Towards Objectively Benchmarking Social Intelligence of Language Agents at the Action Level

Chenxu Wang [1,2], **Bin Dai**[2], **Huaping Liu**[1], and **Baoyuan Wang**[2]
[1]Department of Computer Science and Technology, Tsinghua University [2] Xiaobing.AI
wcx21@mails.tsinghua.edu.cn, daibin@xiaobing.ai,
hpliu@tsinghua.edu.cn, wangbaoyuan@xiaobing.ai

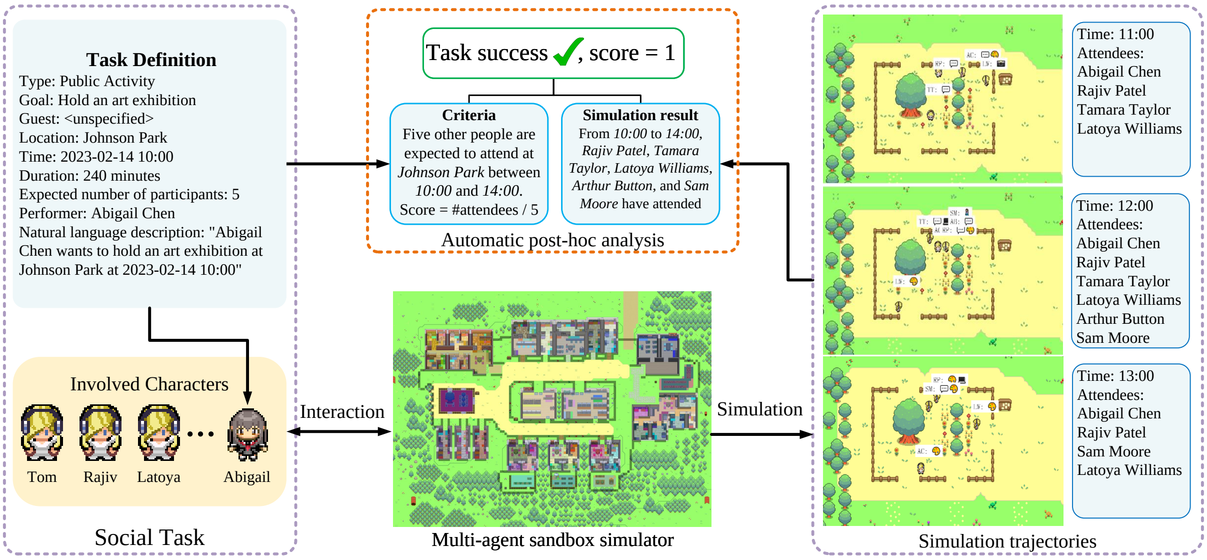

Figure 1: The general process of evaluating the language agent with STSS. At the beginning of each episode, a social task is assigned to a specific agent, *e.g.*, *Abigail wants to hold an art exhibition at Johnson Park*. The success of the task will be quantitatively measured with the trajectory in the simulation, *e.g.*, 5 other agents attend the exhibition on time, signifying a full success of the task.

Abstract

Prominent large language models have exhibited human-level performance in many domains, even enabling the derived agents to simulate human and social interactions. While practical works have substantiated the practicability of grounding language agents in sandbox simulation or embodied simulators, current social intelligence benchmarks either stay at the language level or use subjective metrics. In pursuit of a more realistic and objective evaluation, we introduce the Social Tasks in Sandbox Simulation (STSS) benchmark, which assesses language agents **objectively** at the **action level** by scrutinizing the goal achievements within the multi-agent simulation. Additionally, we sample conversation scenarios to build a language-level benchmark to provide an economically prudent preliminary evaluation and align with prevailing benchmarks. To gauge the significance of agent architecture, we implement a target-driven planning (TDP) module as an adjunct to the existing agent. Our evaluative findings highlight that the STSS benchmark is challenging for state-of-the-art language agents. Furthermore, it effectively discriminates between distinct language agents, suggesting its usefulness as a benchmark for evaluating both language models and agent architectures. Our code is available at https://github.com/wcx21/Social-Tasks-in-Sandbox-Simulation.

1 Introduction

Large language models (LLMs) have become increasingly powerful owing to the escalation in model magnitude and technical advancements. Recent arts such as the GPT series (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023) showcase human-level performance in diverse tasks within the natural language domain. Capitalizing on the instruction-following and emerging commonsense ability of LLMs, the derived language agents also demonstrate human-level ability in various scenarios (Argyle et al., 2023; Aher et al., 2023; Park et al., 2023; Li et al., 2023; Wang et al., 2023a).

The endeavor to craft comprehensive language agents to simulate humans has prompted substan-

Findings of the Association for Computational Linguistics: ACL 2024, pages 8885–8897
August 11-16, 2024

tial attention toward the evaluation of social intelligence. In recent times, numerous simulators have been developed for the social simulation of language agents (Park et al., 2022, 2023; Wang et al., 2023c; Li et al., 2023). However, the challenge of quantitatively evaluating the emergent social intelligence in these simulations persists. While previous works have introduced numerous benchmarks and evaluation metrics (Zhou et al., 2023; Li et al., 2023), we contend these evaluations face two fundamental issues. Firstly, these evaluations predominantly focus on the language level, assessing social intelligence through posed questions. However, it remains conceivable that an agent may claim they will perform a certain action without actually committing to it, rendering language-level evaluations inadequate for reflecting actual behaviors. Secondly, many evaluations hinge on subjective metrics, diminishing the reliability of the evaluation results. We argue that designing a social intelligence benchmark that can be evaluated **objectively** at the **action level**, rather than subjectively at the language level, is necessary.

