, Heidelberg.

Chaithanya Bandi and Abir Harrasse. 2024. Adversarial Multi-Agent Evaluation of Large Language Models through Iterative Debates. *Preprint*, arXiv:2410.04663.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 138595–138631. Curran Associates, Inc.

G Brockman. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2020. On the Utility of Learning about Humans for Human-AI Coordination. *Preprint*, arXiv:1910.05789.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *Preprint*, arXiv:2308.07201.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. 2024a. AutoAgents: A Framework for Automatic Agent Generation. *Preprint*, arXiv:2309.17288.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024b. AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents. *Preprint*, arXiv:2408.08089.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2024c. Put Your Money Where Your Mouth Is: Evaluating Strategic Planning and Execution of LLM Agents in an Auction Arena. *Preprint*, arXiv:2310.05746.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024d. LLMArena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13055–13077.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024e. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Pei Chen, Shuai Zhang, and Boran Han. 2024f. CoMM: Collaborative Multi-Agent, Multi-Reasoning-Path Prompting for Complex Problem Solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1720–1738, Mexico City, Mexico. Association for Computational Linguistics.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024g. Optima: Optimizing Effectiveness and Efficiency for LLM-Based Multi-Agent System. *Preprint*, arXiv:2410.08115.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating Specialized Agents towards a Complex Dialogue Goal. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17853–17861.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Yun-Shiuan Chuang and Timothy T. Rogers. 2023. Computational Agent-based Models in Opinion Dynamics: A Survey on Social Simulations and Empirical Studies. *Preprint*, arXiv:2306.03446.

Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. 2023. Collaborating with language models for embodied reasoning. *Preprint*, arXiv:2302.00763.

José Antonio Siqueira de Cerqueira, Mamia Agbese, Rebekah Rousi, Nannan Xi, Juho Hamari, and Pekka Abrahamsson. 2024. Can We Trust AI Agents? An Experimental Study Towards Trustworthy LLM-Based Multi-Agent Systems for AI Ethics. *Preprint*, arXiv:2411.08881.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-Collaboration Code Generation via ChatGPT. *ACM Trans. Softw. Eng. Methodol.*, 33(7):189:1–189:38.

Silin Du and Xiaowei Zhang. 2024. Helmsman of the Masses? Evaluate the Opinion Leadership of Large Language Models in the Werewolf Game. In *First Conference on Language Modeling*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024a. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 11733–11763, Vienna, Austria. JMLR.org.

Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Yufan Dang, Weize Chen, and Cheng Yang. 2024b. Multi-Agent Software Development through Cross-Team Collaboration. *Preprint*, arXiv:2406.08979.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2023a. Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives. *Preprint*, arXiv:2312.11970.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023b. S3: Social-network Simulation System with Large Language Model-Empowered Agents. *Preprint*, arXiv:2307.14984.

Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, Liuyi Yao, Hongyi Peng, Zeyu Zhang, Lin Zhu, Chen Cheng, Hongzhu Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. AgentScope: A Flexible yet Robust Multi-Agent Platform. *Preprint*, arXiv:2402.14034.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Li Fei-Fei, Jianfeng Gao, and Hoi Vo. 2024. MindAgent: Emergent Gaming Interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3154–3183, Mexico City, Mexico. Association for Computational Linguistics.

Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(2):895–943.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024a. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Thirty-Third International Joint Conference on Artificial Intelligence*, volume 9, pages 8048–8057.

Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. 2024b. StableToolBench: Towards Stable Large-Scale Benchmarking on Tool Learning of Large Language Models. *Preprint*, arXiv:2403.07714.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. LLM Multi-Agent Systems: Challenges and Open Problems. *Preprint*, arXiv:2402.03578.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *Preprint*, arXiv:2308.00352.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024a. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. *Preprint*, arXiv:2310.03302.

