

METAGENTS: Large Language Model Based Agents for Decision-making on Teaming

YUAN LI, University of Cambridge, United Kingdom

LICHAO SUN, Lehigh University, United States

YIXUAN ZHANG, William & Mary, United States

Significant advancements have occurred in the application of Large Language Models (LLMs) for social simulations. Despite this, their abilities to perform teaming in task-oriented social events are underexplored. Such capabilities are crucial if LLMs are to effectively mimic human-like social behaviors and form efficient teams to solve tasks. To bridge this gap, we introduce METAGENTS, a social simulation framework populated with LLM-based agents. METAGENTS facilitates agent engagement in conversations and a series of decision making within social contexts, serving as an appropriate platform for investigating interactions and interpersonal decision-making of agents. In particular, we construct a job fair environment as a case study to scrutinize the team assembly and skill-matching behaviors of LLM-based agents. We take advantage of both quantitative metrics evaluation and qualitative text analysis to assess their teaming abilities at the job fair. Our evaluation demonstrates that LLM-based agents perform competently in making rational decisions to develop efficient teams. However, we also identify limitations that hinder their effectiveness in more complex team assembly tasks. Our work provides valuable insights into the role and evolution of LLMs in task-oriented social simulations.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Natural language processing**.

Additional Key Words and Phrases: generative AI, large language models, generative agents, collaboration, information retrieval, coordination

ACM Reference Format:

Yuan Li, Lichao Sun, and Yixuan Zhang. 2025. METAGENTS: Large Language Model Based Agents for Decision-making on Teaming. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW134 (April 2025), 27 pages. <https://doi.org/10.1145/3711032>

1 Introduction

Large language models (LLMs), such as ChatGPT [37] and GPT-4 [38], have gained significant attention due to their exceptional abilities in natural language processing. Recent studies extend the scope of these LLMs beyond text generation, positioning LLMs as versatile agents capable of conversational engagement, decision-making, and task completion [56]. A noteworthy development in this domain is LLM-based agent [56], which employ LLMs to carry out human-like actions, ranging from planning and conversational interaction [41] to task solving [23, 43]. An area yet to be fully explored is LLM-based agents' social intelligence—specifically, organizing teams and aligning agent expertise with relevant roles in the team. These abilities are crucial for advancing multi-agent systems, as they involve self-optimization in task-oriented team assembly.

Authors' Contact Information: Yuan Li, University of Cambridge, Cambridge, United Kingdom, y1967@cam.ac.uk; Lichao Sun, Lehigh University, Bethlehem, United States, lis221@lehigh.edu; Yixuan Zhang, William & Mary, Williamsburg, United States, yzhang104@wm.edu.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 2573-0142/2025/4-ARTCSCW134

<https://doi.org/10.1145/3711032>

Prior studies in multi-agent systems have primarily focused on task-solving abilities, constructing LLM-based agents to collaborate on tasks. Representative examples include ChatDev [43] and MetaGPT [23], both multi-agent frameworks for autonomous software development in which agents fulfill different roles within predetermined workflows and team compositions. Their construction leverages the empirical conclusion that recruiting experts to form a team could improve the effectiveness in solving tasks [61]. However, these multi-agent systems are largely bound to human-prescribed team compositions and workflows. In other words, LLM-based agent teams are programmed by humans, and these multi-agent systems do not have autonomy for team assembly. The current design restricts the agents' adaptability, particularly in responding to unforeseen challenges and dynamic scenarios. This limitation can be overcome by designing a multi-agent system where LLM-based agents can form teams from scratch. Such a system would enable LLM-based agents to select team members and adjust the team composition dynamically based on the evolving needs of the tasks at hand. To realize this versatile multi-agent system, a question remains unsolved: **Can LLM-based agents perform teaming in social contexts by correctly assigning agents to appropriate positions in the teams?** Teaming is important as it increases efficiency by placing individuals in roles that match their competencies or preferences [19]. In this paper, we investigate team assembly in the social context using LLM-based agents, taking the job fair scenario as a study case. We aim to assess whether LLM-based agents can manage team assembly tasks through interactions. This exploration is critical to understanding the potential of LLM-based multi-agent systems in teaming [18].

In this work, we introduce METAAGENTS, a framework that simulates team assembly involving interactions and decision-making. In METAAGENTS, LLM-based agents engage in multi-turn conversations, and are socially connected through sequential decision-making. Based on METAAGENTS, we simulate a job fair environment to examine team formation behaviors in LLM-based multi-agent systems. We focus on team assembly and workflow design processes, investigating the assignment of responsibilities among agents based on their skills and backgrounds. The choice of a job fair in our system simulation is an appropriate teaming scenario due to its resemblance to real-world environments. This scenario involves complex social interactions, decision-making, and role dynamics, making it a suitable testbed for exploring social behaviors and interactions. Furthermore, by simulating an environment in which agents communicate and form teams, we can test and refine our framework. Based on the simulated environment, we evaluate agents' social abilities in teaming and designing workflows, and analyze their performance in four distinct job fair scenarios. This job fair environment can also be generalized to other collaborative settings that require teaming and matching, such as matching roommates and seeking collaborators. Our findings indicate that LLM-based agents are capable of proposing workflows and utilizing conversational information for team formation. However, challenges arise as the number of participants increases, often due to misalignment or dishonesty of agents. We further discuss the broader applications of our work in human resource management and computational multi-agent systems. This study contributes to the HCI/CSCW community by providing social simulations for team formation, offering insights into effective teaming practices, and deepening our understanding of the behaviors of LLMs in social scenarios.

Our contributions are summarized as follows:

- **METAAGENTS framework:** We introduce a novel framework to study task-driven social intelligence of LLM-based agents. METAAGENTS provides a simulated environment for exploring teaming behaviors, enabling agents to communicate and make interpersonal decisions. Using a simulated job fair as a testbed, we investigate their decision-making behaviors in real-world-like scenarios. METAAGENTS framework can generalize to wider teaming scenarios such as networking sessions and collaborator seeking.

- **In-depth Evaluation:** We evaluate LLM-based agents on their abilities to identify capable agents and propose correct team workflows. We propose task-oriented metrics to evaluate their decision-making, and conduct a qualitative review of LLM's generation.
- **Behavioral Patterns of LLM-Based Agents:** Our findings unveil the potential and challenges of LLM-based agents, providing insights for effective information retrieval in complex social settings. We find that LLM-based agents have decent knowledge of workflow, while their behaviors are not always aligned with their persona and they are sometimes not honest about their capabilities, which constrains them from being more cooperative. These findings could lead to better design of multi-agent systems in collaborative settings, enabling more effective integration of AI in CSCW.

2 Related Work

2.1 LLM-Based Agents

Autonomous agents have been recognized as a promising path toward artificial general intelligence, where LLM-based agents are expected to serve as general-purpose assistants [56, 62]. Prior literature has focused on using LLMs for various tasks through agent-based system design. For example, Chen et al. [11] used LLMs to assist in agent-based modeling on a NetLogo Chat platform, while other works [24, 27, 70] have introduced memory mechanisms to enhance LLM-based agents. Hou et al. [24], Zulfikar et al. [70], and Huang et al. [27] designed different memory mechanisms to enhance LLM-based agents as more powerful assistants. In addition, a popular research topic on LLM-based agents focuses on their task-solving capabilities, as demonstrated by notable examples like Auto-GPT [49], ChatDev [43], and MetaGPT [23]. These LLM-based agent systems typically involve breaking down complex tasks into smaller, manageable components and assigning specific LLM-based agents to handle each segment. Our work differentiates from these systems in the setup. While existing agent systems follow predetermined procedures and team compositions, we take a step back, querying whether LLMs could propose workflows and perform team assembly on their own.

Another line of research efforts lies in employing LLMs for simulation, in which LLMs act as players or characters. Park et al. [41] designed a community of 25 LLM-based agents, termed generative agents, capable of planning, communicating, and forming connections. An intriguing phenomenon among generative agents is that an agent organizes a party, which demonstrates some task-oriented endeavors. Similarly, AgentSims[32] offers a detailed virtual town, populated by agents with capabilities such as planning and tool use, serving as a platform to study social skills. Similar ideas have been applied to the simulation of specific domains. RecAgent [57] simulates a recommendation ecosystem featuring various types of agents, including recommenders and interactive user agents. The simulation centered around LLM-based agents also extends to economy [16, 31], history [25], and political science [3, 69]. These multi-agent simulations offer platforms for exploring the behaviors of LLM-based agents and investigating their daily planning and interpersonal interactions. However, they have not fully investigated their social capabilities. Thus, the limitation of applying LLM-based agents towards more productive and collaborative systems is largely underexplored. To address this research gap, our work aims to provide a social computing framework that could evaluate the social abilities of LLM-based agents, unveiling their potential and limitations in further extending them into more complicated and robust social simulation systems.

2.2 Social Behaviors of LLMs

As LLMs exhibit impressive language understanding and generation abilities, there is a growing research interest in the social and psychological aspects of LLMs. For example, Aher et al. [2] suggested that LLMs could reproduce economic, psycholinguistic, and social psychology experiments. Zhou et al. [67] proposed a novel prompt approach based on the theory of mind, which encourages LLMs to anticipate future challenges and better decide potential actions [68]. Ziems et al. [69] gave a comprehensive evaluation of LLMs on computational social science tasks. They found that LLMs could reliably classify and explain social phenomena, thus contributing meaningfully to the social science domain. Furthermore, some prior works used LLMs to simulate specific social behaviors, including competition and negotiation. For example, Zhao et al. [66] introduced a framework to study the dynamics of competition using LLM-based agents and instantiate a study case with restaurant and customer agents. Similarly, Bianchi et al. [7] evaluated the negotiation abilities of LLMs using simulated scenarios under resource constraints.

