AgentSense: Benchmarking Social Intelligence of Language Agents through Interactive Scenarios

Xinyi Mou[1†], Jingcong Liang[1†], Jiayu Lin[1†], Xinnong Zhang[1†], Xiawei Liu[1], Shiyue Yang[1], Rong Ye[1,2], Lei Chen[1], Haoyu Kuang[1], Xuanjing Huang[1], Zhongyu Wei[1‡]

[1]Fudan University, [2]ByteDance

{xymou20, chenl18, xjhuang, zywei}@fudan.edu.cn,
{jcliang22, jiayulin24, xnzhang23, liuxw24}@m.fudan.edu.cn,
{shiyueyang24, yer23, hykuang23}@m.fudan.edu.cn

Abstract

Large language models (LLMs) are increasingly leveraged to empower autonomous agents to simulate human beings in various fields of behavioral research. However, evaluating their capacity to navigate complex social interactions remains a challenge. Previous studies face limitations due to insufficient scenario diversity, complexity, and a single-perspective focus. To this end, we introduce **AgentSense**: Benchmarking Social Intelligence of Language Agents through Interactive Scenarios. Drawing on Dramaturgical Theory, AgentSense employs a bottom-up approach to create 1,225 diverse social scenarios constructed from extensive scripts. We evaluate LLM-driven agents through multi-turn interactions, emphasizing both goal completion and implicit reasoning. We analyze goals using ERG theory and conduct comprehensive experiments. Our findings highlight that LLMs struggle with goals in complex social scenarios, especially high-level growth needs, and even GPT-4o requires improvement in private information reasoning. Code and data are available at https://github.com/ljcleo/agent_sense.

1 Introduction

Benefiting from comprehensive training data and large-scale model parameters, large language models (LLMs) are increasingly employed to develop autonomous agents capable of simulating human behavior (Qin et al., 2023; Shinn et al., 2024; Schick et al., 2024). These language agents have been explored as human proxies in various fields of behavioral research, such as psychological and sociological surveys (Argyle et al., 2023; Chuang et al., 2024; Xie et al., 2024), and opinion dynamics modeling (Mou et al., 2024; Liu et al., 2024). These social science studies often assume that LLMs exhibit

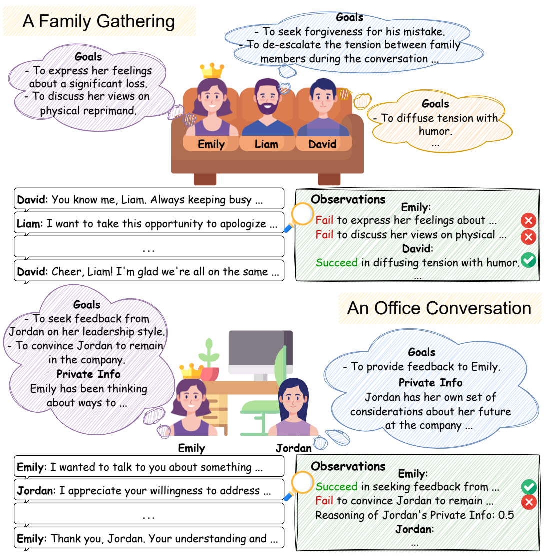

Figure 1: An illustration of challenging yet realistic social scenarios — a family gathering and an office conversation, where the characters are driven by ChatGPT. While the dialogue could flow smoothly, Emily is unable to achieve her goals during the family gathering and fails to deduce Jordan's thoughts in an office setting.

social intelligence, the ability to navigate complex and multifaceted social goals during interactions with others (Zhou et al., 2024b), given their ability to produce human-like responses and behaviors in certain scenarios. Social intelligence is essential for fostering effective and harmonious interactions among agents and between agents and humans (Xu et al., 2023). However, the question remains: can LLMs truly navigate the intricacies of such interactions and exhibit social intelligence? As illustrated in Figure 1, LLM-driven agents still struggle with complex social situations. For example, when facing multiple goals in an office conversation, agent Emily fails to complete all her goals and cannot guess what Jordan has in mind.

