

MINDAGENT: EMERGENT GAMING INTERACTION

Ran Gong^{1†*}, Qiuyuan Huang^{2‡*}, Xiaojian Ma^{1*}, Hoi Vo³, Zane Durante^{4†}, Yusuke Noda³,

Zilong Zheng⁵, Song-Chun Zhu¹⁵⁶⁷, Demetri Terzopoulos¹, Li Fei-Fei⁴, Jianfeng Gao²

¹UCLA; ²Microsoft Research, Redmond; ³Xbox Team, Microsoft; ⁴Stanford; ⁵BIGAI; ⁶PKU; ⁷THU

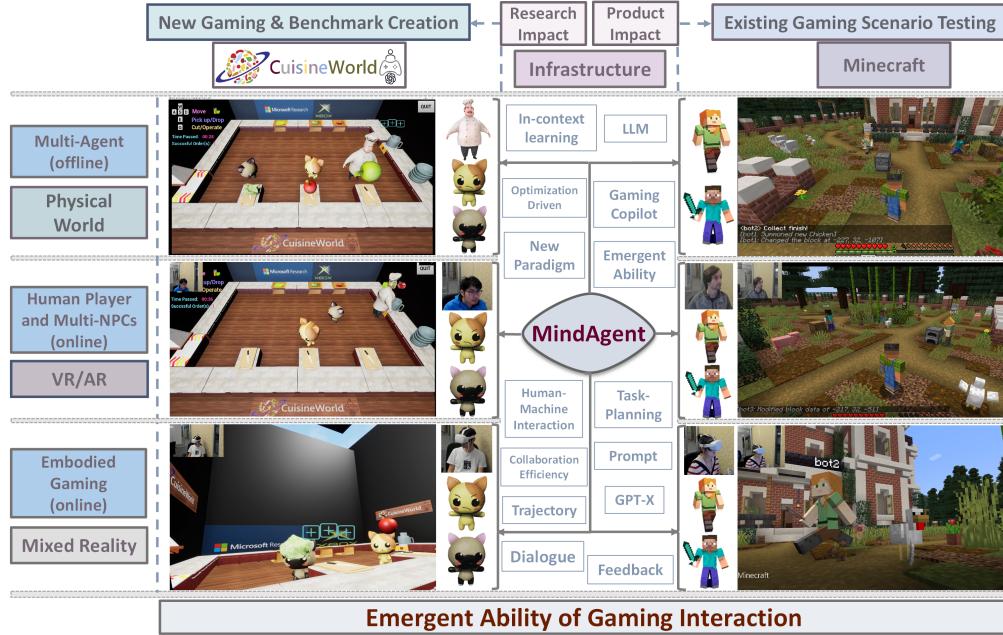


Figure 1: The MINDAGENT system for gaming interactions. MINDAGENT enables complex task planning in a multi-agent system and human-AI collaborated infrastructure across different domains.

ABSTRACT

Large Language Models (LLMs) have the capacity of performing complex scheduling in a multi-agent system and can coordinate these agents into completing sophisticated tasks that require extensive collaboration. However, despite the introduction of numerous gaming frameworks, the community has insufficient benchmarks towards building general multi-agents collaboration infrastructure that encompass both LLM and human-NPCs collaborations. In this work, we propose a novel infrastructure - **MindAgent** - to evaluate planning and coordination emergent capabilities for gaming interaction. In particular, our infrastructure leverages existing gaming framework, to i) require understanding of the coordinator for a multi-agent system, ii) collaborate with human players via un-finetuned proper instructions, and iii) establish an in-context learning on few-shot prompt with feedback. Furthermore, we introduce **CUISINEWORLD**, a new gaming scenario and related benchmark that dispatch a multi-agent collaboration efficiency and supervise multiple agents playing the game simultaneously. We conduct comprehensive evaluations with new auto-metric *collaboration score CoS* for calculating the collaboration efficiency. Finally, our infrastructure can be deployed into real-world gaming scenarios in a customized VR version of **CUISINEWORLD** and adapted in existing broader “Minecraft” gaming domain. We hope our findings on LLMs and the new infrastructure for general-purpose scheduling and coordination can help shed light on how such skills can be obtained by learning from large language corpora. Project webpage: <https://mindagent.github.io>.

* Equal Contribution. ‡ Project Leader.

† Work done while Ran and Zane interning at Microsoft Research, Redmond.

1 INTRODUCTION

Large language Models (LLMs) have been piloting the effort of developing general intelligent machines(Bubeck et al., 2023; Mirchandani et al., 2023) . Although they are trained in large text corpora, their superior problem-solving capacity is not limited to canonical language processing domains. LLMs already demonstrate the potential to tackle complex tasks that were previously presumed exclusive to domain-specific algorithms or human experts, ranging from mathematical reasoning (Imani et al., 2023; Wei et al., 2022; Zhu et al., 2022) to answering questions of professional law (Blair-Stanek et al., 2023; Choi et al., 2023; Nay, 2022) and medicine (Nov et al., 2023; Yang et al., 2023; Jeblick et al., 2022). More recently, some research has shown the possibility of using LLMs to generate complex plans for robots and game AI (Liang et al., 2022; Wang et al., 2023b;a; Yao et al., 2023; Huang et al., 2023), marking an important milestone for LLMs as generalist intelligent agents.

In this work, we would like to further investigate the planning capacity of LLMs. Specifically, we are interested in planning in a multi-agent system (Stone & Veloso, 2000), *i.e.*multi-agent planning. Compared to planning for a single agent, which has been extensively studied by previous research (Wang et al., 2023b;a), multi-agent planning imposes much higher problem-solving complexity due to the exponentially growing action space (w.r.t. number of agents). The planner has to simultaneously control multiple agents, avoid possible conflicts, and coordinate them into completing a shared goal that requires sophisticated collaborations. To understand to which extent can LLMs obtain multi-agent planning skills, we first establish a new benchmark, **CUISINEWORLD** as illustrated in Figure 1.

To incorporate agent AI into video games, we main design an infrastructure - **MINDAGENT** - inspired by multi-agent task allocation optimization theories to facilitate LLM multi-agent planning capabilities. Our infrastructure enables LLMs to perform complex coordination and scheduling with multiple different agents. We conduct comprehensive evaluations with recently introduced LLMs playing our game with our infrastructure, including GPT-4, Claude, and LLaMA. Through the proposed MINDAGENT interactive multi-agent planning framework for LLMs, we make the following key observations: 1) **zero shot multi-agent planning**: Without bells and whistles, powerful pretrained LLMs like GPT-4 are capable of scheduling multiple agents (ranging from 2 to 4) into completing dishes, and even collaborate with human players, by merely reading simple game instructions and recipes; 2) **planning with advanced prompting**: We are able to significantly boost their multi-agent planning performances by leveraging the emergent *in-context learning* capability (Brown et al., 2020; Wei et al., 2021): adding very few expert demonstrations even from different game levels to the prompt, explaining the rationale of certain actions as in Chain-of-Thought prompting (Wei et al., 2022), and providing on-the-fly feedback to the LLMs during planning; 3) **generalist potentials**: LLMs exhibits great potentials of being generalist multi-agent planner as it has strong generalization to coordinate more agents with examples of fewer agents, and adaptation to new game domains like Minecraft.

While compared to canonical domain-specific automated planning systems, multi-agent planning with LLMs can still be bottlenecked by challenging computation cost, context length limitation, non-optimal plans, *etc.*, it has the potential of improving from data without fine-tuning (via *in-context learning*), seamlessly adapting to planning problems from different domains and offering more flexible interfaces. We hope our findings on LLMs for general-purpose scheduling and coordination can help shed some light on how such skills can be obtained by learning from large text corpora, and facilitate the emergence of better LLM planners.

To summarize, our key contributions are as follows:

- We establish a new gaming scenario and related benchmark based on a multi-agent virtual kitchen environment, CUISINEWORLD. It adopts a minimal text-based game format and supports various planning task structures and difficulties, making it an ideal test bed for the emergent multi-agent planning (scheduling and coordination) capacity of LLMs.
- We introduce MINDAGENT, an infrastructure for interactive multi-agent planning with LLMs, which demonstrates the in-context learning multi-agent planning capacity of LLMs and brings several prompting techniques that help facilitate their planning ability, including providing few-shot demonstrations, planning rationals, and environmental feedback.

-
- We conduct extensive evaluations with multiple LLMs and prompting settings on our benchmark. Experimental results confirm their potential on being generalist multi-agent planners in terms of generalizing to more agents.
 - We deploy our system into real-world gaming scenarios and demonstrate its capabilities in human-AI interactions.

2 RELATED WORK

Multi-Agent Coordination. The field of multi-agent collaborations boasts a comprehensive body of literature. Traditionally, such collaborations have been modeled using MDP/POMDP (Lowe et al., 2017; Rashid et al., 2020; Jain et al., 2019) frameworks.

However, there has been a recent shift towards utilizing Large Language Models (LLMs) for these collaborations. For instance, Zhang et al. (2023b) delved into how large language models might communicate and cooperate in a watch-and-help (WAH) task. Meanwhile, Zhang et al. (2023a) investigated a two-agent collaboration game inspired by the simpler dynamics of the two-agent Overcooked-style game. Notably, their research chiefly concentrated on the task success rate, with most studies typically anchored to a singular task objective. In contrast, we emphasize the importance of collaboration efficiency in scenarios encompassing multiple task objectives. Further, our research uniquely focuses on evaluating the collaborative efficiency of more than two agents. Additionally, while other works like Park et al. (2023) simulate each agent individually, we employ a centralized system. This approach not only significantly reduces the number of API calls but also reduces context length, making it more appropriate for gaming applications.

Planning with LLMs. There exists a number of works that leverage LLMs to perform task planning (Huang et al., 2022a; Wang et al., 2023a; Yao et al., 2023). They leverage the LLMs’ internet-scale domain knowledge and emergent zero-shot planning abilities to perform complex task planning and reasoning. Recent works in robotics also leverage LLMs to perform task planning, they decompose a natural language instruction into a sequence of subtasks, either in natural language form or in python code (Ahn et al., 2022; Huang et al., 2022b; Liang et al., 2022). Then they use a low-level controller to execute these subtasks. Additionally, (Huang et al., 2022b; Liang et al., 2022; Wang et al., 2023b) also incorporate environment feedback to improve task performance.

Benchmarks using Games. Numerous games have been developed to study task planning Baker et al. (2022); Carroll et al. (2019), yet only a handful delve into multi-agent collaborations. Even within this limited subset, the focus predominantly remains on two-agent interactions where responsibilities are not evenly distributed. As evidenced by (Wan et al., 2022; Puig et al., 2020), it’s common for one player to assume a dominant role while the other provides support. In contrast, our paper assumes equal responsibilities across agents, and we expand our investigation to encompass collaborations involving more than just two agents, even with human players. While some previous studies have ventured into multi-task settings, none have delved into scenarios where agents must complete multiple distinct tasks using competing resources within a single episode. Furthermore, our game presents tasks with varied levels of difficulty.

Additionally, our work distinguishes itself from Carroll et al. (2019). Contrary to their settings, our game settings feature a diverse array of tools and task objectives, thereby generating an exponentially larger task space. A comparison between our work and other related games is shown in Table 1.

3 THE NEW GAMING CUISINEWORLD DESIGN AND BENCHMARK

We introduce CUISINEWORLD as a novel and flexible game for multi-agent scheduling and coordination in a *virtual kitchen* environment. In this game, a multi-agent system needs to overlook multiple agents and coordinate them, with the goal of completing as many dish orders as possible. It is equipped with a textual interface since our focus is evaluating LLM-based planning agents. Our modularized design separates tasks and game engines, allowing more tasks (type of dishes) and domains (how to implement the “kitchen”: text-based engine, Unity, Minecraft, etc.) to be included.