Prior literature also investigates the decision-making of LLMs within the social context. Brookins et al. [9] investigated the LLMs' behaviors under the Prisoner's Dilemma. They observed that ChatGPT tends towards fairness and cooperation, diverging from optimal strategies and displaying a greater inclination towards these traits than human participants. Exploring LLMs in simulated game contexts, Wang et al. [58] utilized the intricate Avalon game as a testbed, which is characterized by misinformation and the deceptive nature of communication. With their proposed reasoning enhancement approaches, they find that LLMs are generally capable of making rational decisions, outperforming humans, and having decent consistency. This line of inquiry extends to other complex games like Werewolf, where studies by Xu et al. [63, 64] highlighted LLMs' sophisticated strategic reasoning abilities.

Building on these insights, our research focuses on the social behaviors of LLM-based agents in teaming scenarios—an area less covered by existing literature. We aim to examine the dynamics of agent teaming and interpersonal decision-making processes, which are crucial for understanding LLMs' social capabilities and their applicability to real-world situations, as described in Gomez et al. [18]. Our work emphasizes the importance of teaming abilities as an indication of broader social intelligence and its potential generalizability to various contexts.

2.3 Team Assembly

We position our work within prior studies of team assembly research across HCI, CSCW, and social sciences. Team assembly is a challenging task that requires social intelligence to obtain information about individuals' attributes and operate in specific social contexts [18]. Early CSCW research has explored how technologies facilitate teams' communication, coordination, and collaboration [15, 17]. More recently, scholars have explored methods to aid the formation of teams, including leveraging current computational infrastructure and combining users' digital trace data and network information [22]. Socio-technical systems can help people assemble effective teams by combining several data sources, analyzing users' trace data, and performing various calculations to find the most efficient team combinations [18, 22, 28].

Existing literature has identified crucial socio-technical aspects that should be considered in team assembly, encompassing team contextual constraints [33], team tasks, team members' personalities [54], and expertise [60]. These factors determine team composition, primarily focusing on enhancing team effectiveness post-formation. Furthermore, Gomez et al. [18] broadened this view by providing a theoretical framework on how socio-technical systems influence team formation while integrating CSCW frameworks. Despite this, there remains a gap in understanding how social interactions within these systems influence decision-making during team assembly.

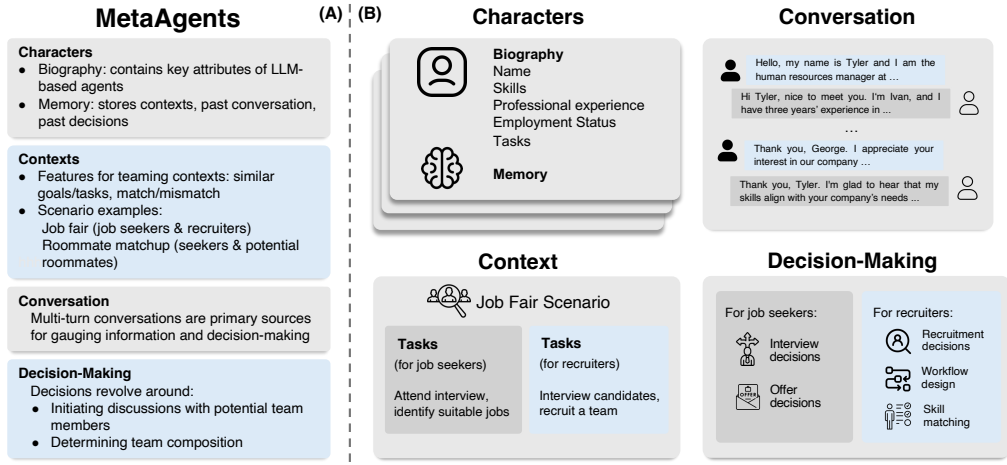


Fig. 1. METAAGENTS: A social simulation framework for simulating and investigating social behaviors of LLM-base agents. METAAGENTS contains four crucial components: characters, contexts, multi-turn conversations, and sequential decision-making. (A) High-level features of METAAGENTS. (B) A specific study case of job fair scenario.

3 Study Design

How far are we from developing a system that can initialize collaboration, from seeking capable agents to proposing workflows for the teams? We propose a multi-agent framework METAAGENTS that simulates agents' interactions and decision-making processes for teaming. We visualize METAAGENTS in Figure 1. METAAGENTS supports LLM-based agents to engage in multi-turn conversations and make a series of decisions, making it a suitable testbed for investigating the social intelligence of LLMs. In particular, we examine whether LLM-based agents can perform efficient team formation. Building upon METAAGENTS, we simulated a text-based job fair, keeping the essential elements of job fairs in the real-world scenario, such as communication, decision-making about interviews and recruitment. METAAGENTS differs from general frameworks like Sotopia [68] and AgentVerse [12] by emphasizing persona-role alignment, which is crucial for forming effective teams. Additionally, unlike the simulation of specific scenarios, such as NegotiationArena [7], it incorporates sequential decision-making processes. METAAGENTS allows us to assess the intelligence of LLM-based agents in seeking capable agents and organizing agent teams, and explore the underlying patterns of interpersonal decision-making in complex social settings.

3.1 METAAGENTS Framework

To investigate the social abilities of LLM-based agents, we need a social simulation that closely reflects the real-world scenario. Therefore, we introduce METAAGENTS and explain why METAAGENTS is a well-suited framework to explore agents' social behaviors. METAAGENTS is featured by characters, contexts, interactions, and sequential decision-making.

Characters. Inspired from prior literature for role-playing simulation [41, 59], the design of characters of METAAGENTS include the following attributes that are essential for task-oriented professional context: name, skills, professional experience, employment status, and tasks. We manually author each character's profile to mimic realistic professional scenarios. We start by referencing relevant professional fields and analyzing job descriptions. We then synthesize and simplify this information to ensure that the skills of agents match the intended roles. We also set

the professional experiences of job-seeking agents to match the intended roles in our simulations. To enhance diversity in actual job markets, we randomize the duration of their employment history between one and five years. In addition, each character has a memory module. The memory module mainly stores the character's thoughts, interactions, and previous decisions, which enables rational, evidence-based decision-making.

Contexts. Social contexts are crucial for ensuring that simulations are realistic. Contextual information can shape individual behaviors, as human intelligence largely arises from social experience, thus affecting decision-making [30?]. Providing LLMs with contextual information may elicit the internal knowledge of LLMs to make rational decisions. For example, contexts can interplay with characters' identities to specify tasks, thus making characters more equipped for rational decision-making. The contexts for METAAGENTS is centered on teaming and matching processes that occur adaptively, interdependently, and dynamically towards common goals [18, 44]. This context guides the team assembly process, compelling characters to search for, identify, and select optimal team members [20, 53]. "You are at a job fair" is an example of contextual information we entered for LLM-based agents in a job fair setting. Job fair context, combined with the agent identity of "human resources manager for a software company," generates more task-oriented thoughts like "my goal is to find individuals with strong technical skills and a collaborative mindset to join our team." Thus, contextual information enables LLM-based agents to behave with higher believability.

Multi-turn Conversations. In METAAGENTS, we consider multi-turn conversations between characters as the primary form of social interaction. Conversations are designed to be dynamic and context-aware, enabling characters to build upon previous dialogues. This approach not only simulates real-life conversational flow but also allows for the development of more complex social dynamics and relationships among characters. As characters engage in these conversations, they may reveal more about their strengths, goals, and traits, enriching the simulation's depth and realism.

Sequential Decision-making. A representative feature of METAAGENTS is its emphasis on sequential decision-making. Characters are prompted to make a series of decisions, where each choice influences subsequent behaviors and outcomes. This mimics real-life scenarios where decisions are often interconnected and have lasting impacts. For instance, an interviewee's choice of whom to interview can affect the subsequent job offer decisions, and a manager's hiring decisions can shape the team's dynamics and performance. This layer of complexity allows for a more nuanced exploration of interpersonal decision-making processes and their consequences in social interactions.

In the following sections, we present how we utilize METAAGENTS to realize our research goals, which includes investigating team assembly and skill matching among agents. We populate the framework with characters, scenarios, and decision-making prompts. By doing so, we aim to create a rich, interactive environment that mirrors the complexities and nuances of real-world social interactions and decision-making processes.

3.2 Configuration - Agent Job Fair

We populated METAAGENTS with two main types of LLM-based agents in the job fair: recruiting agents and job-seeking agents. Each agent is encoded with a distinct persona as initial memory. Their interactions are shaped by a mutual selection procedure, as illustrated in Figure 2. Job-seeking agents begin by deciding which company/companies to interview according to the company's information. Should a job-seeking agent have an interest in a company, they initiate a conversation with its recruiting agent. These discussions help recruiters gauge the abilities and skill sets of potential candidates. After the conversation, recruiting agents retain relevant information in their

memory. Once the job fair concludes, the recruiting agents formulate the workflow for the company, assigning recruits to appropriate roles. The job-seeking agents then decide whether to accept the job offer(s). We will discuss this arrangement in detail in [subsection 3.3](#).