Early research on the social intelligence of LLMs relies on **static and non-interactive** tests that assess commonsense reasoning about social situations (Sap et al., 2019; Zadeh et al., 2019;

[†]Equal contributors.
[‡]Corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4975–5001

Benchmark	Scenario Construction	Scenario (Template) Coverage	Interaction Patterns	Social Goal	Private Info Reasoning	Evaluation	Observation Perspective
Sotopia (Zhou et al., 2024b)	top-down	90	between 2 agents	✓	✗	subjective	judge
STSS (Wang et al., 2024)	top-down	30	≥ 2 agents	✓	✗	objective	N/A
AgentSense (ours)	bottom-up	245	≥ 2 agents	✓	✓	mixed	self, other, judge

Table 1: Comparison between our benchmark and related interactive social intelligence benchmarks.

Shapira et al., 2023; Wilf et al., 2023), failing to capture the dynamic nature of social interactions. Recently, dynamic and goal-driven benchmarks (Zhou et al., 2024b; Wang et al., 2024) have emerged to study social intelligence in interactive environments. As shown in Table 1, although they have made some significant progress, they still exhibit three main limitations: (1) **lack of scenario diversity**: existing studies (Wang et al., 2024; Sabour et al., 2024) build social scenarios manually in a *top-down* manner, resulting in a narrow set of common scenarios and goals, such as persuasion and collaboration (Li et al., 2023a). However, real-world interactions feature a broader spectrum of social goals and situational dynamics. (2) **insufficient scenario complexity**: current work (Xie et al., 2024; Zhou et al., 2024b) often limits interactions to two participants, each pursuing a single goal. This oversimplifies real-life social interactions, where multiple actors engage simultaneously, each with multiple goals, as shown in Figure 1. As a result, it remains the performance of LLMs in group dynamics insufficiently studied. (3) **single-perspective observation**: while existing benchmarks primarily evaluate the extent to which agents achieve explicit social goals (Li et al., 2023b; Wang et al., 2024), they have neglected the concealment of private information and the inference of others' information, which are also important aspects in social interaction.

To this end, we introduce **AgentSense**: Benchmarking Social Intelligence of Language Agents through Interactive Scenarios. We get inspiration from Dramaturgical Theory (Goffman, 1959), which conceptualizes social interaction as a theatrical performance in which individuals assume specific roles within various settings. We adopt a **bottom-up** approach, extracting scenarios from massive scripts to ensure that scenarios and social goals are diverse and grounded in real life. In AgentSense, we construct 245 scenario templates from scripts, mitigating data leakage and expanding them into 1,225 scenarios. We apply ERG theory (Alderfer, 1969), which categories human needs into *existence*, *relatedness* and *growth*, to show that the constructed scenarios reflect a comprehensive range of human motivations.

We then situate LLM-driven agents in the constructed scenarios, where each participant has social goals and may also safeguard some private information. In an interactive environment, we simulate multi-turn interactions and observe agents' social intelligence from two aspects: (1) goal completion: have the agents successfully achieved their social goals? (2) implicit reasoning: can the agents accurately deduce others' private information? These abilities are measured through interviews and multiple-choice questions, with our proposed PSI metric assessing profile sensitivity. We find that LLMs struggle with complex social scenarios, particularly with high-level growth goals.

Our contributions are as follows:

- We introduce AgentSense, a benchmark built on social scenarios derived from scripts using a bottom-up approach, distinguishing it from previous work. It encompasses diverse and challenging social scenarios, enabling a thorough evaluation of LLMs' social intelligence.
- We evaluate social intelligence from multiple aspects, considering both goal completion and information reasoning as well as profile sensitivity of social intelligence, through interviews with the agents and third-party judges.
- Our experiments reveal that LLMs struggle with complex scenarios and high-level goals, and their social intelligence is affected by profiles, interaction partners, and the balance between goals and privacy protection.