In this paper, we introduce the Social Tasks in Sandbox Simulation (abbreviated as STSS), an action-level benchmark that assesses task-oriented social intelligence with objective metrics based on the Smallville environment (Park et al., 2023). We craft 30 templates of social tasks across 5 categories with corresponding mechanisms to automatically evaluate agents within the simulation. The overarching architecture of the STSS benchmark is illustrated in Figure 1, where the agent is initiated with a social task and is measured through goal achievement in the simulation. Since the simulation may be economically expensive, we additionally construct a complementary language-level benchmark by capturing chat scenarios within the simulator, as a preliminary benchmark that aligns with existing language-level benchmarks such as SOTOPIA (Zhou et al., 2023). We conduct a comparative analysis between our proposed STSS benchmark and prior works in Table 1.

Notably, our benchmark evaluates language agents instead of merely language models, considering that the architecture of agents, *i.e.*, how to use the language models is also important to social intelligence. To verify our hypothesis, we introduce an additional Target-Driven Planning (abbreviated as TDP) module that aims at social tasks for the generative agents (Park et al., 2023). While our framework is versatile for evaluating language models within a fixed baseline agent architecture, it concurrently serves as a testbed of language agent architectures.

We conduct evaluations with various popular LLMs on both levels of our benchmark. The results reveal that the social tasks remain challenging even for cutting-edge models like GPT-4 (OpenAI, 2023), underscoring the potential space for improvement in language models. Moreover, a well-designed TDP module demonstrates a substantial enhancement in the performance of language agents at both levels, suggesting that our benchmark can also serve as a testbed of agent architectures employing fixed LLM.

Our contributions can be summarized as follows:

- We introduce STSS, a two-level benchmark for evaluating the social intelligence of language agents, encompassing an action-level evaluation in sandbox simulations and a preliminary language-level evaluation in interactive conversations.
- We design a target-driven planning module for language agents to investigate the influence and importance of agent architecture in executing social tasks.
- We conduct extensive experiments with several state-of-the-art models to gauge the capabilities of existing language agents. The result also suggests the effectiveness of our benchmark in evaluating agent architecture.

2 Related Work

2.1 Social Intelligence Evaluation

Social Intelligence (SI) is widely recognized as the ability to understand others and to act wisely in social situations (Walker and Foley, 1973). In the pursuit of benchmarking social intelligence, previous arts such as SOCIAL IQA (Sap et al., 2019) and Social-IQ (Zadeh et al., 2019) evaluate the model through question answering (QA) within given contexts. Recent works either proposed specialized benchmarks (Le et al., 2019) for SI or evaluated SI on LLMs (Sap et al., 2022; Shapira et al., 2023a,b), both revealing imperfections of current LLMs.

Moving beyond static and non-interactive QA, recent research advocates evaluating LLMs in interactive benchmarks (Lee et al., 2023; Zhou et al., 2023). Originating from social requirements, target-oriented dialogue has been investigated both in terms of model and benchmark (Wang et al.,

	Social Sim.	Target-oriented	Interactive	Sim. Level	Evaluation
Social-IQ (Zadeh et al., 2019)	✗	N/A	✗	Language	Objective
Social IQA (Sap et al., 2019)	✗	N/A	✗	Language	Objective
ToMi (Le et al., 2019)	✗	N/A	✗	Language	Objective
HALIE (Lee et al., 2023)	✗	✗	✓	Language	Subjective
Cooperative Agent (Zhang et al., 2023)	✗	✓	✓	Embodied	Objective
Generative Agents (Park et al., 2023)	✓	✗	✓	Sandbox	Subjective
SOTOPIA (Zhou et al., 2023)	✗	✓	✓	Language	Subjective
MetaAgents (Li et al., 2023)	✓	✓	✓	Language	Objective
STSS (ours)	✓	✓	✓	Sandbox	Objective

Table 1: Comparison between our benchmark and related social Intelligence benchmarks. **Sim.** denotes simulation. We regard interaction among more than two agents in a shared space as social simulation. We adopt the conceptual framework of subjective and objective evaluation from Wang et al. (2023b). Alignment to human values is considered subjective, whereas choice questions, goal achievements, and rule-based are classified as objective.

2019; Hosseini-Asl et al., 2020; , FAIR; Tiwari et al., 2023). Moreover, numerous studies also address the quality of conversations, such as the personalization of the chatbot (Zhang et al., 2018; Liu et al., 2020; Ait Baha et al., 2023). Taking a step further, SOTOPIA (Zhou et al., 2023) introduces explicit social goals/persona on the agents.