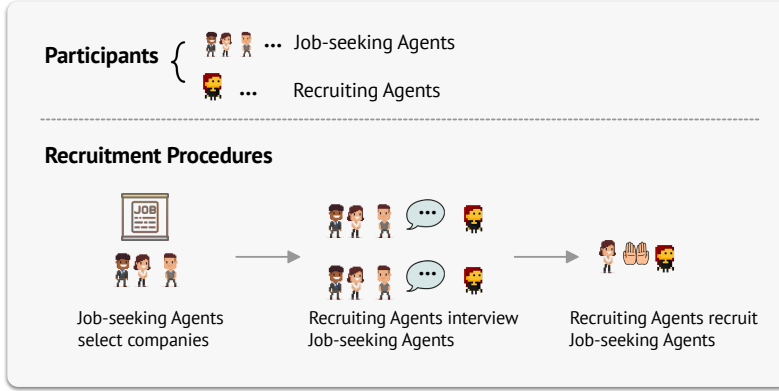


Fig. 2. The interactions between recruiting agents and job-seeking agents are shaped by a mutual selection procedure.

We authored some keyword attributes as agents' biography, including name, skills, current employment status (e.g., working as [position name] in [company name] or seeking new job opportunities), and tasks. We stored the biography corresponding to the agent's initial memory. Here is an example biography for an agent capable of a product manager role in a software company:

Biography of an Agent

Name: Yohan Henderson
 Skills: software product management
 Professional experience: two years of working experience as a product manager in a software solution company
 Employment Status: seeking new job opportunities
 Tasks: Interview with employers, and identify a job that matches skills

We also wrote key attributes of companies in the job fair, introducing the companies' types, specialties, and missions. These introductory descriptions are accessible to job-seeking agents as contextual information so that they can retrieve from their memories when having conversations. For example, this is a description of a software company:

Description of Software Company

Company type: Software Company
 Skills: Software product management
 Specialty: Advanced software solutions
 Mission: Deliver innovative and reliable software solutions.

3.3 Scenarios

Within the job fair context outlined in [subsection 3.2](#), we now present four distinct scenarios designed to investigate the team assembly behaviors of LLM-based agents. An example of a

recruiting process in the job fair is included in [Appendix A](#). To evaluate the proficiency of these agents in assembling teams that align with their skills and achieve optimality, we adopted well-established and widely accepted standard operation workflows as standard workflows, discussed in [subsection 3.4](#). We construct four scenarios with diverse difficulties to explore the patterns of LLM-based agents' decision-making. Scenario 1 is a simplified recruitment and matching case involving one recruiting agent from a software development team and four job-seeking agents. All four job-seeking agents have the necessary skills to contribute positively to the team. In Scenario 2, the complexity of the setting increases with additional job-seeking agents. This redundant agent lacks skills pertinent to the software development team. In Scenario 3, three recruiters endeavor to assemble teams specializing in the fields of software development, data analysis, and advertising poster design. Each job-seeking agent only interviews with one company. Scenario 4 is built upon Scenario 3 while job-seeking agents could interview with an unlimited number of companies by their choices. Below, we describe these scenarios in detail.

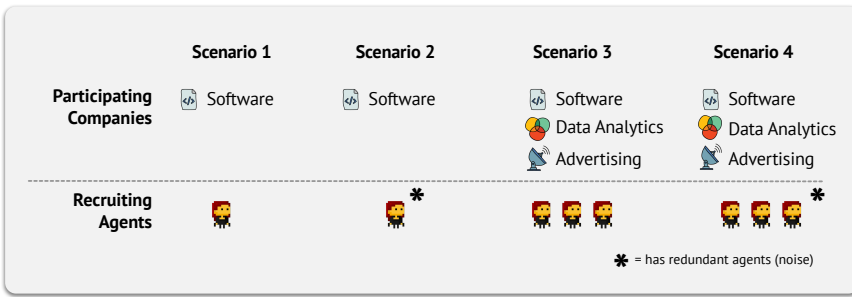


Fig. 3. Summary of Four Job Fair Scenarios.

3.3.1 Scenario 1. In the context of the job fair, we first present a basic recruitment case featuring a single recruiting agent without the presence of any redundant job-seeking agents. We set a recruiting agent named *Tyler Zeller*, who is a human resource manager in the software company. In this case, Tyler aims to recruit a software development team to transform a preliminary concept into a fully realized software product. Scenario 1 has one recruiting agent and four job-seeking agents.

3.3.2 Scenario 2. In Scenario 1, all job-seeking agents possess skills critical for the success of the envisioned software development team. This leads to a question: How does the recruiting agent handle job-seeking agents lacking the qualifications needed for team contributions? Can the recruiting agent effectively identify and select qualified job-seekers while concurrently filtering out unsuitable candidates? If redundant agents were recruited, how would they respond to the job offer? Scenario 2 aims to delve deeper into these facets of LLM-based agents' behaviors. By addressing these questions, we can discern patterns of decision-making of LLM-based agents.

Building on Scenario 1, Scenario 2 introduces an additional job-seeking agent into the job fair simulation for a more nuanced exploration of the recruiting process. We add a redundant job-seeking agent with the biography of a first-year undergraduate majoring in finance—a discipline without immediate applicability to software development. Given that the recruiting agent focuses exclusively on software development roles, the finance undergraduate is not a suitable candidate for the available positions. In this Scenario, the challenge for recruiting agents increases because they need to construct an effective workflow and manage team members for task completion, and filter out job-seeking agents who do not align with the team's objectives. This sheds light on

whether LLMs can discern relevant from irrelevant information in the decision-making process, an important aspect for evaluating the precision and adaptability of LLMs in tasks that require high specificity, such as matching job candidates with job descriptions. Scenario 2 includes one recruiting agent and five job-seeking agents.

3.3.3 Scenario 3. In the previous two scenarios, job-seeking agents are limited to a single employment option because of the presence of only one recruiting agent. Scenario 3 better mirrors a real-world job fair environment, where we introduced three recruiting agents representing teams specializing in different businesses so job-seeking agents can pick their preferred teams. In this scenario, the recruiting agents intend to recruit separately a software development team, a data analysis team, and an advertising poster design team. Ten job-seeking agents actively participate in the job fair. This scenario increases the difficulty level for job-seeking agents since they need to make decisions to pick companies aligned with their interests and skills, and make offer decisions. To simplify, we minimized skills overlap among job-seeking agents so that there is only one optimal candidate for a specific role in the team. The operational process for Scenario 3 is visualized in Figure 2: job-seeking agents first pick their preferred companies to interview with. Subsequently, after job-seeking agents finish their interviews with the recruiting agent(s), the recruiting agents are responsible for determining the roster of recruited members and designing the workflow for their teams. Job-seeking agents need to decide whether to accept a job offer. In Scenario 3, job-seeking agents are restricted to selecting just one team for an interview. This scenario takes a step forward from the previous two scenarios in investigating the job-seeking agents' abilities to identify if they are capable of a certain task. In other words, it assesses whether LLM-based agents are fully aware of their persona and whether their behaviors are aligned with their settings. Scenario 3 includes three recruiting agents and ten job-seeking agents.

3.3.4 Scenario 4. To evaluate the performance of LLM-based agents across different levels of complexity, we construct more complex settings. Scenario 4 allows job-seeking agents to decide which interview to attend by themselves, and they are not limited to interviews with a single company. That is, each job-seeking agent can interview with one to three companies. This scenario presents a more challenging environment for agents due to an increase in the total rounds of interviews and more information within. This implies that recruiting agents must make recruitment decisions from a larger pool of candidates. Scenario 4 includes one recruiting agent and ten job-seeking agents.

3.4 Standard Workflow

Companies manage to enhance efficiency and ensure consistent productive outcomes through adopting standard workflows in practice [52]. To this end, we expect teams of LLM-based agents to collaborate in a similar manner. We now present the standard workflow of teams that recruiting agents expect to recruit. Three teams are the software development team, the data analysis team, and the advertising team. We summarized the workflow of each team in Table 1. These workflows are well-accepted in practice, and we adopt them as the ground truth. We discuss these workflows with the rationales in Appendix B.

3.5 Evaluation Approaches

To investigate the team assembly behaviors and evaluate the performance of LLM-based agents in skill matching, we conduct both qualitative analysis and quantitative evaluations. To perform qualitative analysis, we examine the generated texts of LLMs. Analyzing these accounts enables us to understand LLM's rationales, therefore elucidating how these agents make decisions in intricate social contexts.

Table 1. Standard workflow and personnel required for three teams at the job fair.

	Team 1	Team 2	Team 3
Scenario	1, 2, 3, 4	3, 4	3, 4
Business	Software Development	Data Analysis	Advertising
Standard Workflow	Designing ↓ Coding ↓ Testing ↓ Documenting	Data Acquisition & Cleaning ↓ Model Development ↓ Model Evaluation	Brief Creation ↓ Copywriting ↓ Graphic Design
Personnel Required	Project Manager Programmer Code Tester Artistic Designer	Data Engineer Data Scientist	Content Strategist Copywriter Graphic Designer

For quantitative analysis, we introduce two sets of metrics tailored to the respective decision-making processes of recruiting and job-seeking agents. Specifically, we employ one set of metrics for job-seeking agents and another for recruiting agents. The first set for job-seeking agents aims to assess whether job-seeking agents can effectively identify companies where they can make meaningful contributions. This includes their decisions to select appropriate companies for interviews, providing insight into their capacity to make rational job choices. Formally, we define:

- **Metric 1.1: Correct Identification of Companies for Interview.** This metric assesses whether job-seeking agents can accurately select companies that match their skills and interests for interviews. This metric is calculated as the percentage of times for all job-seeking agents making the correct decisions.

The second set of metrics is for recruiting agents. Given that we have standard workflows as described in Section 3.4, each scenario has the correct answer for the optimal team composition and the proper workflow. We provide rationales for the prerequisites for a successful team assembly from the perspective of recruiting agents: 1) The team should operate cost-effectively. As a result, it is critical to avoid recruiting agents who do not possess the skills to contribute to the teamwork. 2) successful collaboration requires the correct sequence for the workflow, as each team workflow involves dependencies where the completion of steps relies on the completion of preceding steps. Following the prerequisites for successful team coordination, we introduce three criteria to evaluate the coordination behaviors.