Benchmark	Multi-task	Object Interaction	Tool Use	Maximum Agents	Collaboration	Human in-the-loop	Procedural Level Generation
ALFWorld (Shridhar et al., 2020)	✓	✓	✓	1	✗	✗	✗
WAH (Puig et al., 2020)	✓	✓	✗	2	✓	✓	✗
TextWorld (Côté et al., 2019)	✓	✓	✓	1	✗	✗	✓
Generative Agents (Park et al., 2023)	✓	✓	✓	25	✗	✗	✓
EMATP (Liu et al., 2022)	✓	✓	✓	2	✓	✗	✗
Overcooked-AI (Carroll et al., 2019)	✗	✓	✓	2	✓	✓	✗
HandMeThat (Wan et al., 2022)	✓	✓	✓	2	✓	✗	✗
DialFRED (Gao et al., 2022)	✓	✓	✓	2	✓*	✗	✗
TEACH (Padmakumar et al., 2022)	✓	✓	✓	2	✓*	✗	✗
CerealBar (Suhr et al., 2019)	✗	✗	✗	2	✓	✗	✗
LIGHT (Urbanek et al., 2019)	✓	✗	✗	1369	✗	✓	✓
Diplomacy (Bakhtin et al., 2022)	✗	✗	✗	7	✓	✓	✗
CUISINEWORLD (Ours)	✓	✓	✓	4+	✓	✓	✓

Table 1: Comparison between CUISINEWORLD and other related benchmarks. **Multi-task:** The benchmark contains multiple different tasks. **Object Interaction:** Agents have to manipulate or engage with different items or environmental elements to achieve certain goals with irreversible actions. **Tool Use:** Completing tasks necessitates the use of specific tools by the agents. **Maximum Agents:** This denotes the upper limit of agents that can be present in a single experiment. **Collaboration:** Many tasks mandate teamwork and collaboration between different agents. **Human in-the-loop:** The framework allows humans to join the game and collaborate actively with the agents. **Procedural Level Generation:** There’s flexibility in adding new tasks, making the game dynamic and adaptable. *: Notably, even though multiple agents can be present, the second agent is limited to communicating with the first agent. The second agent cannot interact with the environment in an active gaming capacity.

Type	Arguments	Description
goto	agent location	Move agent to location
get	agent location (item)	agent obtain item from location
put	agent location	agent put everything it holds to location
activate	agent location	agent turn on location
noop	agent	not dispatching agent

Table 2: Action space in CUISINEWORLD.

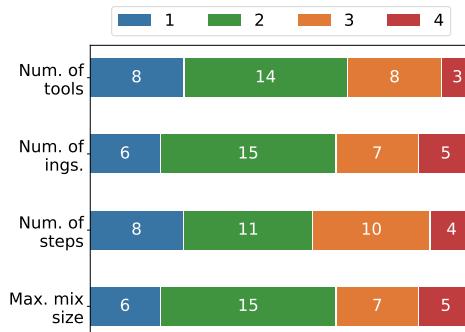


Figure 2: Dish distribution over the number of tools and ingredients (ings.) involved, cooking steps, and maximum mixture size as in the recipe.

3.1 TASK DEFINITION

We follow prior works (Yao et al., 2023; Liu et al., 2023; Deng et al., 2023) to **interactively evaluate LLMs as planning agents**. Overall, the interactive evaluation can be formulated as a *Markov Decision Process* ($\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{G}$), with state space \mathcal{S} , action space \mathcal{A} , (effectively indicating all the possible schedules that can be made at a single time step), transition dynamics \mathcal{T} , reward function \mathcal{R} and task instruction space \mathcal{G} . Note that, although there are multiple agents inside CUISINEWORLD that can be coordinated, as we mentioned above, we adopt a centralized planning scheme and thereby formulate our game as a single-agent and fully-observable decision-making problem. An illustration of the state & action space of our game can be found in Figure 1.

State Space \mathcal{S} . In CUISINEWORLD virtual kitchen, there are two types of entity: *location* and *agent*. For each entity, the game will provide a set of descriptions, the aggregated descriptions of all entities will be the state returned by our game. A *location* can be *storage*, where you could obtain ingredients and dispense waste, a *serving table*, where you should put the completed dish on, or a cooking tool, e.g. *pan*, *blender*. We offer up to two descriptions for each location: *inside(location, items)*, indicating what items (some ingredients, completed dishes, etc.) are now inside the location; and *occupy(location)*, suggesting location is now being used

and cannot be touched, *e.g.* an activated blender. A agent is an entity that can be dispatched to complete the task, and we provide up to three descriptions for each agent: `at(location, agent)`, indicating now agent is at location; `hold(agent, items)`, suggesting what items agent is holding; and finally `occupy(agent)`, implying agent is now operating a tool, *e.g.* chopping some fruits, and will not respond to any dispatching command.

Action Space \mathcal{A} . An action in CUISINEWORLD is a list of dispatching commands. Given N agent entities, a total of N commands need to be generated. The agent provides the following commands (also illustrated in Table 2): 1) `goto(agent, location)`, to let agent move to location; 2) `get(agent, location, item)`, to let agent get a specific item from location; 3) `put(agent, location)`, to put whatever agent is holding into location; 4) `activate(agent, location)`, to let agent turn on location if it is a cooking tool, *e.g.* `blender`; 5) `noop(agent)`, to have agent perform no actions in this round of dispatching. We will provide more detailed illustrations and rules about the action space in *appendix*. Note that, to avoid the possible confusion of multiple agents being dispatched to operate with the same location, the dispatcher also needs to properly order the dispatching commands as they will be executed sequentially.

Tasks and Reward. A task in CUISINEWORLD is a dish order, ranging from the most basic `tunaSashimi`, which can be made by simply chopping some tuna meat, to sophisticated dishes like `porkPasta` that requires various cooking tools. In a game episode with maximum steps of T , every τ_{int} steps (we name this *task interval*), a new task or dish order will be added to the active task list. A task will be viewed as *completed* and removed from the active task list when a matched dish has been put on the serving table. On the contrary, a task will be deemed to have *failed* and removed from the list when it reaches its *lifetime* τ_{lft} . Lifetime depends on the complexity of the dish and details can be found in *appendix*. Along with the tasks, the game provides rewards & penalties or feedback on certain occasions, *e.g.* when a task is just completed, some infeasible commands are dispatched, *etc*. Due to the space limit, we defer details on tasks to Appendix B..

3.2 IMPLEMENTING CUISINEWORLD

The implementation of CUISINEWORLD mostly follows the spirit of *Overcooked!*, a renowned video game. Therefore we refer to many of its game mechanisms while simplifying some of them, *e.g.* we skip low-level control and assume all agent have access to all location at any time (detailed comparisons between CUISINEWORLD and the original video game can be found in *appendix*). Specifically, we crawled the rules and recipes from the community-contributed wiki¹, streamlined them and made necessary modifications, ending up with the basic version of CUISINEWORLD comprising **10** types of location (*serving table*, *storage*, and 8 different cooking tools), **27** types of ingredients, and **33** unique dishes. We group the dishes based on their difficulty to make (primarily the number of cooking tools involved) and design **12** game levels, which are further categorized into 4 classes: *entry*, *simple*, *intermediate* and *advanced*, with 3 levels each. Note that the recipes, dishes, and levels can be easily extended to allow more challenging tasks.

3.3 EVALUATION METRIC

Collaboration Score (CoS). We would like to evaluate to which extent the dispatcher (played by an LLM) can coordinate multiple agents into completing dish orders, across different scenarios. Similar to the original *Overcooked!* game, we are particularly interested in this question: Can the dispatcher still coordinate the agents into efficient collaborations with smaller τ_{int} , *i.e.* more dish orders are flooding in? Our hypothesis is, an ideal dispatcher should be capable of coordinating agents until there are way more tasks than the system can handle. Therefore, we introduce *collaboration score CoS*, defined as below:

$$\text{CoS} = \frac{1}{M} \sum_{i=1}^M \frac{\#\text{completed task } [\tau_{\text{int},(i)}]}{\#\text{completed task } [\tau_{\text{int},(i)}] + \#\text{failed task } [\tau_{\text{int},(i)}]}, \quad (1)$$

where M is the total amount of τ_{int} we evaluate. Effectively, **CoS** is the average task completion rate across different τ_{int} conditions. In our default setting, we use $M = 5$. While the actual values of τ_{int}

¹<https://steamcommunity.com/sharedfiles/filedetails/?id=1769729191>

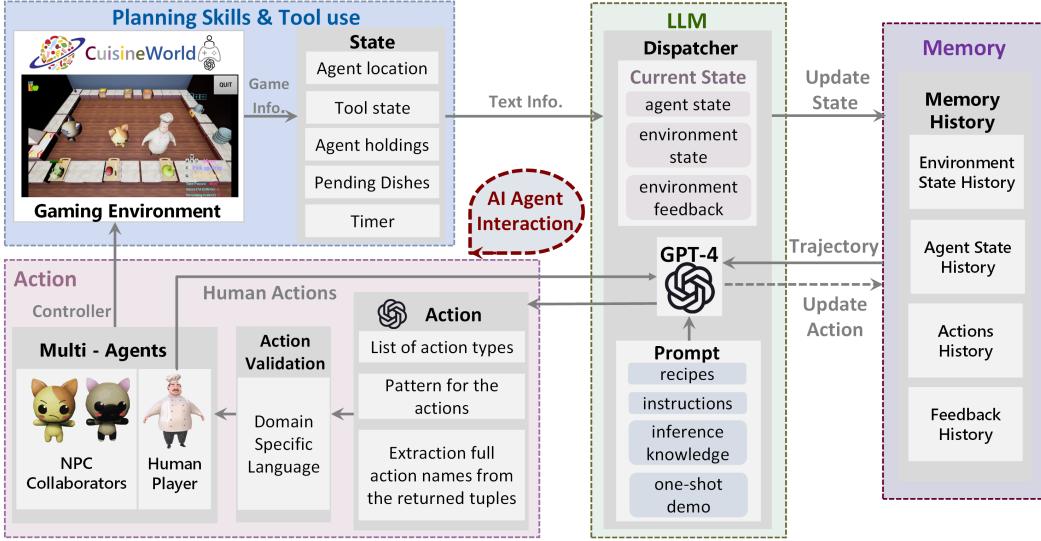


Figure 3: Our overview of our MINDAGENT architecture. **Planning Skill & Tool Use:** The game environment requires diverse planning skills and tool use to complete tasks. It emits related game information. This module also converts relevant game data into a structured text format so the LLMs can process it. **LLM:** The main workhorse of our infrastructure makes decisions, which is a dispatcher for the multi-agent system. **Memory History:** A storage utility that stores relevant information. **Action Module,** extract actions from text inputs and convert them into domain-specific language. Validate DSLs so they don't cause errors when executing.

depend on the game level, we ensure they elicit a wide range of difficulty including both extremely relaxed and intense scenarios.

In a word, CuisineWorld is a game that emulates a virtual kitchen, where several robots are commanded to use various cooking tools and ingredients to prepare as many dish orders as possible in a limited period of time. To facilitate collaboration, new orders will keep flooding in while the existing ones should be completed before expiration. Therefore, LLMs need to properly coordinate these robots to maximize overall productivity. CUISINEWORLD also offers game levels with a wide range of planning difficulty: dishes with different complexity (number of ingredients and tools involved), number of agents, order frequency and lifetime, etc, making it an ideal test bed for LLM-based multi-agent planning.