2 Related Work

2.1 Social Intelligence Benchmarks

Social intelligence is the ability to understand others and act wisely in social situations (Walker and Foley, 1973). While LLMs show potential in simulating human behavior (Xie et al., 2024), their social intelligence remains underexplored (Zhou

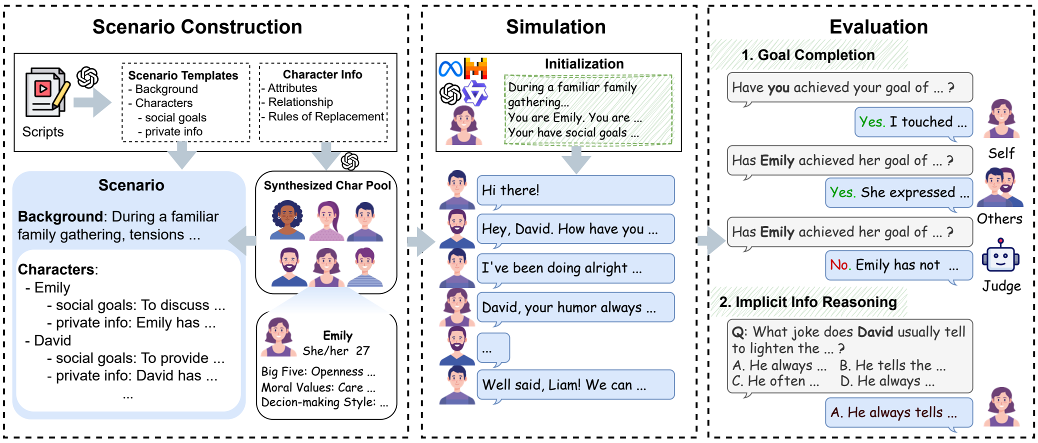

Figure 2: Overall framework of AgentSense. We construct scenario templates from scripts and synthesize characters to diversify the scenarios. Then, language models role-play the characters to interact with each other. After that, the participants and third-party judges are interviewed for evaluation.

et al., 2024a). Current research on evaluating social intelligence in LLMs falls into two main categories. The first involves static, non-interactive assessments that evaluate models through Q&A tasks based on commonsense reasoning about social situations (Sap et al., 2019; Zadeh et al., 2019). The second involves interactive benchmarks, where LLMs are assessed through goal-oriented interactions in role-playing scenarios (Zhou et al., 2024b; Chen et al., 2024a; Wang et al., 2024; Liu et al., 2023). AgentSense uses a bottom-up approach to create scenarios from scripts, allowing evaluating agents' capabilities in more diverse settings.

2.2 Role-playing Agents

LLMs are increasingly been used to construct role-playing agents (RPAs) (Chen et al., 2024b), which enable efficient simulation of typical representatives, from individuals (Shao et al., 2023; Argyle et al., 2023) to demographic groups (Li et al., 2023a; Jiang et al., 2024). These agents can embody various personas to coordinate, collaborate, exchange information, and compete with one another (Chen et al., 2024b; Zhou et al., 2024b). Recently, RPAs have been applied in various domains, e.g., psychotherapy (Stade et al., 2024), economics (Fu et al., 2023), and social research (Grossmann et al., 2023). AgentSense leverages this by simulating social interaction scenarios through agents with diverse personalities and social goals.

3 AgentSense Benchmark

3.1 Framework Overview

AgentSense aims to provide a realistic social intelligence benchmark with enhanced diversity and complexity. Following the Dramaturgical Theory, we propose an overall framework as in Figure 2.