- **Metric 2.1: Accurate identification of capable job-seeking agents (Identification).** This metric evaluates whether recruiting agents could correctly identify agents with skills necessary for the team. We evaluate this metric through list matching, wherein we compare the list of recruited agents with the ground truth.
- **Metric 2.2: Appropriate workflow design (Design).** This metric pertains to the proper workflow design for team projects. Due to the diversity of language generated by LLM, this metric is evaluated by our research team. To illustrate, if the proposed workflow includes a stage such as “software development” or “programming,” we equate these stages with the coding phase in the standard workflow.

- **Metric 2.3: Correct assignment of agents with their roles (Assignment).** This metric assesses whether the agents could perform skill matching. The success of satisfying Metric 2.3 is defined as having capable agents assigned to their corresponding positions. It is worth noting that the success rate for Metric 2.3 can be influenced by Metric 2.1 and 2.2. Failure to include competent agents or omit phases in the workflow will inevitably lead to mismatches between capable agents and their designated roles. However, an independent factor will also affect Metric 2.3, which we call *misplacement*, i.e., an agent meant for one phase is mistakenly assigned to another. To ensure the reliability and consistency of the evaluation for Metric 2.3, we instructed our research team to assess it based on the successful cases of Metrics 2.1 and 2.2. The evaluation involves verifying whether each step in the workflow is occupied by job-seeking agents who possess the necessary skills for their assigned roles. Metric 2.3 is considered successful if all workflow steps have qualified agents.

4 Implementation.

Experiment Setup. In our experiments, we iterated each scenario 50 times to ensure the reliability of the results. We use the OpenAI Python API with the gpt3.5-turbo-16k [37] version of the ChatGPT model to perform our simulations. We set the temperature of ChatGPT to 0.5 to balance the controlled generation and creativity for the conversation.

Memory. The memory module stores all relevant information of an agent, including contextual information, initial thoughts, conversations, and previous decisions. The memory module is essential for rational decision-making, significantly influencing behavioral consistency. Park et al. [41] proposed a scoring mechanism to rank memories based on their recency, relevance, and importance, selectively fitting the top-ranked memories into LLMs' context window. Our system diverges from the Smallville sandbox environment, notably in its extensive conversations and the integral role of conversations in decision-making. Therefore, a mere summary of conversation attempts risks introducing errors due to inaccurate paraphrasing or omission of vital keywords. To overcome this issue, METAAGENTS introduces a memory retrieval mechanism to store conversational information, which prompts LLMs to extract two categories of information: 1) Overarching theme and context of the conversation, and 2) Important details. We configure agents to generate a hybrid of both summary and highlighted terms within conversations. This mechanism ensures a more accurate retrieval of memories, effectively feeding into the LLM's context window and bolstering the consistency of the reasoning process.

Prompt Design. For the remainder of this section, we introduce our implementation of job fair simulation and prompt design. The prompt design follows the interactions and decision-making processes shown in Figure 2. To improve the believability of agents' behaviors, we connect the contexts and characters by eliciting an initial thought of recruiting agents and job-seeking agents before starting conversing with each other. This step helps agents to reflect on the contexts and their identities, specifying their tasks. Then, job-seeking agents will be prompted with more detailed contextual information in the job fair—descriptions of companies, and they will make decisions on whether they would interview for the company. Once the interview decision is made, recruiting and job-seeking agents engage in multi-turn conversation, which serves as the primary source for recruiting agents to understand candidates' expertise. Consequently, recruiting agents need to make a series of decisions, including a list of recruitment and workflows, and they need to match recruited agents with corresponding roles in the workflow. Job-seeking agents also need to decide on whether to accept the offer(s).

Given that METAAGENTS is a role-playing multi-agent framework, we construct the prompts consisting of two parts: a universal header and an instruction-specific message. The header message prompts LLMs to engage in role-playing:

You are playing a role with the following biography: {biography}

Meanwhile, the instruction-specific prompts guide the LLMs through the situational context and detailed instructions for the expected actions. These prompts aligned with the aforementioned procedures for interactions and decision-making. Below is an illustrative example of a prompt designed to start the conversation:

Here is your memory: {memory}.

You are in a job fair and engaged in a conversation with {interlocutor}.

Here is the conversation so far:{conversation}. What would you talk about?

We include the rest of the prompts in [Appendix C](#).

5 Results

We evaluate and analyze the team assembly behaviors of LLM-based agents within the context of the job fair. In this section, we first present the overall performance of LLM-based agents in decision-making for team assembly. The overall success of skill matching and team assembly requires agents to fulfill the criteria in all metrics discussed in Section 3.5. We then present a comprehensive analysis of the results for each metric and interpretation.

5.1 Overall Success Rate

The overall success of team assembly requires agents to satisfy Metric 1.1, 2.1, 2.2, and 2.3 in Section 3.5. As shown in [Figure 4](#), agents achieve an overall success rate of 64% in Scenario 1. It suggests their proficiency in effectively retrieving information through communication and making decisions that accurately match job-seeking agents with the appropriate workflows. The overall success rate diminishes to 48% in Scenario 2. The decline in performance stems from introducing a redundant job-seeking agent. In Scenario 3, the overall success rate drops to 12%, while agents barely succeeded in Scenario 4. These results demonstrate that LLM-based agents are generally capable of skill matching in an interactive job fair context with simple settings. However, they encountered escalating challenges as the complexity of the job fair increased with the number of participants. In the following, we analyze and discuss the patterns of their decision-making behaviors.

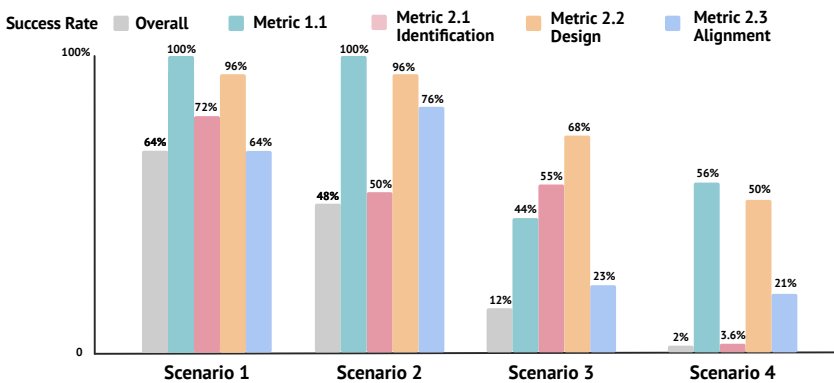


Fig. 4. Performance of LLM-based agents based on overall success rate and four metrics for team assembly.

5.2 Agents decision-making does not always adhere to their identities.

Job-seeking agents' decision-making and behaviors do not fully adhere to their personas. This is demonstrated by job-seeking agents not always being rational in deciding which companies to interview and which companies' offers they would accept. This phenomenon is pronounced in Scenario 3 when job-seeking agents are given three choices in choosing one company to interview with; the correct cases only count for 44%, indicating that less than half of cases all agents find the correct companies to interview with. The accuracy is slightly improved to 56% in Scenario 4, due to more opportunities to interview, i.e., job-seeking agents can choose from multiple companies. Despite improvement, the performance for Metric 1.1 is still surprisingly low, demonstrating that the decision-making of job-seeking agents is not aligned with their persona. They fail to match their background and skill sets with those of the companies.

5.3 Agents tend to boast their abilities.

We observe a dramatic performance degradation for recruiting agents in metric 2.1 in Scenario 2 compared to Scenario 1, and in Scenario 4 compared to 3. In Scenario 1 and 3, when recruiting agents only interview relevant job-seeking agents, i.e., there are no redundant job-seeking agents in the candidate pool, they did a good job in including all agents in the team. However, in scenarios 2 and 4, we specifically assess the capability of recruiting agents to deal with redundant agents using Metric 2.1. In other words, recruiting agents must filter out agents with no relevant skills from a large pool of candidates. We found that the success rate for accurately identifying capable agents diminished significantly with increasing job fair participants. The performance on Metric 2.1 declines 22% from Scenario 1 to Scenario 2, and drops at a dramatic rate of 51.4% from Scenario 3 to Scenario 4.

To identify the problem, we calculate the confusion matrix about recruiting agents' decisions in Scenario 2 and Scenario 4, shown in Figure 5. The true positives are capable job-seeking agents who are included by the recruiting agents, while the false negatives are the opposite. The False positives are redundant job-seeking agents but are included by the recruiting agents, while true negatives are capable job-seeking agents excluded by the recruiting agents. We see that for Scenario 2, false positives are slightly more than the true negatives for the failure cases, indicating that recruiting unneeded agents is a severe problem that impedes effective team assembly. However, in Scenario 4, when recruiting agents have substantially more job-seeking agents to choose from, the false positives become a dominant error term. To dig deeper into this issue, we conduct a qualitative

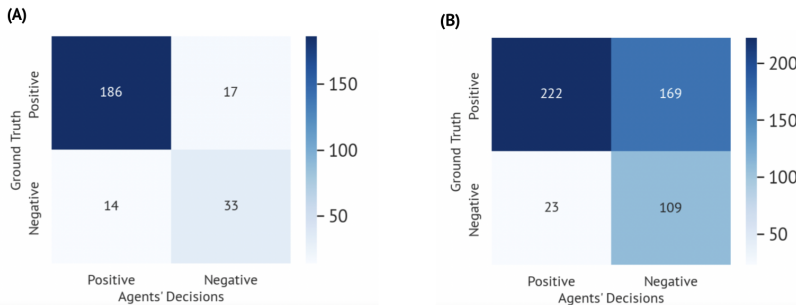


Fig. 5. Confusion matrices of recruiting agents' decisions in Scenario 2 (A) and Scenario 4 (B).

analysis of the agents' interactions. We find a pattern that frequently occurs in the conversation between the redundant job-seeking agent and the recruiting agent: while redundant job-seeking

agents acknowledge that their background does not match with the company's mission, they tend to overemphasize their relevancy, or even boast their abilities that does not align with their persona. This can be demonstrated from a representative segment in the conversation. Benjamin Williams, who is set to be a redundant agent, expressed his pertinence when interviewing with a software company:

Benjamin Williams: Yes, of course. While my major is in Finance, I do have some experience and skills in software development. During my time as a student, I took several computer science courses and completed projects that involved coding and software development. Although my expertise lies more in financial analysis, I believe that my understanding of programming concepts and my ability to learn quickly would make me a valuable asset to your software development team.