4 MINDAGENT: INFRASTRUCTURE FOR GAMING AI

4.1 INFRASTRUCTURE

Our first foray into the challenging CUISINEWORLD benchmark is an interactive multi-agent planning framework for LLMs: MINDAGENT. It adopts a minimalist design for the purpose of demonstrating the emergent capacity of LLMs in scheduling and coordination, while also bringing in exploratory prompting techniques that facilitate better planning and shed some light on future approaches. Our infrastructure follows in-context learning. We will outline the key techniques below:

To facilitate in-context learning, our MINDAGENT infrastructure is composed of three primary components: the prompt, current state, and memory.

Within the prompt component, there are four distinct sub-components: recipes, general instructions, inference knowledge, and a one-shot demo.

Recipes. outline the hierarchical procedure for preparing various dishes at the given level. They specify the necessary ingredients for each intermediate or final product, the appropriate tools required, and the expected outcome post-cooking.

Instructions. detail the foundational rules of CUISINEWORLD. These instructions delineate the array of actions agents can undertake within the game and enumerate the characteristics of every tool available in the current kitchen scenario. Moreover, they inform agents about the base ingredients retrievable from storage, as well as all potential intermediate products they can procure. Agents are also explicitly advised to remain cautious about feedback from the environment.

Inference Knowledge. houses insights and helpful hints for the agent. When utilized appropriately, these hints can guide agents to sidestep potential errors and enhance their collaborative efficiency.

One-shot Demo. presents a step-by-step demonstration of the preparation of a distinct dish, different from other dishes at the current level. This demonstration spans several time steps, each of which is incorporated as part of the prompt. The demonstration illustrates the major procedures for cooking one dish in CUISINEWORLD, including obtaining ingredients, putting ingredients into different tools, transporting intermediate ingredients, and delivering the final dish to the serving table.

Current State. provides a snapshot of the prevailing observations from the environment. It encompasses information such as the agents' locations, the objects currently in the agents' possession, the tools that are accessible within the environment, the ingredients present within each tool, and the tools that are actively in use. Moreover, it includes optional feedback from the environment, triggered when the agents' actions contravene the environment rules—for instance, when assigning two distinct actions to the same agent.

Memory History. archives the interaction history with the environment. Specifically, it chronicles the state of the environment and the state of the agents at every time step.

In addition to the prompt modules, additional modules are implemented to help interface between LLMs and CUISINEWORLD.

Action Extraction. employs a regular expression matching procedure to distill agent actions from the LLM's textual output. This module is indispensable because, on occasion, the LLM's output is not clean. The output contains information reflecting its internal thought processes. At times, the LLM might even issue apologies for prior missteps in reaction to environment feedback.

Action Validation. utilizes a look-ahead checking mechanism. This module parses the proposed actions, assessing their feasibility. Should an action be deemed inexecutable, an error message is promptly returned.

4.2 INFRASTRUCTURE MECHANISM

Assuming a multi-agent system with a total of N agents, the system must complete a sequence of P different tasks. Each task has M_p different sub-tasks. Furthermore, the number and types of tasks are unknown at the beginning of the episode. The environment will sample a task for the agents to finish for a given interval. Then the agents need to complete the designated task along with other tasks in the task queue. In addition, each task has an expiration time. After the expiration time, the task will be marked as a failure. The objective of the multi-agent system is to finish as many tasks as possible and fail as fewer tasks as possible within a given time frame.

We aim to find valid and optimal task planning, scheduling, and allocations. We define q_{pim} and c_{pim} as quality and cost, respectively, for allocating agent i to work on the sub-task m for the p th task in the episode. Then the combined utility for the sub-task is:

$$u_{pim} = \begin{cases} q_{pim} - c_{pim}, & \text{if agent } i \text{ can execute sub-task } m \text{ for the } p \text{ th task in the episode} \\ -\infty, & \text{otherwise} \end{cases}$$

We define the assignment of sub-task m to agent i as

$$v_{pim} = \begin{cases} 1, & \text{agent } i \text{ is assigned to sub-task } m \text{ for the } p \text{ th task in the episode} \\ 0, & \text{otherwise} \end{cases}$$

The goal is to maximize the utility of the episode under a time constraint. Define the execution time for task m by agent i for the p th task in the episode as τ_{pim} , and the maximum time allowed to execute the task as T_{max} , we can express the task decomposition and assignment problem as follows:

$$\arg \max_v \sum_{p=1}^P \sum_{i=1}^N \sum_{m=1}^{M_p} u_{pim} v_{pim} \quad (2)$$

Subject to:

$$\begin{aligned} \sum_p \sum_i \sum_m \tau_{pim} v_{pim} &\leq T_{max} \\ \sum_i v_{pim} &\leq 1 \quad \forall m \in M, \forall p \in P \\ v_{pim} &\in \{0, 1\} \quad \forall i \in N, \forall m \in M, \forall p \in P \end{aligned}$$

As pointed out by (Korsah et al., 2013), this problem cannot be solved in polynomial time. In this work, we tackle this problem by using large-language models.

Our prompt design choices try to help LLM system solve Equation 2. In practice, we reformulate Equation 2 with qualities or rewards expressed in natural languages as environment feedback. For example, when the agent successfully collects an item, the environment emits a signal “collect finish.” When the dispatcher assigns a different task to the same agent, the environment will emit a signal “agent ids cannot be the same.” As rewards are not immediately observable, we borrow sprites from temporal difference learning. We accumulate state-action history into memory history. Due to context length limits, it’s infeasible to fit the entire history into the context window. We select a fixed horizon history as a part of the prompt to guide the model performance. We further express the constraints of the system in natural language formats and repeat important constraints multiple times if necessary.

5 EXPERIMENTS AND RESULTS

Overview. We conduct extensive experiments in CUISINEWORLD. We first introduce the experiment settings and present an analysis of empirical results in CUISINEWORLD. Our experiments focus on addressing the following research questions:

- Q1:** How efficiently can the model dispatch multiple agents?
- Q2:** Can the model dispatch agents for dynamic, on-the-fly goals across different tasks?
- Q3:** How do various components of the input prompt influence the model’s performance?
- Q4:** How do other LLMs perform compared to GPT-4?
- Q5:** To what extent can the existing methods collaborate with human users?
- Q6:** What’s the human perception of collaborating with numerous intelligent agents?

5.1 LLM SETTINGS

We perform experiments on CUISINEWORLD through OpenAI APIs and anthropic APIs. All GPT-4 experiments are using gpt-4-0613 model, and all chat-GPT experiments are using gpt-3.5-turbo-0613. For Llama 2 experiments, we use hugging face inference endpoints Llama-2-70b-chat-hf. We set the temperature for all experiments to 0.1 following (Wang et al., 2023a). We report the average results over three episodes.

5.2 EXPERIMENT SETTING I: LLMs DISPATCH MULTI-AGENTS (NPC)

Collaboration Efficiency (Q1, Q2). Figure 4 and Table 3, Table 4 and Table 5 reports the system performance under different settings. In particular, Table 3 reports the multi-agent collaboration results among two agents. Table 4 reports the multi-agent collaboration results among three agents, and Table 5 reports the multi-agent collaboration results among four agents. Figure 4 displays the collaboration efficiency curve.

As shown in Figure 4, across different task levels, more agents generally lead to better collaboration efficiencies. As the collaboration efficiency curve is generally higher with more agents.

Computing **CoS** by levels also reveals that more agents will lead to better collaboration efficiencies. As shown in the tables, the **CoS** score is the highest when there are two agents in two cases. The

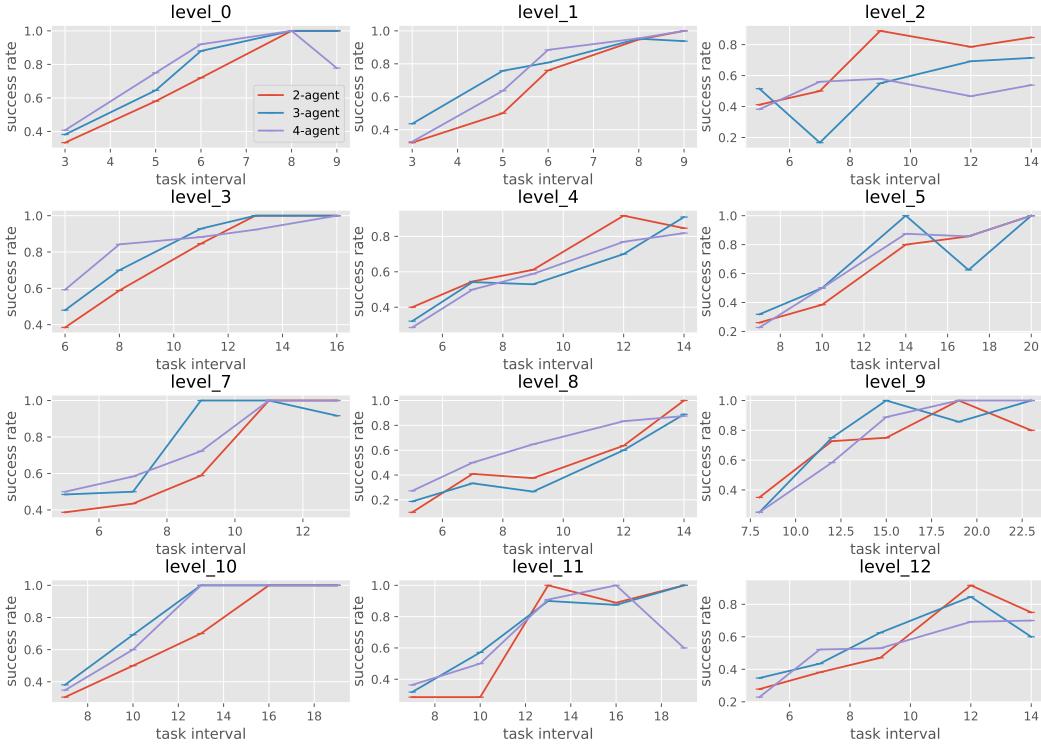


Figure 4: Collaboration Results on Different Tasks

CoS score is the highest when there are three agents in seven cases. The **CoS** score is the highest when there are four agents in three cases. The results also confirm that more agents will lead to higher collaboration efficiencies.

Findings. First, we observe that the system performance is generally better when there are more agents, indicating that LLM dispatcher can coordinate more agents to execute tasks more efficiently. Second, we observe that the system performance degrades with more agents in less demanding conditions, indicating that LLM dispatcher struggles when there are fewer tasks.

5.3 EXPERIMENT SETTING II: HUMAN AND MULTI-NPCs WITH LLMs

5.3.1 HUMAN DATA COLLECTION

Human Testing of Study Protocol. Before starting the experiment, a webpage introduction to the game is handed to the players. It contains rules and the basic controls of the game. Then we randomly assign the playing order. Participants can drop out of the testing at any time as they wise; in that case, their data will be discarded. The human evaluation interface is shown in Appendix D.

Measurement. In the background, we collect the number of failed and successful tasks during the participant's interaction with the game system. In addition, we record the entire action history of players and intelligent agents. Therefore, we can replay action histories for further analysis. After each episode, the participants must complete a survey about their engagement with the system on a 5-point likert chart.

Our objective measure is intended to evaluate the human AI teaming performance, and the subjective measure is designed to evaluate users' perceptions of the system.

5.3.2 EXPERIMENT II SETTING

We conducted a user study in our gaming environment that tries to answer **Q5**, **Q6**.