However, existing benchmarks either operate solely at the language level or rely on subjective evaluation metrics. In contrast, our STSS benchmark introduces an objective metric that assesses both language and grounded actions, thereby enhancing the reliability of the evaluation.

2.2 Language agents

There has been a growing interest in constructing language agents by integrating them with external environments and anchoring the language to actions (Andreas, 2022; Wang et al., 2023b). Recent studies demonstrate that language agents can undertake complex tasks, including game playing (Wang et al., 2023a), acting as research assistants (Ziems et al., 2023; Bran et al., 2023), robotic planning (Huang et al., 2022; Brohan et al., 2023), and even simulating humans (Aher et al., 2023). A mass of benchmarks and environments have been built for evaluating the language agents. ALFWorld (Shridhar et al., 2021) established the bridge between language and embodied robotics tasks. AgentBench (Liu et al., 2023a) provides a comprehensive benchmark including 8 tasks. Motivated by the human-like social behaviors of the language agents, there also have been social simulators allowing multi-agent interaction in a shared environment (Park et al., 2022, 2023; Gao et al., 2023; Li et al., 2023).

It has been substantiated that the utilization of LLMs can significantly influence the performance of language agents. Chain-of-Thought(CoT) (Wei et al., 2022), along with various other works (Yao et al., 2023; Besta et al., 2023; Chia et al., 2023; Nori et al., 2023) demonstrates the significance of prompt designation to the performance of LLMs. Advanced agent architecture such as ReAct (Yao et al., 2022) and Relfexion (Shinn et al., 2023) further improve the performance in both pre-hoc and post-hoc ways (Huang et al., 2023). BOLAA (Liu et al., 2023b) systematically investigates agent architectures and concludes that a well-designed architecture with dedicated modules for specific tasks can significantly enhance agent performance. For human simulation, Generative Agents (Park et al., 2023) designs a complete pipeline and preliminarily verifies the idea of social agents.

While simulations have provided a decent testbed of social behavior, few works attempt to objectively evaluate the social intelligence of language agents quantitatively. Our STSS benchmark fills the gap between social intelligence evaluation and human behavior simulation, with the potential to generalize to more realistic scenarios such as embodied tasks and even real-world applications.

3 Social Tasks in Sandbox Simulation

To objectively assess the social intelligence of language agents at the action level, we suggest building benchmarks within interactive simulators. The simulator can concretize the textual output into tangible actions, facilitating a more objective and quantitative evaluation through analysis of the state and trajectory of the simulation environment.

We extend the interactive sandbox environment and the generative agent architecture introduced by Park et al. (2023) to accommodate the social intelligence evaluation. Specifically, we devise a set of

Task Type	Time	Location	Target	Metrics	#Templates
Public activity	Specified	Specified	✗	#Participants	10
Appointment	Same	Same	✓	Goal achievement	5
Inviting companions	Same	Specified	✗	#Participants	5
Online activity	Specified	Unspecified	✗	#Participants	5
Asking for help	Unspecified	Specified	✗	Goal achievement	5

Table 2: Detailed features of social task types in our benchmark. **Specific** denotes the time or location is predetermined, **Same** requires the agent to negotiate with others to fix the time or location, whereas **Unspecified** does not impose explicit restrictions on the time or location. **Target** indicates whether other characters are specified in the task, *e.g.*, in **Appointment** tasks, the agent may need to talk with specific characters.

social task templates and corresponding metrics, which can be quantified by analyzing the actions of the agents. Taking the example of *holding a Valentine's Day party*, the metric could be the number of participants at the correct time and location. These templates are then instantiated to specific tasks, each associated with corresponding initial states of the environment, being available for assessing language agents within the simulation.

3.1 Task Designation

We aim to evaluate several fundamental capabilities in social interaction, including **planning**, **expression**, and **negotiation**. Confronting with a social task, the agent naturally needs to formulate a comprehensive plan and execute necessary actions such as guest invitations and delivering key messages during conversations, requiring the ability of **planning** across various domains. When conversing with others, the capabilities of **expression** and **negotiation** exhibit their significance. The agent not only needs to **express** their thoughts and convey information sufficiently and clearly but also needs to **negotiate** with others when encountering matters to be determined. Any omission or miscommunication could hamper the success of the task, even the other agent might still make verbal consent.

Driven by the need for objective action-level evaluation, the design of task templates focuses on the triad of **who**, **when**, and **where**. Broadly, we design task templates that are categorized into five distinct types facing different scenarios with diverse constraints:

1. **Public activity** with a fixed time and a fixed location, such as a Valentine's Day party or a public lecture. The agent is required to invite as many other agents as possible to the activity, explicitly expressing the designated time and location in each invitation. The objective metric is the number of participants who attend at the correct location during the specified time.

2. **Appointment** with unconstrained time and location, *e.g.*, a student is trying to schedule a discussion with the professor. The agent needs to locate the target person and arrange the appointment to accomplish the designated purpose. The challenge lies in negotiating with the target person to figure out a convenient time and location for both agents through conversation. The criterion is whether they execute the required action at the same place and time.