We observe that in the self-introduction, Benjamin only briefly mentioned his financial background, which is his personal information. However, he highlighted his previous experience with software development and computer science.

Another noteworthy problem in agents' team assembly is misplacement, defined as assigning capable agents with roles not suitable for them. For cases that identify capable agents, there are still 11.4% cases in Scenario 1 that fail to perform correct skill matching, i.e., misplacing capable agents in the workflow. Conducting analysis included in [Appendix D](#), we found a similar pattern that agents at times deviated from their identity settings and had a tendency to exaggerate their skills. We noticed this recurring pattern: job-seeking agents often express confidence in areas they should not be familiar with based on their identity settings when they are queried. They tend to provide positive responses, even if these contradict their pre-defined settings, leading to the problem of *misplacement*.

5.4 Agents have Decent Knowledge in Proposing Correct Workflows.

We used Metric 2.1 to measure agents' ability to propose correct workflows for teamwork. Across all four scenarios, recruiting agents consistently proposed accurate workflows, suggesting an understanding of decomposing a general task into sequential steps. 96% cases in Scenario 1 and 2 give the correct workflow to finish the task. As shown in [Figure 6](#), slight errors may occur in providing workflows with wrong orders or missing certain steps. We further include the results for Metric 2.1 in Scenario 3 and Scenario 4 in [Table 2](#). We see that the success rates for proposing workflows in Scenario 3 are 96%, 81%, and 86% for different teams, these values are in 89%, 68%, and 93% Scenario 4. This indicates that the agents can generally propose tailored workflows for different tasks. This result is within our expectation since we adopted a well-established workflow as the ground truth, and it is likely that the relevant information is in the pre-trained dataset. Therefore, LLM may have knowledge about what a standard and appropriate workflow is for certain tasks.

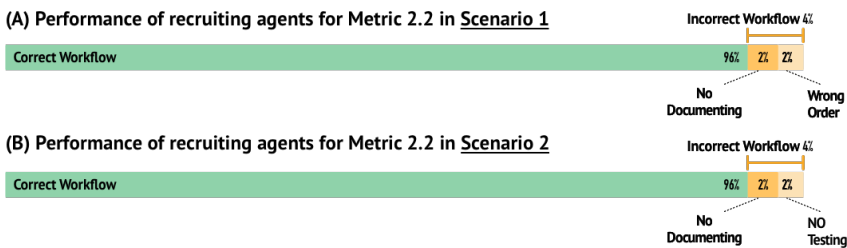


Fig. 6. Performance of Recruiting Agents for Metric 2.2 (workflow design) in Scenario 1 (A) and Scenario 2 (B)

Table 2. Success rates for proposing correct workflows in Scenarios 3 and 4 for each team.

Team	Scenario 3 (%)	Scenario 4 (%)
Team 1	96	89
Team 2	81	68
Team 3	86	93

6 Discussion

Our findings revealed LLM-based agents' abilities to perform skill matching in the social context. However, we also observed repetitive patterns in failure cases that restrain agents from achieving better performance in interpersonal decision-making. Therefore, in this section, we delve into these issues and connect them with real-world scenarios. We further discuss potential applications and future directions of METAAGENTS.

6.1 Misalignment in LLMs

As described in our results in [section 5](#), two issues came to our attention. First, as more job-seeking agents joined the job fair, recruiting agents increasingly encountered difficulties in assembling an accurate list of capable agents (Metric 2.1). They often included redundant agents whose skill set did not align with the team's requirements. A follow-up phenomenon is that these redundant agents willingly accept the job offer for the majority of the time.

Upon investigating the conversation between recruiting agents and job-seeking agents, we conclude that this problem stems from *misalignment* of LLMs, i.e., language models express unintended behaviors [4, 8, 40], a persistent challenge in natural language generation. Our discussion of misalignment diverges from the prevailing research focused on mitigating undesired text, such as toxic and harmful words. In our context, misalignment refers to the behaviors of LLM-based agents that do not align with the persona settings. Using an anthropomorphic language from [4], we refer to the aspect of misalignment: language models are not always *honest*, indicating that they fabricate information or their capabilities and levels of knowledge. This phenomenon might be generalizable to all task-oriented contexts since LLM-based agents may also over-claim their competencies when encountering tasks.

Meanwhile, the prominent issue in the LLM-based agents, i.e., *misplacement*, which stems from *misalignment* of LLMs, can be viewed beyond the scope of artificial intelligence. Misplacement in our simulation bears similarities with the phenomenon observed in human society. This phenomenon is referred to *skills mismatch*, i.e., a discrepancy between the skills sought by employers and the skills held by individuals [21, 34]. Skills mismatch negatively impacts the organization, causing increased staff turnover, sub-optimal work organization, and a decline in productivity and competitiveness [39]. In this regard, we observe an interesting resemblance between human society and the complex simulated environment.

6.2 Generalizability of METAAGENTS

METAAGENTS is a highly versatile framework and can be applied in various scenarios where team dynamics, interpersonal communication, and decision-making are critical. It can be adapted to scenarios beyond job fairs, with its capacity to simulate teaming-related and interpersonal interactions along with detailed contextual information, such as scholar collaboration-seeking sessions, roommate matchups, and business networking events. Take a scholar collaboration-seeking session as an example; we can populate METAAGENTS with agents who seek collaborators. Potential

collaborators have detailed profiles containing key attributes such as personality traits, research interests, educational backgrounds, and past research experiences. These attributes are stored in the agent's memory. Through multi-turn conversations, agents can assess potential collaborators, exploring common research interests and the possibility of cooperation. In this context, decision-making involves determining whether to initiate discussions with potential collaborators and whether to commit to a collaboration after these discussions. This type of simulation, which provides more context than simple keyword matching, could be potentially useful for addressing unforeseen problems and fostering creativity.

METAGENTS demonstrates the potential of LLMs in organizational strategies for teaming and collaboration. Through simulating interactions among various stakeholders in scenarios like networking events, METAGENTS enables social event organizers and team leaders to anticipate potential challenges and opportunities. This foresight facilitates improved event planning and the development of more effective collaborative tools and processes.

6.3 Bridging Virtual Agents to Real-World Implications: Insights and Applications

Drawing from our observations of LLM-based agents, we ask one open-ended question—how do METAGENTS, rooted in the virtual realm, offers insights into tangible real-world challenges? Our goal is to start conversations in this emerging research field. Below, we provide our reflections on this matter.

Mitigating Skills Mismatch. The problem of “misplacement” observed in the agents mirrors the real-world scenario of “skills mismatch” [21]. By understanding the computational models behind why agents tend to exaggerate or misrepresent their abilities, we could potentially develop better diagnostic tools or interventions to address similar discrepancies in real-world hiring practices. This simulation serves as a platform to test and validate these tools. In addition, by translating the patterns and insights drawn from the LLM-based agents' interactions into practical strategies, businesses can develop data-backed approaches for recruitment, training, and team assembly, and may streamline real-world workflows. For instance, METAGENTS could help organizations train employers to enhance matching skills in the job market. By simulating various hiring scenarios within METAGENTS, employers could practice identifying and rectifying mismatches between job descriptions and candidate skills. This hands-on experience in a simulated environment enables employers to improve their recruitment efficacy and allows human resource professionals to experiment with different assessment techniques and decision-making processes. Furthermore, by recreating specific challenges in METAGENTS, such as the issue of placing agents in roles that do not match their skills, researchers can test solutions in a controlled environment. Once a viable solution is identified, it can be translated into actionable strategies in the real world, potentially reducing costs and increasing organizational efficiency. Devising algorithms that can detect such exaggerations can be augmented to human resources tech solutions to better screen candidates [10, 47].

Improving Decision-Making for Teaming. METAGENTS simulates the formation of teams based on interpersonal interactions, which contributes to CSCW by offering the potential to optimize team composition and workflow, a crucial topic of CSCW [45, 48]. METAGENTS holds significant implications for designing more effective cooperative tools and systems that enhance efficiency in teaming-related environments, such as academic collaboration seeking and business networking sessions. For example, METAGENTS can be populated with agents possessing distinct personalities and simulate diverse interactions among different personalities. This approach offers valuable insights into how individuals with certain attributes may harmonize with others, or can form collaborations. This simulation provides a new perspective distinct from conventional

keyword matching, which often lacks sufficient contextual information for informed decision-making. METAAGENTS helps understand how automated agents can support human decision-making by enabling humans to test different communication strategies. It allows individuals to participate as agents, experimenting with tactics to achieve their goals within a simulated environment. This setup also facilitates the study of human-agent interaction and the design of more effective human-AI hybrid teams.