2-agent	very simple			simple			intermediate			advanced			Avg.
	level 0	level 1	level 7	level 2	level 4	level 8	level 3	level 9	level 10	level 5	level 11	level 12	
GPT4 $\tau_{\text{int},(1)}$	18/54	18/56	12/31	14/34	12/30	3/30	10/26	7/20	7/23	6/23	6/21	10/36	0.318
GPT4 $\tau_{\text{int},(2)}$	18/31	17/34	10/23	13/26	12/22	9/22	10/17	8/11	6/12	5/13	4/14	8/21	0.486
GPT4 $\tau_{\text{int},(3)}$	18/25	19/25	10/17	16/18	11/18	6/16	11/13	6/8	7/10	8/10	9/9	8/17	0.709
GPT4 $\tau_{\text{int},(4)}$	18/18	18/19	12/12	11/14	11/12	7/11	12/12	8/8	9/9	6/7	8/9	11/12	0.912
GPT4 $\tau_{\text{int},(5)}$	18/18	17/17	12/12	11/13	11/13	9/9	11/11	4/5	7/7	8/8	8/8	9/12	0.937
CoS	0.727	0.706	0.682	0.687	0.664	0.504	0.764	0.725	0.701	0.661	0.692	0.559	0.673

Table 3: 2 agents performance on different tasks

3-agent	very simple			simple			intermediate			advanced			Average
	level 0	level 1	level 7	level 2	level 4	level 8	level 3	level 9	level 10	level 5	level 11	level 12	
GPT4 $\tau_{\text{int},(1)}$	21/55	24/55	16/33	17/33	9/28	6/32	12/25	5/20	8/21	7/22	7/22	9/26	0.368
GPT4 $\tau_{\text{int},(2)}$	20/31	25/33	11/22	4/24	13/24	7/21	14/20	9/12	9/13	7/14	8/14	10/23	0.549
GPT4 $\tau_{\text{int},(3)}$	22/25	21/26	17/17	11/20	9/17	4/15	13/14	8/8	12/12	7/7	9/10	10/16	0.791
GPT4 $\tau_{\text{int},(4)}$	22/22	20/21	14/14	9/13	7/10	6/10	10/10	6/7	10/10	5/8	7/8	11/13	0.846
GPT4 $\tau_{\text{int},(5)}$	20/20	15/16	11/12	10/14	10/11	8/9	12/12	6/6	8/8	5/5	8/8	6/10	0.914
CoS	0.781	0.778	0.780	0.528	0.600	0.455	0.822	0.771	0.815	0.689	0.733	0.570	0.694

Table 4: 3 agents performance on different tasks

4-agent	very simple			simple			intermediate			advanced			Average
	level 0	level 1	level 7	level 2	level 4	level 8	level 3	level 9	level 10	level 5	level 11	level 12	
GPT4 $\tau_{\text{int},(1)}$	22/54	18/55	17/34	13/34	8/28	9/33	16/27	5/20	8/23	5/22	8/22	8/35	0.349
GPT4 $\tau_{\text{int},(2)}$	24/32	21/33	14/24	14/25	12/24	11/22	16/19	7/12	9/15	7/14	6/12	12/23	0.590
GPT4 $\tau_{\text{int},(3)}$	23/25	23/26	13/18	11/19	10/17	11/17	15/17	8/9	11/11	7/8	10/11	9/17	0.785
GPT4 $\tau_{\text{int},(4)}$	22/22	21/22	14/14	7/15	10/13	10/12	12/13	9/9	10/10	6/7	8/8	9/13	0.875
GPT4 $\tau_{\text{int},(5)}$	14/18	20/20	14/14	7/13	9/11	7/8	12/12	5/5	7/7	6/6	3/5	7/10	0.859
CoS	0.771	0.761	0.761	0.505	0.592	0.626	0.848	0.744	0.790	0.692	0.675	0.534	0.692

Table 5: 4 agents performance on different tasks

The user study evaluates the LLM dispatcher’s capabilities of collaborating with humans, where participants are collaborating with 1,2,3 agents or working alone on the virtual cooking tasks. We consider the most general setting, where the LLM works on the unseen task, level_3.

5.3.3 EXPERIMENT II DESIGN

Hypotheses. The user study tests the following hypotheses:

- **H1: Task productivity.** Participants have higher productivity if collaborating with AI agents.
- **H2: Task productivity with more agents.** Participants have higher productivity if collaborating with more AI agents.
- **H3: Perception of the robot.** Participants would have higher perceived task efficiency and have more fun playing the game due to collaboration.

Manipulated Variables. We use a within-subject design for our experiment. In particular, every user tries to finish the task by himself or collaborates with different numbers of robots with varying degrees of competency. We randomize the order of the treatment to mitigate practice effects, fatigue effects, and carryover effects.

- **Single agent:** Participants work on the task by themselves.
- **LLM powered multi-agent system:** Participants collaborate with the multi-agent system powered by LLM.
- **Random agent:** Random agents execute random actions from a pool of valid actions. Participants collaborate with random agents.

Main Results. We recruited 12 subjects for our study. Among them, there are two females and 10 males.

We use ANOVA to test the effects of different experimental conditions on collaboration performance and subjective perception of the AI agents. Tukey HSD tests are conducted on all possible pairs of experimental conditions.

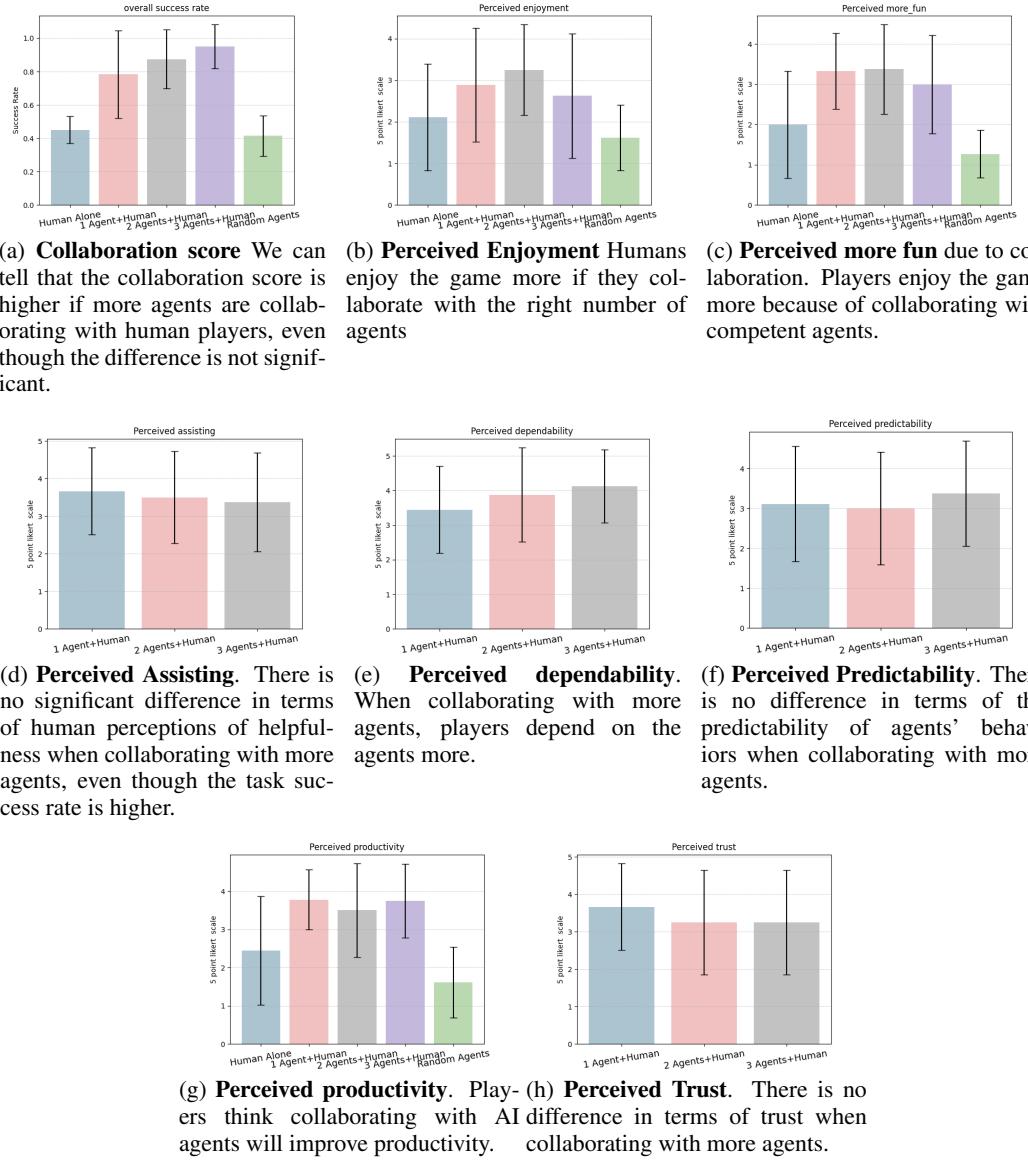


Figure 5: Human Evaluations

Findings. We find significant effects on team collaboration success rate $F(4, 55) = 28.11, p < 0.001$. Post-hoc comparisons using the Tukey HSD tests revealed that the team of the player with LLM agents achieves a higher success rate than a human working alone, $p < 0.001$ across different numbers of agents, **confirming H1**. Even though the success rate is generally higher when collaborating with more agents, there is no significant effect compared with collaborating with one agent, collaborating with two agents $p = 0.774$, or collaborating with three agents $p = 0.231$. We observe that human players have more fun playing the game when collaborating with LLM-powered intelligent agents than playing alone, $p = 0.0126$. Players feel that collaboration with intelligent agents leads to higher productivity, $p = 0.0104$, thus **confirming H3**.

In addition, when playing with intelligent agents, human players will take their actions based on other players' actions $p = 0.00266$. Human players also found that intelligent agents are more predictable compared with random agents $p < 0.001$.

Further insights from player feedback highlighted an intriguing trade-off: while more agents improved overall task success rates, it reduced the game's enjoyment. Often, players felt sidelined and less involved. Thus, game developers should adjust AI performance to maintain player engagement

and fun. As indicated by Yuan et al. (2022), aligning human values with AIs might be a promising way to solve this problem.

5.4 VISUALING "CUISINEWORLD"

To implement CUISINEWORLD into a real game system, we built on top of Gao et al. (2020). In our game, as visually depicted in Figure 6, players are given the opportunity to engage in collaborative interactions with NPCs. In this game, human players' actions can be obtained from an inverse dynamic model by checking preconditions and post-effects. This introduces a unique dynamic to the gameplay, enabling users to experience a more immersive cooperative environment. Additionally, the game's interface is versatile, allowing players multiple ways to interact within the game world. They can either use a standard keyboard setup, which is more conventional and likely familiar to most PC gamers, or they can immerse themselves even further using a Virtual Reality (VR) device. This VR functionality ensures a more tactile and realistic interaction, as players can physically move, gesture, and engage with the NPCs and other in-game elements in a 3D environment.