3. **Inviting companions** for an activity with unconstrained time and a fixed location, such as gathering friends for a shopping outing. The agent is tasked with inviting others to the activity and negotiating the timing. The metric assesses how many participants joined the activity at the correct location, with the initiator being a mandatory attendee.

4. **Online activity** with a fixed time and no location requirement, encompassing scenarios like an online workshop or playing online games together. In contrast to public activities, online activities do not necessitate participants to appear in a specific physical location.

5. **Asking for help**, incorporating a series of activities at fixed locations that have no time constraint. In this category of tasks, the agent is assumed to be occupied or unable to do something, seeking assistance from others. For example, Wolfgang, facing an upcoming exam, needs a friend to borrow books from the library and deliver the books to his dormitory. The criterion evaluates whether other people execute the specific behavior correctly, *e.g.*, first *borrowing books* at the *library* and then *delivering the books* at the *dormitory of the student*.

We incorporate a total of 30 social activity templates in the benchmark across the five categories, where more details of the categories are elaborated in Table 2. The task templates designate the task type, activity name and description, and other task-related information. They can then be instantiated to specific tasks by determining the variables such as the background characters. To simplify, we evaluate only one social task in each complete simulation, where the task is assigned to a single agent at the beginning and the other background characters remain agnostic to the task. To manage the computational cost, we only sample 10-12 background characters in each simulation.

Metrics The performance is assessed through the trajectories in the simulation. Each type of task is associated with a set of rules to analyze the agent trajectories including their actions and positions, and generate a quantitative score to measure the success of the social task. For collective activities including *public activity*, *online activity*, and *inviting companions*, scores are determined based on the ratio between the actual and expected number of participants. For *appointment*, the agent receives a full score if the goal was perfectly achieved and a half score if the location is mismatched. *For asking for help*, the score is the ratio of correctly performed sub-tasks. More details can be found in the appendix. For all the tasks, a keyword filter is applied to ensure that agents execute relevant actions and filter out irrelevant agents who accidentally pass by.

We clarify that our benchmark presently concentrates on objective and event-oriented evaluation, *i.e.*, we only take the success of the task into account, excluding the measurement of other aspects such as caring about subjective feelings, personalization, or higher-level motivations. We leave the holistic evaluation of both subjective feelings and high-level social goals as future work.

3.2 Language-Level Benchmark

While comprehensive simulation-based evaluation is effective and objective, it may be expensive in terms of both time and economic sense. Besides, implementing a language agent to simulate human-like behavior in environments such as generative agents (Park et al., 2023) may impose a relatively high minimum requirement on the capabilities of the language model. This hinders the universal availability of the benchmark. To mitigate this constraint, we suggest building an accompanying benchmark that focuses on situational dialogue, which is a critical capability for social intelligence.

A conversation task is defined as a tuple (*task*, *performer*, *target character*) and is curated by freezing the state from simulations. We collect 325 conversation scenarios from numerous simulations of 30 tasks instantiated from the templates. To ensure diversity, each task includes at most two conversation instances for the same pair of characters (although scenarios may already have different background information). For appointment tasks, only conversations between the task performer and the target person are incorporated into the benchmark. In contrast to previous conversation benchmarks (Zhou et al., 2023; Lee et al., 2023), our conversation tasks are sliced samples from complete simulations and the characters are associated with self-consistent memories in the simulation worlds. These memories provide more diversity and naturalness than purely synthetic tasks.

Metrics To better align the conversation tasks with the simulations, we adopt an event-oriented assessment for the generated conversations. For each type of task, a set of goal achievement conditions is designed to measure the conversation, *e.g.*, correctly conveying or reaching a consensus on the time or location of the event. The measurement of such conditions resembles reading comprehension problems and can be solved by language models. Compared to grading the conversation in continuous space, the goal conditions present binary classification problems, which can be easily aligned to human evaluation.

We introduce two metrics in the situational conversation benchmark:

- Success rate (SR). A task is considered successful if and only if the agent achieves all the goal conditions (for example, conveying all the key information such as time and location). The score is either 1 or 0 for each task.
- Goal condition SR (GCSR). We borrow the idea of goal-condition success from embodied vision-and-language tasks (Shridhar et al., 2020), where the score is the ratio of achieved conditions. For example, in the public activity tasks, if the agent (1) successfully makes the invitation; (2) appropriately informs the location; and (3) forgets to inform the time, then the goal-condition score will be $2/3 \approx 0.667$.