Yizhe Huang, Xingbo Wang, Hao Liu, Fanqi Kong, Aoyang Qin, Min Tang, Xiaoxi Wang, Song-Chun Zhu, Mingjie Bi, Siyuan Qi, and Xue Feng. 2024b. AdaSociety: An Adaptive Environment with Social Structures for Multi-Agent Decision-Making. *Preprint*, arXiv:2411.03865.

Md. Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. MapCoder: Multi-Agent Code Generation for Competitive Problem Solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4912–4944, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Jiang, Zhijun Zhuang, Shreyas S. Shivakumar, Dan Roth, and Camillo J. Taylor. 2024. Multi-Agent VQA: Exploring Multi-Agent Foundation Models in Zero-Shot Visual Question Answering. *Preprint*, arXiv:2403.14783.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. *Preprint*, arXiv:2310.14985.

Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 2024. MACM: Utilizing a Multi-Agent System for Condition Mining in Solving Complex Mathematical Problems. *Preprint*, arXiv:2404.04735.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel:

Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024a. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *Preprint*, arXiv:2405.02957.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024b. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Renhao Li, Minghuan Tan, Derek F. Wong, and Min Yang. 2024c. CoEvol: Constructing Better Responses for Instruction Finetuning through Multi-Agent Cooperation. *Preprint*, arXiv:2406.07054.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024d. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023b. MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. *Preprint*, arXiv:2310.06500.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *Preprint*, arXiv:2305.19118.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. 2023. AgentSims: An Open-Source Sandbox for Large Language Model Evaluation. *Preprint*, arXiv:2308.04026.

Leilei Lin, Yumeng Jin, Yingming Zhou, Wenlong Chen, and Chen Qian. 2024. MAO: A Framework for Process Model Generation with Multi-Agent Orchestration. *Preprint*, arXiv:2408.01916.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2023a. Training Socially Aligned Language Models on Simulated Social Interactions. In *The Twelfth International Conference on Learning Representations*.

Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. 2024a. Autonomous Agents for Collaborative Task under Information Asymmetry. *Preprint*, arXiv:2406.14928.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023b. AgentBench: Evaluating LLMs as Agents. *Preprint*, arXiv:2308.03688.

Yuhan Liu, Esha Choukse, Shan Lu, Junchen Jiang, and Madan Musuvathi. 2024b. DroidSpeak: Enhancing Cross-LLM Communication. *Preprint*, arXiv:2411.02820.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2023c. BOLAA: Benchmarking and Orchestrating LLM-augmented Autonomous Agents. *Preprint*, arXiv:2308.05960.

Volodymyr Mnih. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4789–4809, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI. 2024a. OpenAI o1. https://openai.com/o1/.

OpenAI. 2024b. Openai/swarm. OpenAI.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722.

Xuchen Pan, Dawei Gao, Yuexiang Xie, Yushuo Chen, Zhewei Wei, Yaliang Li, Bolin Ding, Ji-Rong Wen, and Jingren Zhou. 2024. Very Large-Scale Multi-Agent Simulation in AgentScope. *Preprint*, arXiv:2407.17789.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *Preprint*, arXiv:2304.03442.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. *Preprint*, arXiv:2208.04024.

Bhrij Patel, Vishnu Sashank Dorbala, Amrit Singh Bedi, and Dinesh Manocha. 2024. Multi-LLM QA with Embodied Exploration. *Preprint*, arXiv:2406.10918.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. *Preprint*, arXiv:2404.16698.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024a. ChatDev: Communicative Agents for Software Development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024b. Scaling Large-Language-Model-based Multi-Agent Collaboration. *Preprint*, arXiv:2406.07155.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. 2024. Small LLMs Are Weak Tool Learners: A Multi-LLM Agent. *Preprint*, arXiv:2401.07324.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2024. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. *Preprint*, arXiv:2408.12574.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Tao Song, Man Luo, Linjiang Chen, Yan Huang, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, Fei Zhang, Weiwei Shang, Jun Jiang, and Yi Luo. 2024. A multi-agent-driven robotic AI chemist enabling autonomous chemical research on demand.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. 2023. ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases. *Preprint*, arXiv:2306.05301.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024a. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *Preprint*, arXiv:2311.10537.