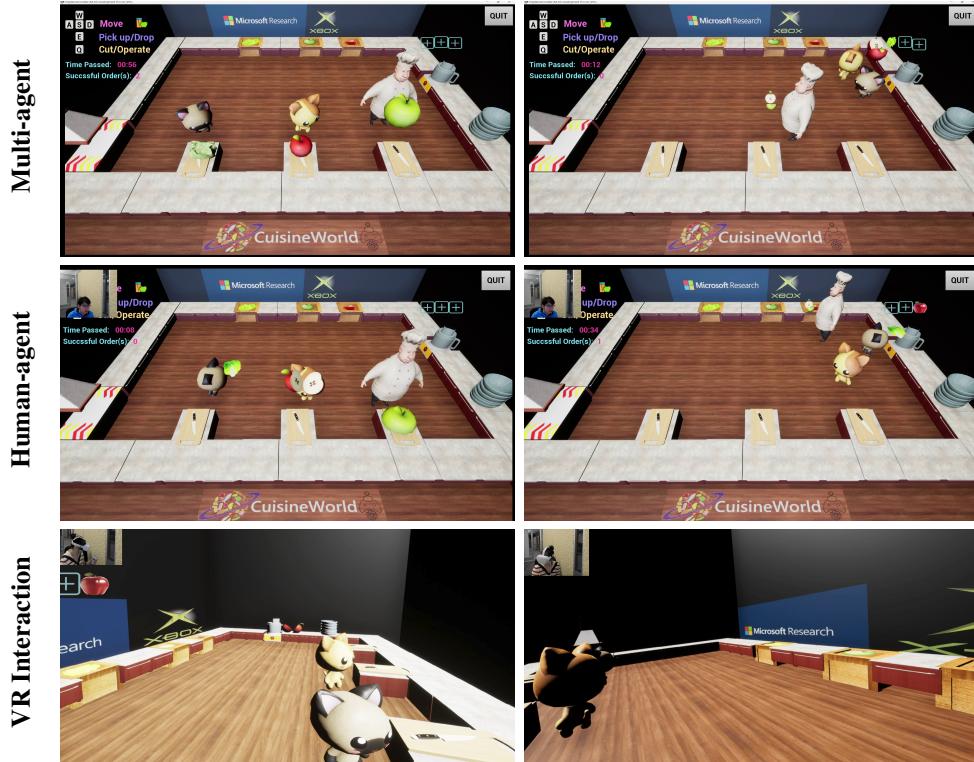


Figure 6: The top two images show a multi-agent collaboration example in CuisineWorld, the three agents are preparing a mixed juice together. The middle two images show a human player as the head chef instructing the agents to cook mixed juice. The bottom two images show a human player collaborating with collaborative agents in VR.

6 ANALYSIS AND EMERGENT GAMING ABILITIES

6.1 ABLATION STUDY FOR MULTI-AGENTS

Study on the Prompt Components Q3. In Table 7, we elucidate the performance of LLM dispatchers with certain components of the prompt omitted. Details about prompt can be found in Appendix Figure 9 and Figure 8. Specifically, for these tests, we excluded individual components like inference knowledge, reduced the prompt example to a mere two steps instead of the complete demonstration, and evaluated the model without environment feedback. For context, our principal experiments, varying in the number of agents, incorporate a one-shot example for the correspond-

ing number of agents. Our ablation studies further probe how varying the number of agents can influence model performance, with details in Table 8.

Findings: From Table 7, a significant drop in performance is observed when environment feedback is excluded, underscoring its pivotal role in the efficacy of the LLM dispatcher. Replaying action sequences reveals that, without feedback, the LLM dispatcher tends to repeat mistakes and gets stuck in specific states for prolonged durations. Another key takeaway is that a succinct two-step demonstration of input and output format can still achieve commendable performance for unseen tasks with dynamic objectives. Notably, in these two-step instances, there’s no explicit guide to finish any tasks. Yet, the model doesn’t merely complete the task but continually performs additional tasks within the same episode. Furthermore, we also observe that integrating human-crafted inference knowledge bolsters the LLM dispatcher’s performance. Lastly, even with few-shot demonstrations involving fewer agents, the LLM dispatcher retains satisfactory performance as shown in Table 8.

Study on Other LLMs’ Performance Q4. To study how other LLMs perform on our tasks, we tested the collaboration performance of GPT-3.5, Claude-2 and LLaMA in Table 6. For a fair comparison, all tests employed identical prompt inputs.

Findings: We observe that while other LLMs tend to underperform, models such as Claude-2 still manage to complete the task to a considerable extent.

6.2 EMERGING CAPABILITIES

Across our experiments, we observe the following emergent properties under our MINDAGENT framework.

Emergent Collaboration Tasks Understanding. As shown in Table 7, especially in the few-step ablation entries, GPT-4 exhibits its proficiency even when not provided with a full demonstration for specific tasks. To clarify, a “full few-shot demo” typically refers to a comprehensive demonstration of a task, detailing each step and procedure involved. In contrast, we use provide GPT-4 with only a partial demonstration or a glimpse of the task only executing two steps.

Yet, despite this limited input, GPT-4’s performance is remarkable. This underscores GPT-4’s impressive **emergent zero-shot multi-agent planning** capabilities. Beyond simply completing unseen tasks, GPT-4 also demonstrates adaptability by dynamically prioritizing multiple different tasks as they arise, emphasizing its **emergent multi-task, on-the-fly planning** skills.

Emergent Multi-agent Reasoning Capabilities. Referencing Table 8, GPT-4 has the capability to deploy more agents based on demonstrations of fewer agents. For instance, GPT-4 can effectively dispatch four agents having only seen demonstrations involving two agents. Moreover, the efficiency of collaboration is higher as the number of agents increases, spotlighting its **emergent collaboration** prowess.

	2 agent				3 agent				4 agent			
	GPT-4	Claude-2	LLaMA	ChatGPT	GPT-4	Claude-2	LLaMA	ChatGPT	GPT-4	Claude-2	LLaMA	ChatGPT
$\tau_{int,(1)}$	10/26	3/24	0	0/24	12/25	5/26	0	0/24	16/27	9/25	0	0/24
$\tau_{int,(2)}$	10/17	3/16	0	0/15	14/20	4/16	0	0/15	16/19	4/15	0	0/15
$\tau_{int,(3)}$	11/18	3/12	0	0/12	13/14	3/12	0	0/12	15/17	4/12	0	0/12
$\tau_{int,(4)}$	11/13	3/9	0	0/9	10/10	5/11	0	0/9	12/13	6/11	0	0/9
$\tau_{int,(5)}$	11/11	4/6	0	0/6	12/12	5/7	0	0/6	12/12	6/7	0	0/6
CoS	0.686	0.3125	0	0.822	0.372	0	0	0.848	0.473	0	0	0

Table 6: Performance of Other LLMs on Level 3

2 agent	GPT-4	GPT-4 w/ few-step	GPT-4 w/o inference knowledge	GPT-4 w/o feedback
$\tau_{int,(1)}$	10/26	8/26	8/25	4/25
$\tau_{int,(2)}$	10/17	11/19	9/17	4/17
$\tau_{int,(3)}$	11/13	11/13	10/12	4/12
$\tau_{int,(4)}$	12/12	9/11	8/9	1/9
$\tau_{int,(5)}$	11/11	10/10	9/9	5/7
CoS	0.764	0.710	0.714	0.311

Table 7: Additional Ablation

level_3	4agent using 4agent module	4agent using 2agent module	3agent using 3agent module	3agent using 2agent module
GPT4 $\tau_{\text{int},(1)}$	16/27	14/27	12/25	11/25
GPT4 $\tau_{\text{int},(2)}$	16/19	16/20	14/20	11/19
GPT4 $\tau_{\text{int},(3)}$	15/17	15/16	13/14	12/14
GPT4 $\tau_{\text{int},(4)}$	12/13	13/13	10/10	12/12
GPT4 $\tau_{\text{int},(5)}$	12/12	12/12	12/12	11/11
CoS	0.848	0.851	0.822	0.775

Table 8: Using different numbers of agent demos

7 NOVEL GAME ADAPTATION

In line with our ongoing efforts to create collaborative, in-game, multi-agent systems, we ventured beyond CuisineWorld and made strides in integrating our infrastructure into the widely popular sandbox game, Minecraft. In this new adaptation, we designed several unique cooking tasks where two in-game agents, Alex and Steve, are assigned the responsibility of cooking various types of meat as shown in Figure 7. After cooking, agents need to deposit the items into a chest. More details can be found in Appendix C. The experiment results are presented in Table 9.

We define the following actions for the multi-agent system in our Minecraft game: 1) `goto(agent, location)`; 2) `killMob(agent, mobType)`; 3) `mineBlock(agent, blockType)`; 4) `putFuelFurnace(agent, fuelType)`, to put the item from agent’s inventory to the furnace’s bottom slot. 5) `putItemFurnace(agent, itemType)`, to put the item from agent’s inventory to the furnace’s top slot; 6) `takeOutFurnace(agent)`, take out the cooked item from the furnace 7) `putInChest(agent, itemType)` ;

The state space in Minecraft contains the following: 1) nearby blocks for each agent 2) nearby entities for each agent. 3) each agent’s inventory 4) items inside the furnace 5) items inside the chest. 6) human player’s inventory if a human player is involved.

To ensure reproducibility, we modify the game mechanism. A killed mob will respawn nearby, and a mined block will also respawn nearby.

The empirical data we collected from these game sessions provided us with compelling evidence that the multi-agent collaboration infrastructure we’ve developed has the robustness to be extrapolated and adapted across multiple distinct games, paving the way for broader applications in the gaming industry.

Going a step further, we bridged the gap between human players and in-game (NPC) agents by integrating Microsoft’s Azure speech-to-text API into the Minecraft environment. This addition allows human players to communicate and collaborate with in-game NPC agents using voice chat. Human players can express their intents and desired goals to NPCs in real-time through voice chat. This real-time vocal interaction enriches the gameplay experience, fostering a deeper level of immersion and synergy between human players and AI agents. Moreover, this integration opens the door for research into the efficacy of voice-assisted AI learning and how real-world human interactions can shape AI behavior in virtual domains.

In the case of the human player chatting with the multi-agent system, the prompt contains additional human instructions and human dialog history components.

In addition, by integrating Minecraft VR mode with our infrastructure, we can bring the player interactive experiences to the next level.

GPT-4 minecraft	$\tau_{\text{int},(1)}$	$\tau_{\text{int},(2)}$	$\tau_{\text{int},(3)}$	$\tau_{\text{int},(4)}$	$\tau_{\text{int},(5)}$	CoS
Performance	0.195	0.381	0.704	0.792	0.833	0.581

Table 9: Performance of our framework in Minecraft



Figure 7: The top two images show a multi-agent collaboration example in Minecraft. In the left image, Alex and Steve are killing different animals, and in the right image, Alex and Steve are cooking meat in a furnace together. The middle two images show a human player instructing the agents to perform certain actions. The bottom two images show a human player collaborating with agents in VR.

8 CONCLUSION

In this paper, we presented MINDAGENT, an infrastructure for multi-agent collaboration through LLMs across multiple gaming domains. We investigated the multi-agent planning capabilities of MINDAGENT, and we deployed our infrastructure into real-world video games to demonstrate its effectiveness for multi-agent collaboration and human-AI collaboration. Beyond its practical applications, we hope that our endeavor serves as a beacon, guiding the development of future gaming systems where human-AI collaboration is seamless and intuitive. Furthermore, we are optimistic that our insights and findings might catalyze innovations in crafting games that are not only technologically advanced but also significantly more engaging and enjoyable for players.

ACKNOWLEDGMENTS

We are especially grateful to Johannes Gehrke, Ryon White, Haiyan Zhang, Kareem Choudhry for their enormous advice, support and encouragement of the work. We appreciate Katja Hofmann, Andrzej Banburski-Fahay, Jianwei Yang, Michel Galley, Nebojsa Jojic, Bill Dolan for the early insightful discussions, suggestions and comments. The authors gratefully acknowledge Adrian Brown from X-Box team for his discussion, feedback and pointers to the modeling generation and literature. We thank Rohan Taori, Janardhan Kulkarni, Ziheng Zhou, Yu Wang, Eloi Moliner Juanpere, Xiaofeng Gao, Collin Huang, Xiaodong Yu, and Shuwen Qiu for their help on the human experiment setup.