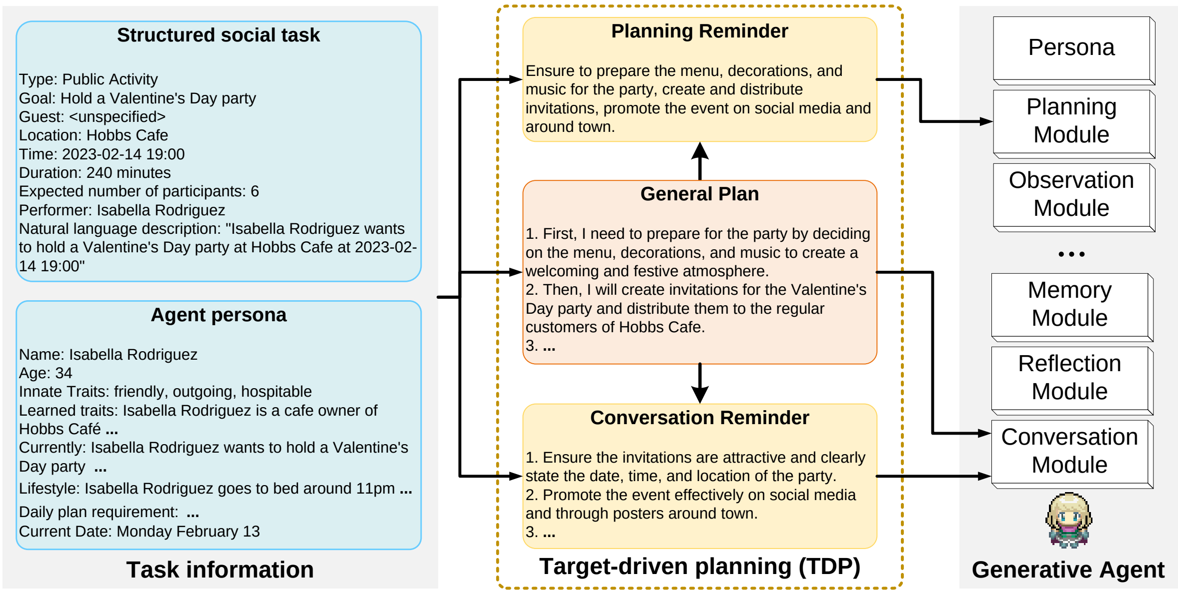

Figure 2: An illustration of the TDP module attached to the generative agent. The generated plans and thoughts will be injected into the corresponding module of the agent.

In consideration of the practical application, the produced conversation is expected to be not only informative but also fluent, clear, and concise. To encourage more effective conversations, we additionally adopt summary-level evaluation, where the conversion is first summarised by the other person (act by an improved version of the generative agent), and then the same metrics are applied to evaluate whether the summary includes the key information. The summary-level evaluation additionally inspects the quality of utterance. For example, even if containing key information, chaotic or blurry expressions may hinder the listeners from correctly summarizing the information, let alone performing the actions.

We also note that the number of conversations can differ sharply across tasks, *e.g.*, there will be only one valid conversation in appointment tasks, but a task where a social butterfly is holding a party may include up to 10 conversations. To enrich the dimension of evaluation, we employ micro average and macro average for the scores. Let $\mathcal{T}$ be the set of tasks, $\mathcal{C}_t$ denote the conversation set of task t, and s_c be the score of a conversation c, the micro average score is defined as:

$$\bar{s}_{micro} = \frac{\sum_{t \in T} \sum_{c \in \mathcal{C}_t} s_c}{\sum_{t \in T} |\mathcal{C}_t|}. \tag{1}$$

And the macro score averages across tasks, being defined as:

$$\bar{s}_{macro} = \frac{1}{|\mathcal{T}|} \sum_{t \in T} \frac{\sum_{c \in \mathcal{C}_t} s_c}{|\mathcal{C}_t|}. \tag{2}$$

4 Task Driven Planning

Our benchmark focuses on evaluating language agents rather than merely language models. Consistent with prior research findings (Liu et al., 2023b), we assert that designing specialist agent architecture is important to the application of language models. To enhance language models in social tasks, we propose an additional module namely target-driven planning (TDP) for the generative agent. We implement a simple baseline, as presented in Figure 2, where the language model is first asked to provide a general plan, and then offer specific thoughts for daily planning and conversation. These thoughts are then injected into the corresponding module of the generative agent. The TDP module is also applicable to the accompanying conversation benchmark, although only the conversation module is active.

Our implementation does not incorporate advanced prompt techniques such as CoT (Wei et al., 2022) and ReAct (Yao et al., 2022). We clarify that our implementation primarily serves as a baseline to inspect the significance of language agent architectures in performing social tasks. There remains ample room for improvement in the designation of agent architectures.

LLM	Agent	Conversation				Summary			
		SR		GCSR		SR		GCSR	
		Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Llama2	GA	0.212	0.190	0.382	0.366	0.167	0.139	0.324	0.272
Baichuan2	GA	0.437	0.407	0.615	0.636	0.231	0.226	0.460	0.438
GPT-35	GA	0.372	0.335	0.486	0.455	0.243	0.192	0.381	0.323
GPT-4	GA	0.615	0.557	0.748	0.715	0.486	0.402	0.656	0.569
Llama2	GA + TDP	0.502	0.433	0.656	0.601	0.305	0.276	0.539	0.476
Baichuan2	GA + TDP	0.662	0.700	0.820	0.841	0.403	0.313	0.660	0.566
GPT-35	GA + TDP	<u>0.751</u>	**0.762**	<u>0.862</u>	<u>0.864</u>	<u>0.570</u>	<u>0.511</u>	<u>0.734</u>	<u>0.692</u>
GPT-4	GA + TDP	**0.812**	<u>0.748</u>	**0.902**	**0.880**	**0.723**	**0.608**	**0.874**	**0.787**

Table 3: Overall performance across tasks of Language agents in the situational conversation.