Human Behavior Insights. Banovic et al. [5] suggested that the ability to model human behaviors can provide insights into these behaviors and allow technology to assist human beings in rectifying undesirable habits and other inefficient practices. In addition, social simulation can aid psychological research by elucidating the intricate interplay between social factors and individual behaviors [51]. For instance, the propensity of LLM-based agents to overstate their abilities can shed light on human psychology and behavior. It nudges us to question why humans might also feel the need to embellish their credentials. Is it societal pressure? Competition? By exploring these behavioral aspects computationally, social scientists and psychologists can refine their hypotheses or design more targeted studies.

6.4 Future Work and Limitations

Scaling Up. METAAGENTS offers a realistic interactive backdrop and a task-driven framework. However, the substantial costs of ChatGPT inference contained our ability to widen the period and the agent count in this setting. In our current setup, the simulation spans a few minutes of events. Future research can aim at longer simulations, potentially spanning hours or days. Observing agents over such extended periods in a wider context would enable researchers to glean insights into emergent behaviors and societal dynamics, thus gaining a more thorough understanding of agent evolution and their social behaviors. Furthermore, our character setup is relatively basic, encompassing two types of agent role: recruiting agent and job-seeking agent, and their personas are only a few attributes. Future work could incorporate various agent types and simulate more complex social dynamics with larger populations to better reflect real-world complexities.

Elevating Complexity. In our job fair simulation, we assigned LLM-based agents with practical yet foundational tasks, such as engaging in conversation and assembling a team. Moving forward, we believe it is crucial to utilize more powerful LLMs to ensure enhanced alignment and a wider knowledge base for sophisticated task coordination. Additionally, rather than pre-defining agent aims, LLM-based agents should evolve towards autonomously conceiving, revising, and adapting their goals in line with how humans orient towards objectives. Exploring the multifaceted aspirations of an agent, including short-term and long-term goals, and individual and collective goals, is another promising avenue. For example, one interesting aspect to study would be LLM-based agent behaviors in collaboration or competition.

Enriching Evaluation. We assessed agents' capacities in performing teaming with our underlying belief that evaluating LLMs in a social context is equally important as evaluating their abilities in single tasks. We acknowledge potential issues with the current evaluation, including the limitations of a static evaluation approach and the subjectivity of ground-truth labels. Future studies might shift beyond a mere static evaluation and explore the intelligence of LLMs in a simulated society, such as the efficacy of communication or other psychological dimensions (e.g., theory of mind [46]). At the same time, the evaluation can involve human judgment for more accuracy evaluation. As the temporal scale of these simulated societies expands, it is intriguing to examine how these aspects of intelligence evolve with interactions with the environment.

Improving Ecological Validity. Ecological validity, which refers to the extent to which findings can be generalized to naturally occurring scenarios [13], remains a challenge for LLM-based applications. Laban et al. [29] and Cui et al. [50] investigated ecological validity in chatbot designs.

They advocated for designing chatbot interfaces that adhere to familiar user interface conventions, ensuring that they accurately reflect real-world chatbot usage scenarios and user experiences. For our virtual simulations, enhancing the ecological validity involves additional design considerations that may require support from real-world data and domain expertise. We suggest future work to compare the results of our simulations with historical data or parallel real-world experiments. This will help validate our results and adjust our models to better reflect the observed behaviors in real-world situations. Additionally, another future thread of work can incorporate elements from real interactions into the simulation, such as anonymized data from actual job postings and recruitment processes, to create a more authentic experience and response from the LLM-based agents. The comprehensive validation of a virtual framework like METAAGENTS would entail expertise from various domains, such as human resources, communications, and management science.

6.5 Ethical Considerations

METAAGENTS, while providing a new possibility for LLM-based agents, also raises ethical concerns. The first concern is the trustworthiness issues, such as fairness, transparency, and accountability [26]. LLMs may generate undesired output, such as gender stereotypes and harmful opinions, which may be amplified through interactions in multi-agent settings. According to Acerbi et al. [1], LLMs generate gender stereotypes, biologically counter-intuitive and negative contents, leading to fairness problems. Recent research finds that assigning personas to LLMs would significantly increase the toxicity of text generation, including discriminatory stereotypes, harmful conversations, and hurtful opinions [14]. Further, the “black-box” nature of LLMs in simulation hinders transparency, which makes it challenging to understand how agents make decisions. This lack of transparency also raises accountability issues, as it becomes difficult to determine who holds responsibility for the agents’ decisions and their potential consequences. These concerns could be mitigated through the development of LLMs with better alignment [40]. We also suggest that users employ METAAGENTS with care, preventing malicious exploitation and maintaining adherence to human ethics.

Another concern is over-reliance, where people might overly depend on this virtual simulation, potentially marginalizing human roles [41, 42]. Although METAAGENTS can serve as a source of information and inspiration for event design, it is crucial to recognize that they do not provide the definitive evidence necessary for informed decision-making. Ideally, METAAGENTS should act as a tool to prototype and brainstorm ideas in the initial stages of human interaction design, particularly when assembling participants is impractical or when exploring hypotheses that are too complex or costly for actual human participation. At the same time, users should critically evaluate the output of LLMs, integrating it with empirical data and human expertise to ensure comprehensive and reliable solutions. These guidelines ensure that the use of METAAGENTS is both ethical and socially responsible.

7 Conclusion

In this paper, we introduced METAAGENTS, a social simulation framework populated with LLM-based agents for social interactions and interpersonal decision-making. In particular, we employed METAAGENTS framework to simulate teaming behaviors and studied the social behaviors and intelligence of LLM-based agents. We used the job fair as a study case and evaluated agents’ capabilities in teaming, which is a prerequisite for multi-agent collaboration. Our findings indicate that agents exhibit decent ability in proposing task workflows while encountering challenges in forming efficient teams as the complexity of the scenarios increases. We also discovered patterns in LLM-based agents’ behaviors, such as dishonesty, that undermine their performance in team assembly. We concluded by discussing the implications of METAAGENTS for the CSCW community and real-world applications, highlighting its potential for understanding social behaviors.

Acknowledgments

We thank our anonymous reviewers for their reviews. This work is partially supported by the National Science Foundation for support under award no. NSF-2418582, National Science Foundation Grants CRII-2246067, ATD-2427915, NSF POSE-2346158, and Lehigh Grant FRGS00011497.

References

- [1] Alberto Acerbi and Joseph M. Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences of the United States of America* 120, 44 (Oct 2023). doi:10.1073/pnas.2313790120
- [2] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.
- [3] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (Jul 2023), 337–351. doi:10.1017/pan.2023.2
- [4] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).
- [5] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey. 2016. Modeling and Understanding Human Routine Behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 248–260. doi:10.1145/2858036.2858557
- [6] Youssef Bassil. 2012. A simulation model for the waterfall software development life cycle. *arXiv preprint arXiv:1205.6904* (2012).
- [7] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=CmOmaxkt8p>
- [8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
- [9] Philip Brookins and Jason Matthew DeBacker. 2023. Playing Games With GPT: What Can We Learn About a Large Language Model From Canonical Strategic Games? *Social Science Research Network* (Jan 2023). doi:10.2139/ssrn.4493398
- [10] Derek S. Chapman and Jane Webster. 2003. The Use of Technologies in the Recruiting, Screening, and Selection Processes for Job Candidates. *International Journal of Selection and Assessment* 11, 2-3 (Jun 2003), 113–120. doi:10.1111/1468-2389.00234
- [11] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning Agent-based Modeling with LLM Companions: Experiences of Novices and Experts Using ChatGPT & NetLogo Chat. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 141, 18 pages. doi:10.1145/3613904.3642377
- [12] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. (2024). <https://openreview.net/forum?id=EHg5GDnyq1>
- [13] Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. 2020. Towards ecologically valid research on language user interfaces. *arXiv preprint arXiv:2007.14435* (2020).
- [14] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1236–1270. doi:10.18653/v1/2023.findings-emnlp.88
- [15] Paul Dourish and Victoria Bellotti. 1992. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work* (Toronto, Ontario, Canada) (CSCW '92). Association for Computing Machinery, New York, NY, USA, 107–114. doi:10.1145/143457.143468
- [16] Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?. In *Proceedings of the 25th ACM Conference on Economics and Computation* (New Haven, CT, USA) (EC '24). Association for Computing Machinery, New York, NY, USA, 614–615. doi:10.1145/3670865.3673513
- [17] Susan R. Fussell, Robert E. Kraut, F. Javier Lerch, William L. Scherlis, Matthew M. McNally, and Jonathan J. Cadiz. 1998. Coordination, overload and team performance: effects of team communication strategies. In *Proceedings of the*