REFERENCES

- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. 3
- Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022. 3
- Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. 4
- Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*, 2023. 2
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 2
- Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019. 3, 4
- Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. Available at SSRN, 2023. 2
- Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers* 7, pp. 41–75. Springer, 2019. 4
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023. 4
- Xiaofeng Gao, Ran Gong, Yizhou Zhao, Shu Wang, Tianmin Shu, and Song-Chun Zhu. Joint mind modeling for explanation generation in complex human-robot collaborative tasks. In *2020 29th IEEE international conference on robot and human interactive communication (RO-MAN)*, pp. 1119–1126. IEEE, 2020. 12
- Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022. 4
- Qiuyuan Huang, Jae Sung Park, Abhinav Gupta, Paul Bennett, Ran Gong, Subhajit Som, Baolin Peng, Owais Khan Mohammed, Chris Pal, Yejin Choi, et al. Ark: Augmented reality with knowledge interactive emergent ability. *arXiv preprint arXiv:2305.00970*, 2023. 2

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9118–9147. PMLR, 17–23 Jul 2022a. URL <https://proceedings.mlr.press/v162/huang22a.html>. 3

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*, 2022b. 3

Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*, 2023. 2

Unnat Jain, Luca Weihs, Eric Kolve, Mohammad Rastegari, Svetlana Lazebnik, Ali Farhadi, Alexander G Schwing, and Aniruddha Kembhavi. Two body problem: Collaborative visual task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6689–6699, 2019. 3

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022. 2

G Ayorkor Korsah, Anthony Stentz, and M Bernardine Dias. A comprehensive taxonomy for multi-robot task allocation. *The International Journal of Robotics Research*, 32(12):1495–1512, 2013. 8

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *arXiv preprint arXiv:2209.07753*, 2022. 2, 3

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023. 4

Xinzhu Liu, Xinghang Li, Di Guo, Sinan Tan, Huaping Liu, and Fuchun Sun. Embodied multi-agent task planning from ambiguous instruction. *Proceedings of robotics: science and systems, New York City, NY, USA*, pp. 1–14, 2022. 4

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017. 3

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023. 2

John J Nay. Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *Nw. J. Tech. & Intell. Prop.*, 20:309, 2022. 2

Oded Nov, Nina Singh, and Devin M Mann. Putting chatgpt’s medical advice to the (turing) test. *medRxiv*, pp. 2023–01, 2023. 2

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2017–2025, 2022. 4

Joon Sung Park, Joseph C O’Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. 3, 4

-
- Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020. 3, 4
- Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1):7234–7284, 2020. 3
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020. 4
- Peter Stone and Manuela Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8:345–383, 2000. 2
- Alane Suhr, Claudia Yan, Charlotte Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. *arXiv preprint arXiv:1910.03655*, 2019. 4
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*, 2019. 4
- Yanming Wan, Jiayuan Mao, and Josh Tenenbaum. Handmethat: Human-robot communication in physical and social environments. *Advances in Neural Information Processing Systems*, 35: 12014–12026, 2022. 3, 4
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a. 2, 3, 8
- Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023b. 2, 3
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2
- Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*, 2023. 2
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 4
- Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183, 2022. 12
- Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. Proagent: Building proactive cooperative ai with large language models. *arXiv preprint arXiv:2308.11339*, 2023a. 3
- Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*, 2023b. 3

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang.
Solving math word problem via cooperative reasoning induced language models. *arXiv preprint arXiv:2210.16257*, 2022. 2

APPENDIX

A PROMPT EXAMPLES

We provide some prompt examples for CuisineWorld. Figure 8 shows an example of the system prompt info. Figure 9 shows an example of a partial demonstration.

The available actions are :

- 0) get: get some object from a location
- 1) put: put some object into a tool
- 2) activate: activate the tool to cook all ingredients inside the tool into a different tools

Sometimes the system will give you error messages. Please consider these error messages when executing actions.

You need to specify action for all of the agents, **except human**. They all have different agent numbers. Do not assign actions to the same agent more than once.

When the tools reach its capacity, you need to take stuff out. Otherwise, you cannot put items inside.

When you are holding objects, you cannot get any more objects.

When you are holding objects, you cannot activate tools.

If you are holding dishes, you need to put it at the servingtable.

You can only pick up objects from the tool location, if you are located at the tool location.

When you activate any tools, make sure all the items inside the tool are respecting the recipes. Otherwise, you will cook waste. Avoid waste at all cost.

*** You should mix salad in the mixer. To make salad you should chop vegetables first. ***

*** When the tools are activated, they will eat up the ingredients. You cannot take them off from it or put objects into it. ***

*** The food orders are keep coming. You should finish as many dishes as possible and finish every dish as soon as possible. Please deliver the order to the servingtable when it is finished. ***

*** The dish will expire after the lifetime reaches 0 and it's not at the servingtable. Please avoid this. *** Here are the recipes:

```

Cook porkMeatcake at:
-- location: blender
-- with ingredients: pork, flour,
Cook salmonSashimi at:
-- location: choppboard
-- with ingredients: salmon,
Cook tunaSashimi at:
-- location: choppboard
-- with ingredients: tuna,
Cook mixedSashimi at:
-- location: mixed
-- with ingredients: salmonSashimi, tunaSashimi,
The following objects are available:
[-1] salmonSashimi
[-2] tunaSashimi
[-3] mixedSashimi
[-4] tunaSashimi
[-5] porkMeatcake
[-6] salmon
[-7] flour
[-8] pork
The objects are cooked using tools or are just base ingredients.
Among them, the following are base ingredients:
[-1] tuna
[-2] salmon
[-3] flour
[-4] pork
You can only obtain base ingredients from the storage initially.
Additional objects:
You can place up to infinite item into the storage@ You can place up to infinite item into the storage@ You can place up to infinite item into the servingtable@ You can place up to infinite item into the servingtable@ You can place up to 1 item into the choppboard@ You can place up to 1 item into the choppboard@ You can place up to 1 item into the mixer@ You can place up to 5 item into the mixer@ You can place up to 5 item into the mixer@ You can place up to 5 item into the mixer@ ** Only ** the following tools are available:
storage@, servingtable@, choppboard@, chopper0, mixer0, mixer1, You cannot pick up these tools. You can only use those tools at the corresponding location.

```

Figure 8: The MINDAGENT system prompt example

```

There are 2 agents available, so you can execute 2 actions at a time.

Goal: porkMeatcake
t=0
-state:
at(agent0, servingTable0)
at(agent1, servingTable0)
hold(agent0, None)
hold(agent1, None)
inside(blender0, None)
inside(chopboard0, None)
inside(servingTable0, None)

-action:
***  

goto_agent0_storage0  

goto_agent1_storage0  

***  

Goal: porkMeatcake
t=1
-state:
at(agent0, storage0)
at(agent1, storage0)
hold(agent0, None)
hold(agent1, None)
inside(blender0, None)
inside(chopboard0, None)
inside(servingTable0, None)

-action:
***  

get_agent0_flour_storage0  

get_agent1_pork_storage0  

***  

Goal: porkMeatcake
t=2
-state:
at(agent0, storage0)
at(agent1, storage0)
hold(agent0, flour)
hold(agent1, pork)
inside(storage0, None)
inside(chopboard0, None)
inside(chopboard0, None)
inside(chopboard0, None)
inside(pan0, None)
inside(servingTable0, None)

-action:
***  

goto_agent0_blender0  

goto_agent1_blender0

```

Figure 9: The MINDAGENT system partial one-shot demo example.

B TASK DETAILS IN CUISINEWORLD

Here we visualize different task graphs in CUISINEWORLD. In CUISINEWORLD, we provide tasks of different complexities to holistically evaluate the multi-agent system's performance. In addition, the environment is highly customizable and extendable. Users only need to modify the JSON files to add more tasks or modify existing tasks.

B.1 LEVEL 0

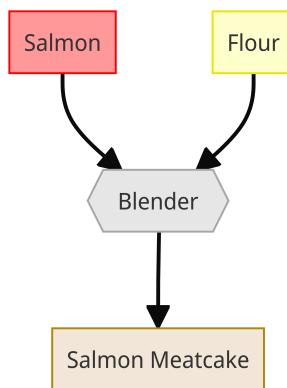
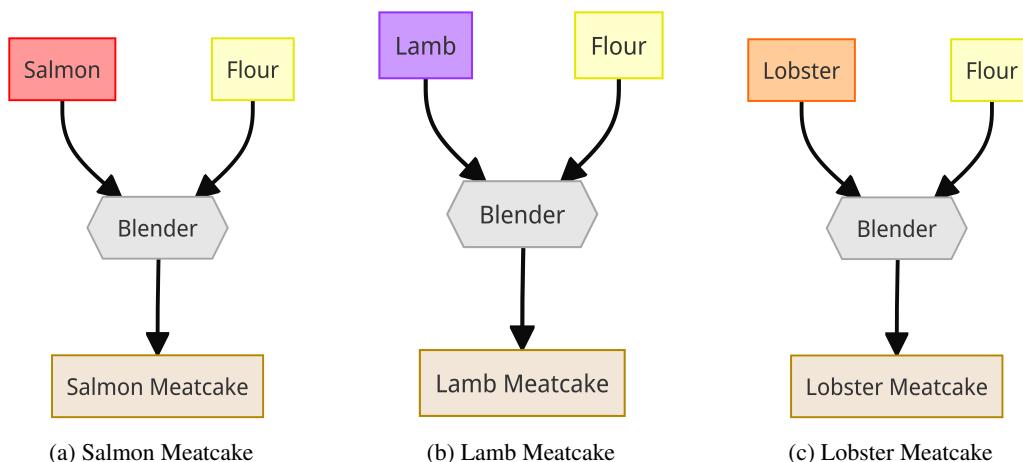
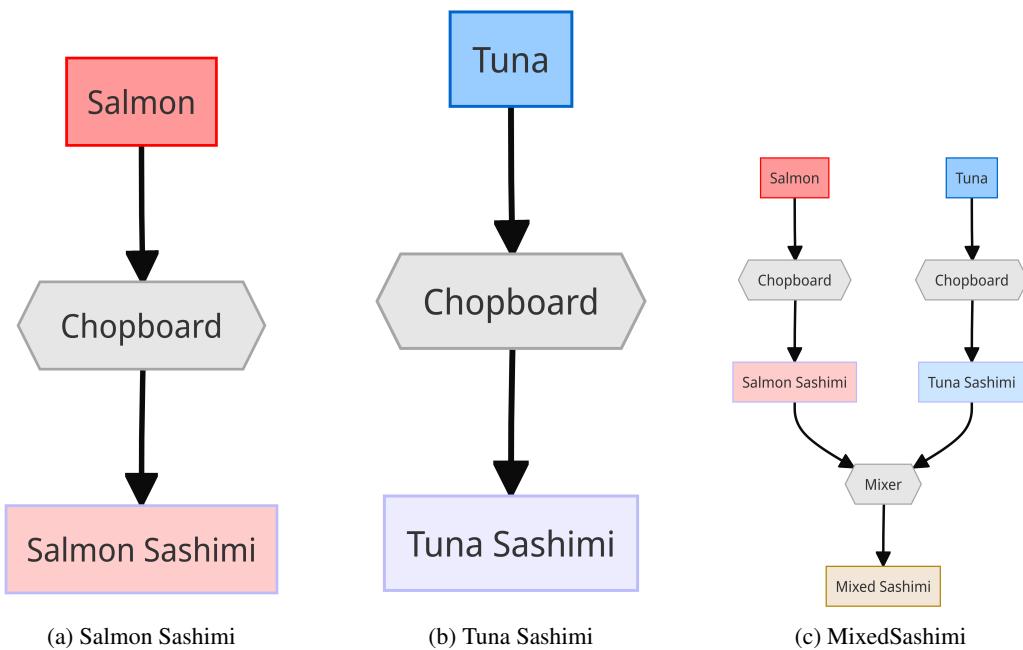