Xunzhu Tang, Kisub Kim, Yewei Song, Cedric Lothritz, Bei Li, Saad Ezzini, Haoye Tian, Jacques Klein, and Tegawende F. Bissyande. 2024b. CodeAgent: Autonomous Communicative Agents for Code Review. *Preprint*, arXiv:2402.02172.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2024b. User Behavior Simulation with Large Language Model based Agents. *Preprint*, arXiv:2306.02552.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024c. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *Preprint*, arXiv:2203.11171.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024d. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023a. Urban Generative Intelligence (UGI): A Foundational Platform for Agents in Embodied City Environment. *Preprint*, arXiv:2312.11813.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. 2023b. MAgIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. *Preprint*, arXiv:2311.08562.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2024a. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. *Preprint*, arXiv:2309.04658.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2024b. Language agents with reinforcement learning for strategic play in the Werewolf game. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 55434–55464, Vienna, Austria. JMLR.org.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023.

ReAct: Synergizing Reasoning and Acting in Language Models. *Preprint*, arXiv:2210.03629.

Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023. Skill Reinforcement Learning and Planning for Open-World Long-Horizon Tasks. *Preprint*, arXiv:2303.16563.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Synergistic Multi-Agent Framework with Trajectory Learning for Knowledge-Intensive Tasks. *Preprint*, arXiv:2407.09893.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, Xiaojun Chang, Junge Zhang, Feng Yin, Yitao Liang, and Yaodong Yang. 2023a. ProAgent: Building Proactive Cooperative AI with Large Language Models. *CoRR*.

Chi Zhang, Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. 2023b. AppAgent: Multimodal Agents as Smartphone Users. *Preprint*, arXiv:2312.13771.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024a. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. 2024b. Towards Efficient LLM Grounding for Embodied Multi-Agent Collaboration. *Preprint*, arXiv:2405.14314.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arik. 2024c. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. *Preprint*, arXiv:2406.02818.

Zaibin Zhang, Yongting Zhang, Lijun Li, Jing Shao, Hongzhi Gao, Yu Qiao, Lijun Wang, Huchuan Lu, and Feng Zhao. 2024d. PsySafe: A Comprehensive Framework for Psychological-based Attack, Defense, and Evaluation of Multi-agent System Safety. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15202–15231, Bangkok, Thailand. Association for Computational Linguistics.

Jun Zhao, Can Zu, Xu Hao, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024a. LONGAGENT: Achieving Question Answering for 128k-Token-Long Documents through Multi-Agent Collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16310–16324, Miami, Florida, USA. Association for Computational Linguistics.

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024b. CompeteAI: Understanding the Competition Dynamics in Large Language Model-based Agents. *Preprint*, arXiv:2310.17512.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024c. An Electoral Approach to Diversify LLM-based Multi-Agent Collective Decision-Making. *Preprint*, arXiv:2410.15168.

Yuan Zhuang, Yi Shen, Zhili Zhang, Yuxiao Chen, and Fei Miao. 2024. YOLO-MARL: You Only LLM Once for Multi-agent Reinforcement Learning. *Preprint*, arXiv:2410.03997.

Zhao Zhuo, Rongzhen Li, Kai Liu, Huhai Zou, KaiMao Li, Jie Yu, Tianhao Sun, and Qingbo Wu. 2024. KAOS: Large Model Multi-Agent Operating System. *Preprint*, arXiv:2406.11342.

Hang Zou, Qiyang Zhao, Lina Bariah, Mehdi Bennis, and Merouane Debbah. 2023. Wireless Multi-Agent Generative AI: From Connected Intelligence to Collective Intelligence. *Preprint*, arXiv:2307.02757.

A Use of AI Assistance

This paper has been edited with the assistance of an AI-powered writing assistant. The tool was used to refine the clarity, coherence, and flow of the text, ensuring that the language was precise and well-structured while maintaining the integrity of the original content.