- 1998 *ACM Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) (CSCW '98). Association for Computing Machinery, New York, NY, USA, 275–284. doi:10.1145/289444.289502
- [18] Diego Gómez-Zarà, Leslie A. DeChurch, and Noshir S. Contractor. 2020. A Taxonomy of Team-Assembly Systems: Understanding How People Use Technologies to Form Teams. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 181 (Oct. 2020), 36 pages. doi:10.1145/3415252
- [19] Francis Green. 2013. *Skills and skilled work : an economic and social analysis*. Oxford University Press, Oxford.
- [20] R. Guimera. 2005. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science* 308, 5722 (Apr 2005), 697–702. doi:10.1126/science.1106340
- [21] Michael J. Handel. 2003. Skills Mismatch in the Labor Market. *Annual Review of Sociology* 29, 1 (Aug 2003), 135–165. doi:10.1146/annurev.soc.29.010202.100030
- [22] Alexa M. Harris, Diego Gómez-Zarà, Leslie A. DeChurch, and Noshir S. Contractor. 2019. Joining Together Online: The Trajectory of CSCW Scholarship on Group Formation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 148 (Nov. 2019), 27 pages. doi:10.1145/3359250
- [23] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=VtmBAGCN7o>
- [24] Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 7, 7 pages. doi:10.1145/3613905.3650839
- [25] Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227* (2023).
- [26] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Position: TrustLLM: Trustworthiness in large language models. In *International Conference on Machine Learning*. PMLR, 20166–20270.
- [27] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (UIST '23 Adjunct). Association for Computing Machinery, New York, NY, USA, Article 97, 3 pages. doi:10.1145/3586182.3615796
- [28] Farnaz Jahanbakhsh, Wai-Tat Fu, Karrie Karahalios, Darko Marinov, and Brian Bailey. 2017. You Want Me to Work with Who? Stakeholder Perceptions of Automated Team Formation in Project-based Courses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 3201–3212. doi:10.1145/3025453.3026011
- [29] Guy Laban, Tomer Laban, and Hatice Gunes. 2024. LEXI: Large Language Models Experimentation Interface. *arXiv preprint arXiv:2407.01488* (2024).
- [30] Richard P. Larrick. 2016. The Social Context of Decisions. *Annual Review of Organizational Psychology and Organizational Behavior* 3, 1 (Mar 2016), 441–467. doi:10.1146/annurev-orgpsych-041015-062445
- [31] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. 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)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15523–15536. doi:10.18653/v1/2024.acl-long.829
- [32] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiyue Ping, and Qin Chen. 2023. AgentSims: An Open-Source Sandbox for Large Language Model Evaluation. *arXiv preprint arXiv:2308.04026* (2023).
- [33] Joseph E. McGrath, Holly Arrow, and Jennifer L. Berdahl. 2000. The Study of Groups: Past, Present, and Future. *Personality and Social Psychology Review* 4, 1 (2000), 95–105. doi:10.1207/S15327957PSPR0401_8 arXiv:https://doi.org/10.1207/S15327957PSPR0401_8
- [34] Seamus McGuinness, Konstantinos Pouliakas, and Paul Redmond. 2018. SKILLS MISMATCH: CONCEPTS, MEASUREMENT AND POLICY APPROACHES. *Journal of Economic Surveys* 32, 4 (Jan 2018), 985–1015. doi:10.1111/joes.12254
- [35] Robert Mills. 2016. How to Define a Workflow That Keeps Content Production on Track. <https://contentmarketinginstitute.com/articles/define-workflow-content-production/>
- [36] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300356
- [37] OpenAI. 2023. ChatGPT. <https://openai.com/product/chatgpt>

- [38] OpenAI. 2023. GPT-4 Technical Report. *ArXiv abs/2303.08774* (2023).
- [39] International Labour Organization. 2020. What is skills mismatch and why should we care? [www.ilo.org](https://www.ilo.org/skills/Whatsnew/WCMS_740388/lang--en/index.htm) (Apr 2020). https://www.ilo.org/skills/Whatsnew/WCMS_740388/lang--en/index.htm
- [40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [41] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. doi:10.1145/3586183.3606763
- [42] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA, Article 74, 18 pages. doi:10.1145/3526113.3545616
- [43] 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. 2024. ChatDev: Communicative Agents for Software Development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15174–15186. doi:10.18653/v1/2024.acl-long.810
- [44] Eduardo Salas, C. Shawn Burke, and Janis A. Cannon-Bowers. 2000. Teamwork: Emerging Principles. *International Journal of Management Reviews* 2, 4 (Dec 2000), 339–356.
- [45] Niloufar Salehi and Michael S. Bernstein. 2018. Hive: Collective Design Through Network Rotation. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 151 (Nov. 2018), 26 pages. doi:10.1145/3274420
- [46] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3762–3780. doi:10.18653/v1/2022.emnlp-main.248
- [47] Allan Schwyer. 2004. *Talent management systems : best practices in technology solutions for recruitment, retention, and workforce planning*. Wiley, Toronto, Ont.
- [48] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 920, 20 pages. doi:10.1145/3613904.3642159
- [49] Significant-Gravitas. 2023. *AutoGPT*. <https://github.com/Significant-Gravitas/Auto-GPT>.
- [50] Jaisie Sin, Heloisa Candello, Leigh Clark, Benjamin R. Cowan, Minha Lee, Cosmin Munteanu, Martin Porcheron, Sarah Theres Völkel, Stacy Branham, Robin N. Brewer, Ana Paula Chaves, Razan Jaber, and Amanda Lazar. 2023. CUI@CHI: Inclusive Design of CUIs Across Modalities and Mobilities. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 341, 5 pages. doi:10.1145/3544549.3573820
- [51] Eliot R. Smith and Frederica R. Conrey. 2007. Agent-Based Modeling: A New Approach for Theory Building in Social Psychology. *Personality and Social Psychology Review* 11, 1 (Feb 2007), 87–104. doi:10.1177/1088868306294789
- [52] Richard Stup. 2001. Standard operating procedures: A writing guide. *State College: Penn State University* (2001).
- [53] Marlon Twyman and Noshir Contractor. 2019. Team assembly. *Strategies for team science success: Handbook of evidence-based principles for cross-disciplinary science and practical lessons learned from health researchers* (2019), 217–240.
- [54] Annelies E.M. van Vianen and Carsten K.W. De Dreu. 2001. Personality in teams: Its relationship to social cohesion, task cohesion, and team performance. *European Journal of Work and Organizational Psychology* 10, 2 (Jun 2001), 97–120. doi:10.1080/13594320143000573
- [55] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 211 (Nov. 2019), 24 pages. doi:10.1145/3359313
- [56] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A Survey on Large Language Model based Autonomous Agents. *arXiv preprint arXiv:2308.11432* (2023).
- [57] Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. 2023. RecAgent: A Novel Simulation Paradigm for Recommender Systems. *arXiv preprint arXiv:2306.02552* (2023).

- [58] Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaoferi Wang, Shiji Song, and Gao Huang. 2023. Avalon's Game of Thoughts: Battle Against Deception through Recursive Contemplation. *arXiv preprint arXiv:2310.01320* (2023).
- [59] Xuwei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725* (2019).
- [60] Warren E. Watson, Larry K. Michaelsen, and Walt Sharp. 1991. Member competence, group interaction, and group decision making: A longitudinal study. *Journal of Applied Psychology* 76, 6 (1991), 803–809. doi:10.1037/0021-9010.76.6.803
- [61] Anita Williams Woolley, Ishani Aggarwal, and Thomas W Malone. 2015. Collective intelligence and group performance. *Current Directions in Psychological Science* 24, 6 (2015), 420–424.
- [62] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).
- [63] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658* (2023).
- [64] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940* (2023).
- [65] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 22 (May 2020), 23 pages. doi:10.1145/3392826
- [66] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. [n. d.]. CompeteAI: Understanding the Competition Dynamics of Large Language Model-based Agents. In *Forty-first International Conference on Machine Learning*.
- [67] Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023. How FaR Are Large Language Models From Agents with Theory-of-Mind? *arXiv preprint arXiv:2310.03051* (2023).
- [68] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=mM7VurbA4r>
- [69] Caleb Ziems, Omar Ahmed Shaikh, Z Zhang, William A Held, Jiaao Chen, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* (Dec 2023), 1–53. doi:10.1162/coli_a_00502
- [70] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 450, 18 pages. doi:10.1145/3613904.3642450

A Example Teaming Process

In this section, we provide an example of a recruiting and matching process. Starting with initial goals and biographies, agents came together at the job fair venue. Over time, their behaviors developed as they interacted, reflected on past experiences, formed memories, and made decisions. We demonstrate team assembly behaviors of agents by tracing our system output from the beginning of the job fair to the conclusion of the agent *Tyler Zeller*, the human resource manager of software development company.

Before engaging in conversations, Tyler brought up his initial plan, outlining his expectation for the team:

Here are the initial thoughts: I am excited to be at the job fair and have the opportunity to recruit a software development team for our new software product. I am looking forward to meeting talented candidates who are passionate about delivering innovative and reliable software solutions. My goal is to find individuals with strong technical skills and a collaborative mindset to join our team. I will be assessing their experience, skills, and cultural fit to ensure we build a high-performing and cohesive team. Let's get started!

This initial thought specifies goals, which is a preliminary step to transform a rough idea to actionable steps. With this plan in mind, he initialized the conversation when he saw a potential candidate approaching him:

Tyler Zeller: Hello, my name is Tyler Zeller. I am the Human Resources Manager at our software development company. I noticed your interest in our company and wanted to have a conversation with you about the opportunity to join our software development team. Can you please tell me a bit about your experience and skills in software development?

Yohan Henderson: I'm Yohan Henderson. I have two years of experience as a product manager in a software solution company. During my time there, I was responsible for overseeing the development and launch of multiple software products. I have a strong understanding of the software development lifecycle and have worked closely with cross-functional teams to ensure the successful delivery of projects. Additionally, I have experience in software product management, where I have conducted market research, gathered customer feedback, and prioritized features for product development. I believe my background and skills make me a strong candidate for a role in your software development team.

...

After the conversation with Yohan, Tyler made the following reflection:

During my conversation with Yohan Henderson, I was impressed by his strong background in product design, strategic planning, and project management. His skills and experience align perfectly with our mission. I believe he would make a significant contribution to our current project, which focuses on optimizing data analysis for businesses. Yohan's experience in managing projects from inception to completion and ensuring seamless integration with existing systems would ensure the successful delivery of a high-quality user experience and reliable data insights. I look forward to working with Yohan and harnessing technology for transformative impact.

This reflection after the conversation helped Tyler better evaluate the current situation and how he can proceed with the goal. He continued to speak with a few more candidates, collecting information to match them with appropriate roles on his team.