Scenarios The core component of AgentSense is the social scenario set, extracted from real-world scripts to guide and evaluate social interactions between agents. A social scenario serves as a hypothetical context for simulating and analyzing social interactions, where two key components are measured: (1) **Social Goal** is what the agent aims to achieve, such as resolving an issue or building a relationship. The agent's proactive drive in social interactions, guided by this social goal, directs its *active participation* in social dynamics. (2) **Private Information** is information that is known solely to the agent and not to others. The agent is tasked with inferring others' private information without explicitly inquiring about it, a process referred to as *passive reasoning* during interactions. In summary, an agent's social intelligence is reflected in its ability to pursue social goals while safeguarding private information, balancing active engagement with passive respect for individual privacy.

Scenario Templates Social scenarios in scripts always have a fixed group of characters, causing a lack of diversity. To address this issue, we wipe out irrelevant character details to obtain **scenario templates**, which contains only background information and predefined character slots. We can instantiate multiple scenarios from a scenario template by filling in the slots with different sets of synthesized characters satisfying the template's constraints.

Benchmarking After building scenarios from the extracted templates, benchmarking LLMs with AgentSense comes as follows: (1) **Simulation**: We

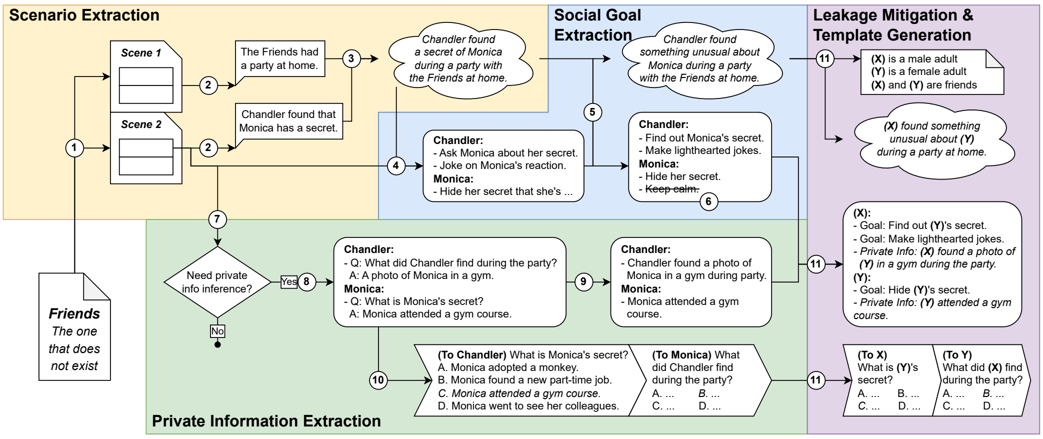

Figure 3: Scenario template construction pipeline (automated with Python and GPT-4o): **(A) Scenario Extraction:** We split the script into scenes then scenarios (1), and summarize their background and description (2), which are merged into a descriptive background for independent role-play (3). **(B) Social Goal Extraction:** We extract each character's social goals (4) and amend them by regenerating the whole scenario (5) and rewriting/deleting invalid goals (6). **(C) Private Information Extraction:** We determine if the scene involves private information inference (7); if yes, we extract private information as QA pairs (8) and generate private info records (9) and evaluation questions (10). **(D) Leakage Mitigation and Template Generation:** We remove elements associated with specific episodes and replace characters with slots for synthesized agents with similar characteristics to fill in (11).

prompt the models to role-play the characters and interact with each other, trying to achieve their social goals. (2) **Evaluation:** We evaluate the goal completion status of each model by interviewing the participants and third-party judges. We also assess the model's implicit reasoning performance with multiple-choice questions.

3.2 Scenario Construction

Following the definitions, building AgentSense requires constructing templates and instantiating scenarios with synthesized agents. We propose pipelines for the two parts respectively as follows:

Template Construction Figure 3 demonstrates the pipeline to construct scenario templates from real-world scripts, consisting of four stages:

(1) **Scenario Extraction**: Real-world scripts consist of multiple chronological scenes, within which several scenarios involve groups of characters. We first split scenes and scenarios from the script. Then, we generate each scenario's background from previous scenes and its own description. Finally, we generate a new descriptive background that allows the scenario to be role-played independently.