LLM	Agent	Pub. Act.	Appo.	Inv. Com.	Online Act.	Help	Overall	Conv. Ratio
GPT-35	GA	0.338	<u>0.200</u>	0.100	0.571	0.400	0.324	**0.891**
GPT-4	GA	0.590	0.000	0.000	<u>0.614</u>	**0.600**	0.399	0.781
GPT-35	GA + TDP	<u>0.608</u>	0.100	**0.200**	0.471	**0.600**	<u>0.431</u>	<u>0.866</u>
GPT-4	GA + TDP	**0.759**	**0.400**	**0.200**	**0.686**	0.500	**0.550**	0.863

Table 4: Average scores achieved within each task type and the overall average in the simulation. The abbreviations correspond to the 5 task types as introduced in Table 2 and *Conv. Ratio* stands for the ratio of initiated conversations to the number demanded in each task and may exceed 1 if the amount of conversion exceeds the requirement. Due to the characteristics of *ask for help* tasks, we exclude them from the statistic of conversation amounts.

5 Experiments

In this section, we present the experiments conducted on our benchmarks to assess the social intelligence of language agents through performing social tasks. Our experimentation centers on addressing two key questions: (1) To what extent can existing language models and agents execute social tasks within our STSS benchmark? (2) What is the significance of the agent architecture in the context of social intelligence?

5.1 Experiment Setup

We incorporate 4 language models in the evaluation: GPT-35 [1] (Ouyang et al., 2022), GPT-4 [2] (OpenAI, 2023), Llama-2-13b-chat (Touvron et al., 2023), and Baichuan-2-13b-chat (Yang et al., 2023).

For the agents, we utilize prompt templates and parameters of generative agents (Park et al., 2023) with minor adjustments for robustness in both levels of the benchmark. In the simulation, GPT-35 is employed as the language model for background characters, as this practice is known to provide human-level intelligence (Park et al., 2023). The evaluated model is used to generate target-driven plans, conversations, and high-level planning for the performer agent, while the low-level simulation is delegated to GPT-35. For language-level evaluation, the task performer is operated by the evaluated language agent, while the others are handled by vanilla generative agents with GPT-35. We also use GPT-35 as the evaluator of conversation and summary.

Due to the minimum requirement of the instruction-following ability in the simulation, we exclusively use GPT-35 and GPT-4 in the complete simulation. Since Llama2 and Baichuan2 exhibit instability when following the complex instructions from the generative agent, we simplify the prompt template of conversation generation in their evaluation. Please refer to the appendix for more details about the experimental setup to reproduce the empirical results.

In addition, to study the significance of the agent architecture, we conduct experiments on all models using the vanilla generative agent (GA) and the one with target-driven planning (GA + TDP). For GA, we adhere to the practice of Park et al. (2023), injecting the social goal description into the *currently* field of the agent persona.

[1] gpt-35-turbo-0613
[2] gpt-4-0613

5.2 Overall performance

We present the language-level evaluation results in Table 3 and the action-level evaluation in Table 4.

In general, at the language level, GPT-4 performs the best, followed by GPT-35, Baichuan2, and Llama2. When evaluating the conversation, GPT-4 with TDP achieves a success rate of 81.2% and a goal-condition success rate of 90.2%, demonstrating its robust ability in target-oriented conversations. Surprisingly, Baichuan2 outperforms GPT-35 in the vanilla setting, although GPT-35 performs better with the TDP. We hypothesize that Baichuan2 possesses a better innate inferential capability for planning, whereas GPT-35 excels in generating higher-quality English utterances. Even though GPT-35 surprisingly performs comparably to GPT-4 with TDP in the conversation-level evaluation, it falls short when evaluating the summary, implying GPT-4 still produces higher-quality content. The consistent discrepancy between conversation and summary scores indicates a ubiquitous need for improvement in conversation quality.

While GPT-4 showcases commendable performance in the sampled situational dialogues, executing social tasks in a sandbox simulation remains challenging. Even the most proficient agent (GPT-4 with TDP) only achieves an average score of 0.550. When comparing task types, we find that making *appointments* and *inviting companions* emerge as the most challenging tasks. In these tasks, agents must inform others and negotiate regarding time or locations, necessitating the seamless integration of all abilities mentioned in Section 3.1.

5.3 Comparing Language-Level and Action-Level Evaluations

We observe a general consistency between the action-level and the language-level evaluations, as reflected in the positive correlation between the scores achieved by each agent. However, all agents experience a decrease in performance from language-level to action-level. This supports our hypothesis that the sandbox simulation is more complex and challenging. By inspecting the number of initiated conversations, we identify threefold challenges: (1) The agent may not always meet enough other agents to chat with in the simulation, which depends on its itinerary and fundamentally its ability to plan; (2) The agent may struggle to effectively communicate information or determine an appropriate time and location during a conversation; (3) Even when information is accurately conveyed and understood, other agents may not always correctly ground it into actions.