Figure 10: Salmon Meatcake

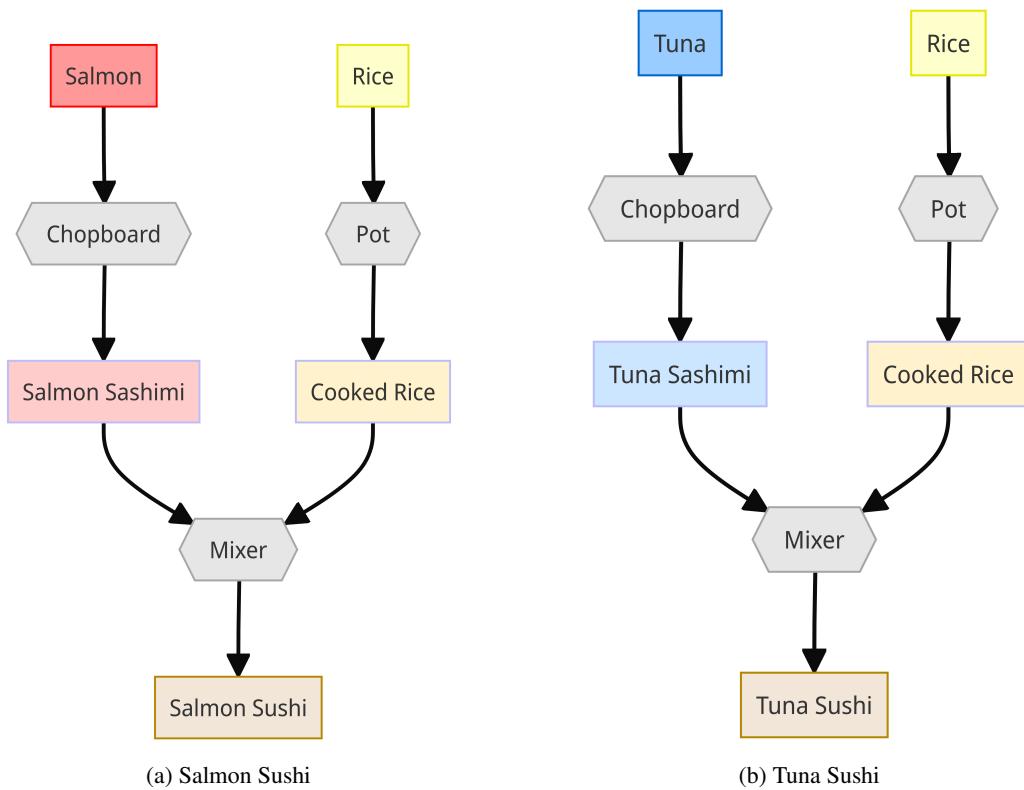
B.2 LEVEL 1



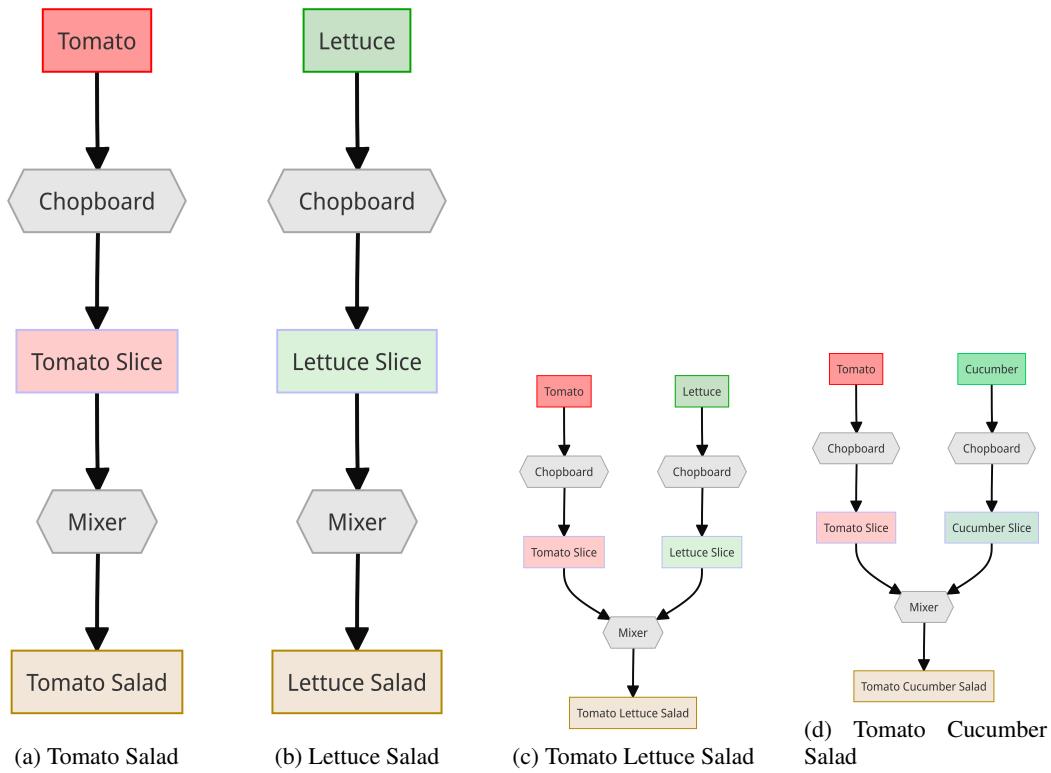
B.3 LEVEL 2



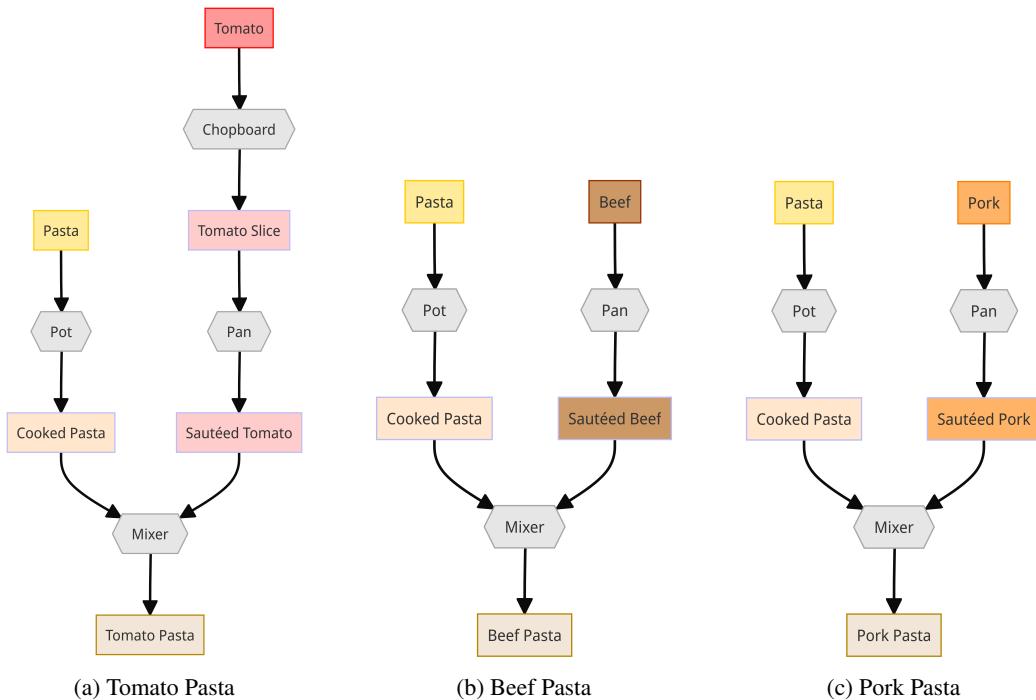
B.4 LEVEL 3



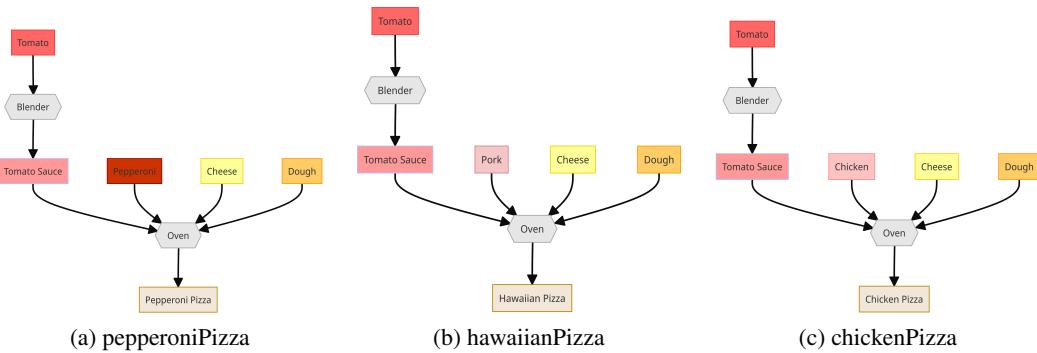
B.5 LEVEL 4



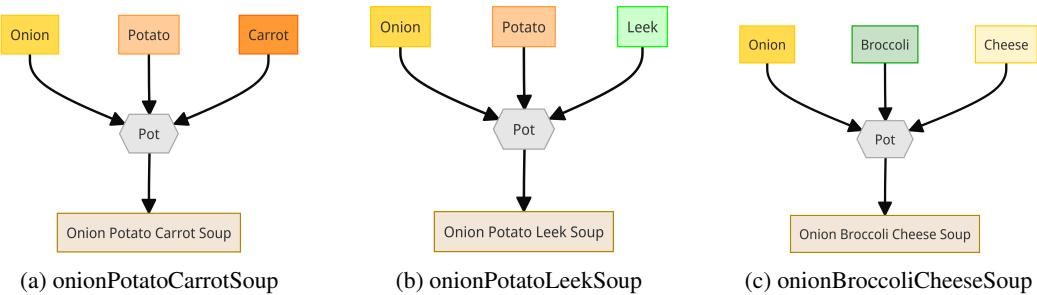
B.6 LEVEL 5



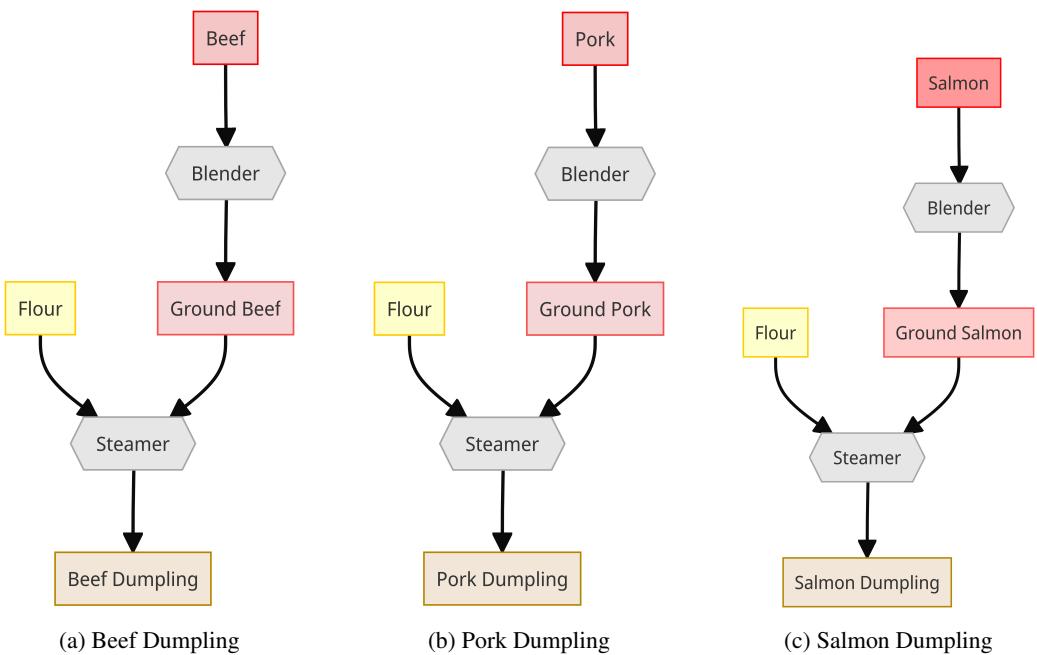
B.7 LEVEL 6



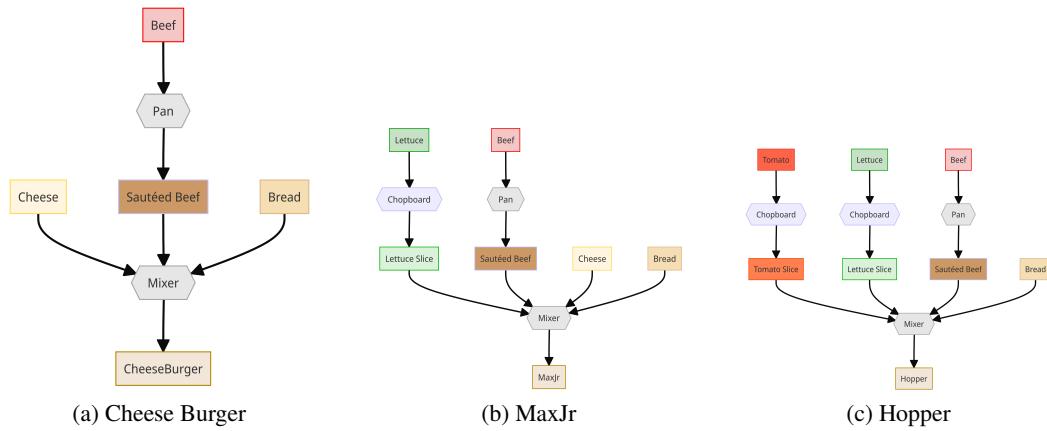
B.8 LEVEL 7



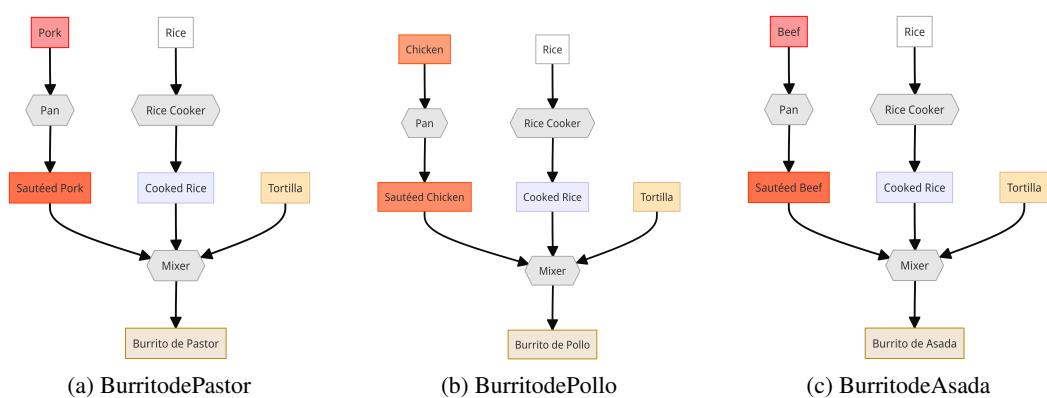
B.9 LEVEL 8



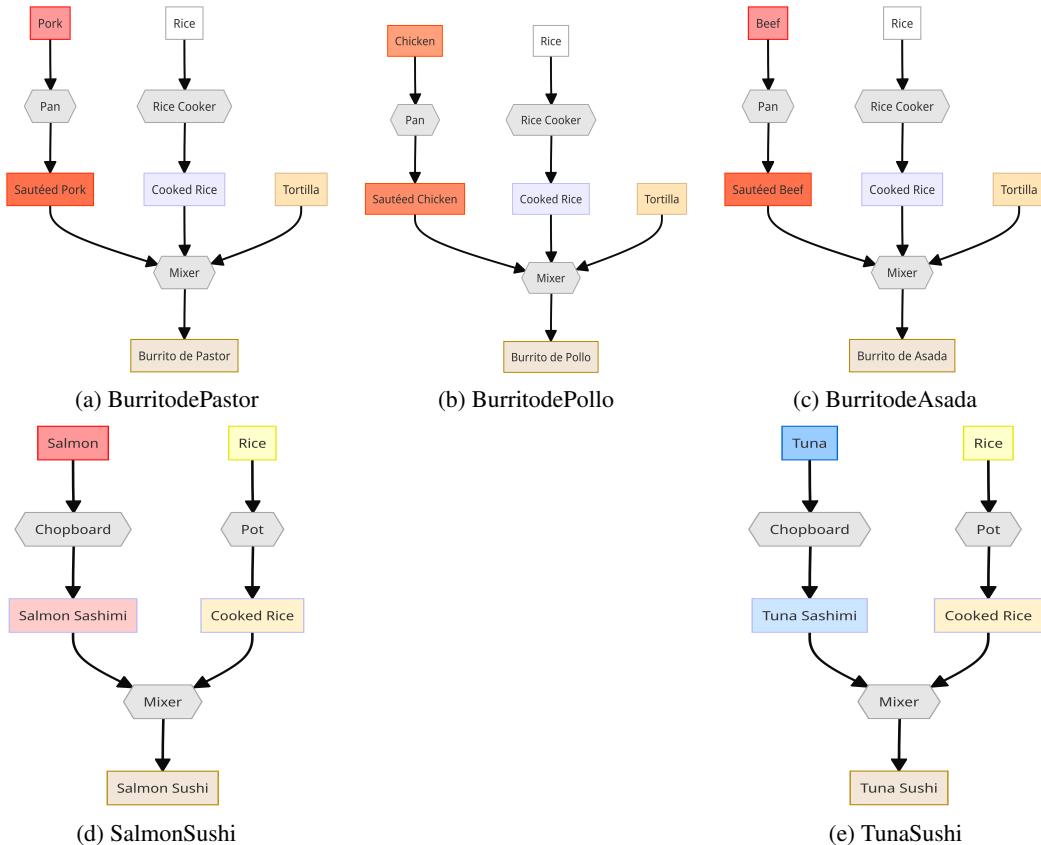
B.10 LEVEL 9



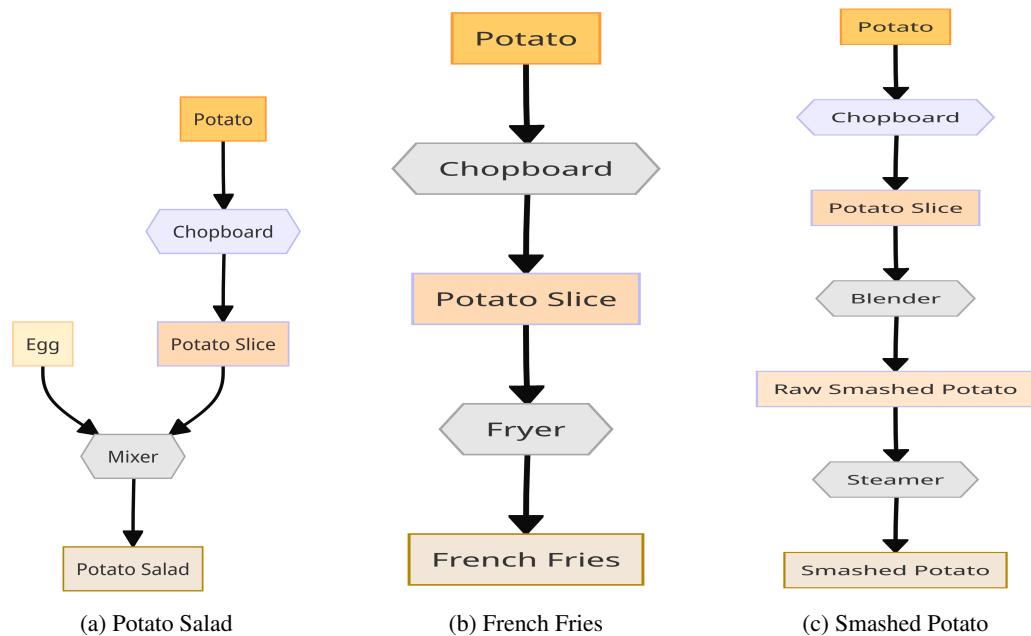
B.11 LEVEL 10



B.12 LEVEL 11

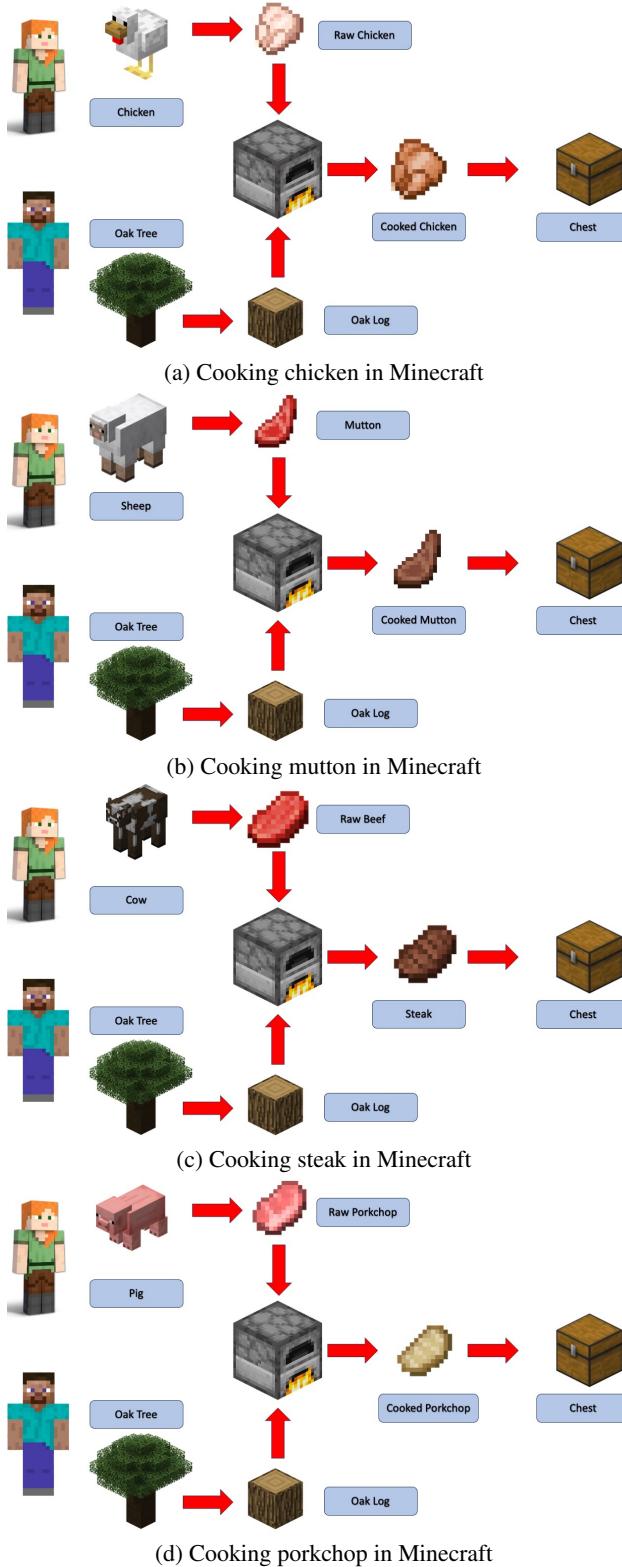


B.13 LEVEL 12



C MINECRAFT

Here we visualize the task graphs for different tasks in Minecraft.



D HUMAN EVALUATION INTERFACE

We use the human evaluation interface to test the human's perception of collaborative agents. This gives us a more controlled environment so users' perception of collaborative agents does not depend on their ability to control the keyboard and mouse, and their perception of collaborative agents does not depend on the latency and rate limits of GPT-4.

(a) Welcom Screen for human evaluation

(b) Human Evaluation Example

(c) Human Evaluation Example

(d) Human Instructions