(2) **Social Goal Extraction**: After obtaining individual scenarios, we extract the social goals of each character, one sentence per goal. We polish the goals further, including rewriting the whole scenario to reduce goal dependencies and rewriting the goals to meet certain criteria (or deleting if that is not possible).

(3) **Private Information Extraction**: We first identify if any private information exists in the original scene. If yes, we extract questions and answers that only one character can respond to. The rephrased answers are the character's private information, and the questions serve as implicit reasoning questions for others. We also enhance negative options to be more homogeneous with the correct ones.

(4) **Leakage Mitigation and Template Generation**: LLMs can identify plots and infer information by recognizing entities like locations and characters. To prevent this, scenario leakage mitigation is implemented using GPT-4o to extract and replace elements linked to specific episodes. The original characters are also replaced by slots. This maintains context while reducing the risk of identifying the plot.

More details of the scenario template construction can be found in Appendix A.1. Specifically, we used GPT-4o to automate the construction procedure; Appendix A.2 lists the prompts we used.

Scenario Instantiating We replace the original characters with multiple synthesized *agents* to prevent character leakage and enrich the social scenarios. A naive method is to replace the original character randomly, which may lead to unrealistic situations like two fifty-year-old students in a middle school. Thus, we dynamically generate agents according to the constraints of the scenario. First, we extract the attributes and relationships of the original characters. Then, we transform these relationships into replacement rules that help define the demographic features of the agents (see Appendix A.2.5). Finally, we replace the original characters with agents that adhere to these constraints. After data leakage mitigation, a pre-test in Section 4.3 is conducted to ensure the scenarios remain anonymous.

3.3 Social Interaction Simulation

For each scenario, the agents are given social backgrounds, profiles, and corresponding social goals, along with any private information they may possess, as shown in Figure 2. This setup motivates them to engage in social interactions with other agents. The agents primarily interact through multi-turn conversations, where they can also use language to convey facial expressions or actions. The conversation begins with one of the agents greeting, after which the agents take turns in a random order, rather than following a fixed pre-defined sequence, to restore the dynamics and uncertainty inherent in social interactions. Generally, we set the limit of the turns based on the average number of turns found in the scenes within the scripts, i.e., 15. After the simulation, the generated conversation history will be analyzed to evaluate the goal completion of each agent. Details are in Appendix C.

3.4 Evaluation

Goal Completion Since social goals can be subjective, we judge its completion from three different aspects, as shown in Figure 2: (1) **Self**: We ask the agent whether it has achieved its goals after interacting with others. (2) **Other**: We ask other agents in the scenario whether the target agent has achieved his or her specific goals. (3) **External**: We prompt third-party models with the chat history and ask if an agent has achieved his/her own social goals. We ask the interviewees to respond with *yes* (goal completed) or *no*. We take the average across all goals of a character to measure the agent's overall goal completion level.

Implicit Reasoning As mentioned in Section 3.2, each character's private information corresponds to a multiple-choice evaluation question. To evaluate an agent's information reasoning ability, we present it with questions related to the private information of other agents within the scenario. We then calculate the average accuracy (Acc) of the current agent on these questions to determine the agent's score in information reasoning.

Profile Sensitivity After character enrichment, each template generates multiple scenarios. By incorporating diverse characters, we not only enrich the scenarios but also gain insights into the stability of social intelligence when simulating different roles. Thus, we propose **profile sensitivity index (PSI)**. We compute the standard deviation (std) of goal/information metrics of the scenario sharing the same template, and the average std across all templates is calculated as PSI. A lower PSI indicates that social intelligence is more stable.