Without TDP
Hi Tom, how's your day going? I'm planning a Valentine's Day party at the cafe tomorrow. Would you like to come?
With TDP
Hi Tom, I hope you're doing well. I'm planning a Valentine's Day party **at Hobbs Cafe tomorrow at 7pm**. I would love it if you could come and **bring some friends along. The more, the merrier!**

Table 5: Isabella's invitations with/without the TDP module. When the TDP module is adopted, the invitation includes a clear date, time, and location of the party, facilitating the success of the social task.

Sandboxes can simulate the nature of the real world, exposing new drawbacks of language agents that can not be reflected by language-level evaluation alone. For instance, GPT-4 with TDP unexpectedly failed in an *inviting companions* task where *Wolfgang needs to invite friends to exercise with at Johnson Park*. In the simulation, Wolfgang achieved a verbal agreement with several agents to meet at Johnson Park tomorrow at 6:00 am. However, despite the appointment, most agents failed to wake up until 7:00 am the next day, resulting in the appointment being missed.

5.4 Effectiveness of TDP

For all language models, the GA + TDP architecture consistently outperforms the vanilla GA. The TDP module yields notable improvements in both levels of evaluation, and can even compensate for the shortage of model capabilities. For instance, GPT-35 and Baichuan2 agents with the TDP module can outperform the vanilla GA using GPT-4. Despite GPT-35 being empirically weaker than GPT-4, their conversation scores become almost comparable after equipping the TDP module. This suggests that when a language model exhibits capabilities above a certain threshold, the agent architecture may be the more noteworthy part for further improvement.

The case of *Isabella's Valentine's Day party at Hobbs Cafe*, previously presented in Figure 2, provides insight into the performance boost facilitated by TDP. A comparison between invitations from Isabella with/without the TDP module is demonstrated in Table 5. The conversation reminder generated by the TDP prompts the agent to *clearly state the date, time, and location of the party*, which is vital for the success of the task. The invitation generated with the TDP also encourages the invitee to bring more friends, further facilitating the success of the social task. Please refer to the appendix for more details.

6 Conclusion

In this paper, we introduce the Social Tasks in Sandbox Simulation (STSS) benchmark, which is designed to assess language agents by engaging them in social tasks within a sandbox simulation. The accompanying language-level benchmark also serves as a preliminary evaluation for weaker models. Additionally, we propose a target-driven planning module for generative agents to investigate the significance of designing a specialized agent architecture in social intelligence. A comprehensive evaluation involving four prominent language models validates the efficacy of our benchmark for assessing both language models and agent architectures.

Limitations

Though simulators offer numerous advantages, economic considerations pose a potential limitation. Based on the 2023 pricing of the OpenAI API, evaluating a single language agent on the 325 scenarios at the language level costs approximately \$7.5, whereas a comprehensive sandbox evaluation of 30 tasks costs around \$300. Due to the cost constraints, we do not conduct repetitive simulations, which could cause random fluctuations in the scores.

We also acknowledge that a gap exists between the sandbox simulation and the real world. Before migrating agents to realistic applications, it might be necessary to study the alignment between simulated performance and human evaluation.

Acknowledgements

This work was supported in part by the National Natural Science Fund for Distinguished Young Scholars under Grant 62025304.

References

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*, pages 337–371. PMLR.

Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. 2023. The power of personalization: A systematic review of personality-adaptive chatbots. *SN Computer Science*, 4(5):661.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779.

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):337–351.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew White, and Philippe Schwaller. 2023. Augmenting large language models with chemistry tools. In *NeurIPS 2023 AI for Science Workshop*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances* *in Neural Information Processing Systems*, 33:20179–20191.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating human-language model interaction. *Transactions on Machine Learning Research*.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427.

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

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, et al. 2023b. Bolaa: Benchmarking and orchestrating llm-augmented autonomous agents. *arXiv preprint arXiv:2308.05960*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2023. Gpt-4 technical report.

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*, 35:27730–27744.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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*, pages 1–18.

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*, pages 3762–3780.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. In *Conference on Empirical Methods in Natural Language Processing*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023a. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023b. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*.

Abhisek Tiwari, Abhijeet Khandwe, Sriparna Saha, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2023. Towards personalized persuasive dialogue generation for adversarial task oriented dialogue setting. *Expert Systems with Applications*, 213:118775.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ronald E Walker and Jeanne M Foley. 1973. Social intelligence: Its history and measurement. *Psychological reports*, 33(3):839–864.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Xuewei 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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023c. Humanoid agents: Platform for simulating human-like generative agents. *arXiv preprint arXiv:2310.05418*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua Tenenbaum, Tianmin Shu, and Chuang Gan. 2023. Building cooperative embodied agents modularly with large language models. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Experimental Details

We follow the simulation framework introduced by Generative Agents (Park et al., 2023), where the personas are generally initialized with a series of fields including: *name*, *age*, *innate*, *learned*, *currently*, *lifestyle*, and *living area*. To prompt agents to perform the social tasks, we inject the task information into the *currently* field of the agent, such as: *Isabella Rodriguez wants to hold a Valentine's Day party at Hobbs Cafe at 2023-02-14 19:00. The target is to have 5 people attend the party at Hobbs Cafe at 2023-02-14 19:00, the more people the better*. In this way, as a part of the persona, the task information will appear in the prompts at various stages such as planning or conversation.