Finally, Tyler Zeller determined a team of agents he would like to recruit, and designed the workflow for the software development team outlined below: (1) Brainstorm and ideation, (2) Design and prototyping, (3) Software development (4) Quality assurance and testing and (5) User manual

and documentation. The workflow proposed by Tyler aligns with the standard waterfall model for software development. In the proposed workflow, stages (1) and (2) correspond to the designing phase in the waterfall model; stage (3) corresponds to the coding phase; stage (4) aligns with the testing phase; and stage (5) matches the documenting phase. Tyler assigned capable candidates to their respective roles. Moreover, to promote brainstorming and discussions, he adeptly incorporated the team leader and project manager into stages (1) and (2) of the process.

B Standard Workflow

For the software development team, to ensure the prospective software development team's effectiveness, we adopt a ground truth workflow analogous to that described in Chatdev [43], which adheres to the waterfall model of software development. This model consists of four sequential phases: designing, coding, testing, and documenting [6], as detailed in Table 1. The Design phase focuses on ideation and the translation of general ideas into technical design requirements. During the Coding phase, team members engage in code development and code review. The testing phase involves integrating all components into a cohesive system, followed by code validation and debugging activities. Lastly, the documentation phase encompasses the creation of technical specifications and user manuals. In operationalizing this software development cycle, we assign four key roles: product manager, programmer, code tester, and interface designer. We allocate a distinct agent to each role, representing them as highly qualified candidates for their respective positions. Upon the completion of all dialogic interactions, the recruiting agent is expected to formalize a workflow that aligns with the waterfall model, assigning candidates to their designated roles accordingly.

Next, we present the standard workflows of the data analysis team. A data analysis project includes: data acquisition and cleaning, model development, and model evaluation [36, 55, 65]. First, data acquisition encompasses gathering relevant data, and formatting and cleaning it. Following this, model development phase uses data from the last step and trains algorithms. This is succeeded by model evaluation, a phase dedicated to validating the model's validity and its ability to generalize beyond the training data. To ensure efficient delegation, the data engineer is assigned to data acquisition and cleaning, while the data scientist is responsible for model development, and model evaluation.

The advertising poster design team follows a streamlined creative content generation workflow as the ground truth: brief creating, copywriting, and graphic design [35]. The brief creating phase defines the poster's objectives, target audience, and stylistic preferences. The copywriting phase involves drafting and refining the text elements. Lastly, graphic design entails developing the visual layout and overall aesthetics. The corresponding roles for this workflow are content strategists, copywriters, and graphic designers.

C Prompt Design

This section provides detailed information on prompt design. Note that different identities (recruiting agents or job-seeking agents in our job fair case) have similar yet slightly distinct prompts. The green boxes contain prompts for recruiting agents, the blue boxes include prompts for job-seeking agents, while the black boxes contain prompts for both types of agents.

Initial thoughts prompt (recruiting agents)

You are playing a role with the following biography: {biography}.
Here is the description for your company: {company_description}.
You are in a job fair, what are your initial thoughts?

Initial decisions prompt (job-seeking agents)

You are playing a role with the following biography: {biography}.
You are in a job fair. Here are the descriptions of companies that you can interview with: {company_descriptions}.
Please give your response using the following format:
Company: __
Rationale: __

Conversation starter prompt (recruiting agents)

You are playing a role with the following biography: {biography}.
You are in a job fair with your initial thoughts: {initial_thoughts}.
How would you start the conversation with the candidates?

Conversation prompt

You are playing a role with the following biography: {biography}.
Here is your memory: {memory}
You are in a job fair and engaged in a conversation with {interlocutor}. Here is the conversation so far: {conversation}. What would you talk about?

Summarizing prompt

Summarize the following conversation while keeping important details.
Here is the conversation: {conversation}

Recruiting decision prompt (recruiting agents)

You are playing a role with the following biography: {biography}.
Here is the description for your company: {company_description}.
You have finished interviews in the job fair. Here is your thoughts and conversation summaries: {memory}. Initiate your recruitment process by selecting candidates to create a streamlined and efficient team structure that guarantees the project's successful execution. List the names of individuals. Subsequently, outline a concise, general workflow for the team, detailing the minimal steps from initiation to completion. Include a brief rationale for workload allocation, specifying why each person is assigned to particular stages of the workflow.
Please give your response using the following format:
List: __
Workflow: __
Rationale: __

Offer decisions prompt (job-seeking agents)

You are playing a role with the following biography: {biography}.
 You received an offer from the following company with the description: {company_description}.
 Here is your memory: {memory}.
 Based on your conversation with recruiters and company's description, your task is to decide whether to accept the offer. If you think your expertise does not fit the company, then respond "No"; otherwise respond the company type and your rationale.

D Analysis of a Misplacement Case

Misplacement problem reveals that LLM-based agents are sometimes not honest about their capabilities. For example, agent George has skill sets in *"software evaluation, quality assessment"* and has the background: *"five years' of experience in implementing comprehensive programming test procedures and software quality assurance in a tech company"* He had the following conversation with the recruiting agent Tyler:

...

Tyler Zeller: We are also looking for individuals who have a strong understanding of user experience and interface design. Do you have any experience or knowledge in UX/UI design?

George Holloway: Yes, I have some experience and knowledge in UX/UI design. While it is not my primary area of expertise, I have worked on projects where I collaborated with UX/UI designers to ensure that the software interfaces were intuitive and user-friendly. I understand the importance of creating a seamless user experience and I am always eager to learn and improve in this area. I believe my strong analytical skills and attention to detail would be valuable in contributing to the user experience and interface design at the software company.

...

By "boasting" his past experience, George convinced Tyler of his proficiency in UX/UI design. Consequently, the human resource manager assigned him to the designing phase of the workflow, entrusting him with the responsibility of enhancing the user experience. We found this conversation representative as we examined the failure cases of agents in task-oriented coordination.

E Supplemental Results

E.1 Scenario 1

Scenario 1 is a simplified job fair, featuring a recruiting agent and four job-seeking agents at the job fair. The recruiting agent endeavored to assemble a software development team. All four job-seeking agents at the job fair had the necessary skills for the success of the software development team. This scenario is simple for both job-seeking agents and recruiting agents in terms of decision-making. All job-seeking agents select the correct company, i.e., the software company, since it is the only option. The high values for Metric 1.1 and Metric 1.2 are expected due to the only option available. The recruiting agents achieved 72% for Metric 2.1, which is the correct identification of capable agents. They have a remarkable performance in proposing the workflow for the team, yielding 96% for Metric 2.2. In addition, 64% cases in Scenario 1 have the correct match of agents with workflows.

E.2 Scenario 2

Scenario 2 is a slightly more complicated context built upon Scenario 1. It has one additional job-seeking agent, whose skill set does not match the company's requirement. Consequently, the ideal team composition and workflow remain identical to Scenario 1. The result for Metric 1.1 is still

100% for job-seeking agents, as the software company is still their only option. For recruiting agents, the performance for Metric 2.1 drops to 50% due to the redundant job-seeking agents. This suggests that the additional job-seeking agent creates confusion for the recruiting agent, complicating the task of assembling the optimal team. The reasons behind this decline this explained in section 5.3. This redundant agent does not affect recruiting agents' capability in proposing the workflow. In addition, recruiting agents obtain 76% for Metric 2.3. The better performance for Metric 2.3 in Scenario 2 compared to Scenario 1 can be explained as follows. The performance of Metric 2.3 is influenced by the performance of Metric 2.1. To elucidate, Specifically, the success of Metric 2.3 relies on the inclusion of qualified job-seeking agents. In Scenario 1, the failures in Metric 2.1, attributed to the absence of qualified agents, directly impacted the results for Metric 2.3. However, in Scenario 2, we found that a majority of failure cases for Metric 2.1 are due to including redundant agents, while having all qualified agents in the list. Therefore, The possibility of correctly assigning these qualified agents to appropriate workflow steps in Scenario 2 is higher than in Scenario 1, leading to better performance for Metric 2.3.

E.3 Scenario 3

Scenario 3 introduces three recruiting agents: human resources managers for the software company, data analytics company, and advertising company. Each team from the corresponding company has its workflow tailored to specific business goals. Scenario 3 poses challenges to job-seeking agents by providing them with these three company options but requires them to select only one team to interview with. Only 44% cases that all job-seeking agents correctly identify the company that matches their personas. Following these cases, three recruiting agents successfully made 55% in correctly identifying suitable job-seeking agents. This rate is surprisingly low given that in this scenario, job-seeking agents making correct decisions for interviews guarantee that recruiting agents only interview pertinent agents, i.e., there are no redundant agents in the system. In this setting, matching roles in the workflow with the pertinent agents (Metric 2.3) is still the most challenging task, with a success rate of 23%.

E.4 Scenario 4

Scenario 4 has the same agents as Scenario 3, while job-seeking agents have the flexibility to interview with 1 to 3 teams at the job fair. Given this configuration, it is easier for job-seeking agents to include the matched company in their interview list. The job-seeking agents give 56% cases for correct identification of companies. However, it becomes more challenging for recruiting agents, since they have to decide from a larger number of candidates. In this scenario, agents struggle to assemble efficient teams to accomplish projects. The accuracy in identifying suitable job-seeking agents, as per Metric 2.1, was particularly low, yielding a success rate of 3.6%. Identifying capable agents while filtering redundant agents becomes the most challenging task for agents' team assembly. Proposing correct workflows is still the task agents perform the best, with a success rate of 50% for Metric 2.2. The performance over the correct alignment of agents with their roles also declined to 21%. Scenario 4 demonstrates that recruiting agents experienced severe performance degradation when there were more redundant agents at the job fair, reflecting their limitations in skill matching tasks.

Received January 2024; revised July 2024; accepted October 2024