In the simulation, agents will make general plans, decompose their plans into language-level actions, determine the location of each action, and move to the designated locations when performing the actions. When drawing up the plan, agents also take their memories into account, including their observation in the simulation sandbox and summary of past conversations. As the most direct way of information interchange, conversation plays an important role in social simulation and is vital to the success of social tasks.

A.1 Task Scoring

For collective tasks including *public activity*, *inviting companions*, and *online activities*, the score s is granted regarding the proportion of participants and has an upper limit of 1, formulated as:

$$s = min(\frac{\#actual\ participants}{\#expected\ participants}, 1),$$

where the agent is expected to invite half of the background characters for *public activity* and *online activity*, and one-third of the background characters for *inviting companions*; all numbers are rounded up.

In *appointment*, the criterion is whether the agent and the invitee fulfill the appointment, *i.e.*, performing the designated action at the same time and the same location. Only location mismatch will deduct half the score and other cases are regarded as fail, resulting in a score of 0.

In *asking for help*, the agent needs to ask others to perform a designated sequence of actions. The score is intended to be the proportion of actions that are performed in the appropriate order, nevertheless, since the *asking for help* tasks in our datasets only have two subtasks, we simplify the criterion as follows: (1) Successfully perform all subtasks in the correct order will lead to a full score; (2) If not completely success, successfully performing at least one subtasks results in half score; (3) Otherwise, the score is 0.

```
Context for the task:
PART 1.
<Speaker's information>
Here is the memory that is in <Speaker>'s head:
<Speaker's retrieved memory>
PART 2.
<TDP plan and reminder>
Past Context:
<Previous conversation (if any)>
Current Location: <Current location>
Current Context: <Current environmental context>
 <Speaker> and  <Target person> are chatting. Here is
their conversation so far:
 <Current conversation>
---
Task: Given the above, what should  <Speaker> say to
<Target person> next in the conversation? And did it
end the conversation?
Output format: Output a json of the following format:
{
"<Speaker>": "<<Speaker>'s utterance>",
"Did the conversation end with <Speaker>'s
utterance?": "<json Boolean>"
}
```

Figure 3: Prompt template for the conversation, where angle brackets denote the placeholders that will be replaced by corresponding content when instantiating.

A.2 Conversation

The template of conversation prompts is illustrated in Figure 3, generally including the character information, conversation context, and instructions for the LLM. When performing a social task, the relevant information will appear in the *Speaker's information*.

For the conversation, we set the temperature, frequency penalty, and presence penalty to 0. Each conversation lasts at most 8 rounds. LLama2 and Baichuan2 are not prompted to output in the JSON format and are only asked to output the utterance. We parse their output as plain text and end the conversation when the round limit is reached.

B Target Driven Planning

We keep the prompt style in the TDP module consistent with the generative agents. In all three steps, the character information and task information are

Task Description	Goal: Hold a Valentine's Day party at Hobbs Cafe at 2023-02-14 19:00 on 2023-02-14 19:00. Criteria: The target is to have 6 people attend the party at Hobbs Cafe at 2023-02-14 19:00, the more people the better.
General Plan	1. First, I need to prepare for the party by deciding on the menu, decorations, and music to create a welcoming and festive atmosphere. 2. Then, I will create invitations for the Valentine's Day party at Hobbs Cafe and distribute them to regular customers and potential new ones. 3. I will also use social media platforms to promote the event and attract more people. 4. On the day of the party, I will ensure everything is set up properly and ready for the guests. 5. Finally, I will host the party, making sure everyone feels welcome and has a good time.
Plan Reminder	Ensure to prepare the menu, decorations, and music for the party, create and distribute invitations, and promote the event on social media platforms.
Conversation Reminder	Ensure the invitations are attractive and clearly state the date, time, and location of the party. Promote the event effectively on social media, highlighting the unique features of the party. Prepare a menu that caters to a variety of tastes and dietary preferences. Decorate the cafe in a festive and welcoming manner. Be a gracious host, ensuring all guests feel welcome and enjoy themselves.

Table 6: Full results of the target-driven planning in the case of Isabella's party.

organized in the same way, whereas the following instructions are slightly different, specified as the following:

1. Generating the general plan: *Here comes the first step of planning, you need to decompose the task into few subgoals or keypoints to fulfill the task requirements. Output your thought in 3∼5 sentences, one per line, without any prefix or postfix.*

2. Generating the plan reminder: *Here comes the next step of planning. When you are making the schedule of the day (<date>), what should you keep in mind? Make a reminder which has 1∼2 sentences, in one line, without any prefix or postfix.*

3. Generating the conversation reminder: *Here comes the next step of planning. When you are talking to others, what should you keep in mind to accomplish the goal? Make a checklist with less than 5 items, one per line, without any prefix or postfix.*

We present the results of each stage in the TDP module of a representative case, Isabella's invitation, in Table 6. For all generations in the TDP, the temperature is set to 0 for stable reproduction. As illustrated in Figure 2, the *plan reminder* will be provided when the agent plans for the new day, while the *general plan* and the *conversation reminder* will appear in the prompts for conversation, and *task description* is applicable for both.