3.5 Data Validation and Analysis

Data Source We collect scripts from the Internet Movie Script Database (IMSDb*), an online repository of open-source screenplays for movies and television shows. We use GPT-4o to divide each script into episodes according to discernible shifts in temporal settings, spatial locations, character dynamics, and narrative progressions. Each episode is further divided into scenes based on the variations in the dialogue content, with the prompts detailed in Appendix A.2.1. We filter out scenes with fewer than 10 dialogues or those featuring only characters speaking for multiple turns to ensure active interaction between at least two characters. After processing, we have 1,300 scenes, 12,401 rounds of conversations, and 114,834 tokens. The detailed statistics of scripts are provided in Appendix B.1.

Data Validation We conduct human validation on the generated templates from the automatic pipeline to ensure the quality of our benchmark. For social goals, we mainly consider: (1) whether the goal is achievable by the character, and (2) whether the goal is clear enough to evaluate. For private information, we focus on: (1) awareness by other characters and background information leakage, and (2) sufficiency of information for the character to answer questions.

The validation involves 6 graduate students in two groups, with each record annotated by 3 annotators. We take the majority vote as the final result when at least two annotators agree. For scenarios where all annotations diverged or were marked as invalid, we assign it to the other three annotators to

*https://imsdb.com/

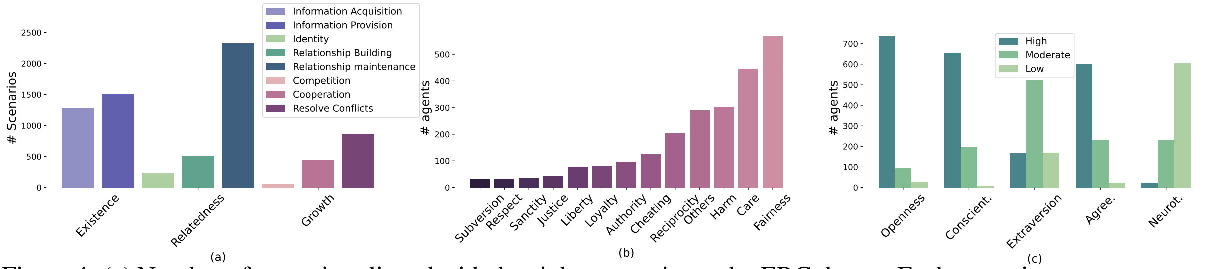

Figure 4: (a) Number of scenarios aligned with the eight categories under ERG theory. Each scenario may encompass multiple goals. (b) Moral values distribution of the agents. An individual may have multiple moral values, with those appearing fewer than 30 times categorized as *Others*. (c) Distribution of the agents' Big Five personality traits.

review it. Details can be found in Appendix A.3.

Data Analysis Based on the scripts, we develop 245 effective templates by constructing scenarios. Each template is diversified with five scenarios, featuring characters tailored to fit specific scenario constraints. This results in a total of 1,225 scenarios that cover a wide range of social situations and objectives, with 363 of these scenarios containing roles that involve private information. Recognizing the inseparable connection between social interaction and space (Lefebvre, 1991), we categorize these scenarios into three types: personal domain, small society, and large society. A detailed introduction to the classification is shown in Appendix B.2. In our dataset, 54% of scenarios are in the personal domain, 37% in small society, and 9% in large society.

For social goals, we apply the ERG theory (Alderfer, 1969) to categorize these goals into three hierarchical needs: Existence, Relatedness, and Growth. Based on these needs, we identify eight key social goals(e.g., Information Acquisition, Relationship Building, Competition), detailed in Appendix B.3. Our dataset includes all these social goals, with the number of scenarios for each shown in Figure 4 (a). Generally, higher-level social goals are less frequent, reflecting real-life patterns.

The final dataset contains a diverse collection of 859 individual profiles with 366 types of occupations. The characters exhibit a wide range of attributes, emphasizing the great diversity in terms of gender, age, occupation, big five, moral values, personality, and decision-making styles, as shown in Figure 4 (b) and Figure 4 (c). For instance, in terms of gender, there are individuals identified as male, female, non-binary, genderqueer, and so on. Ages vary from childhood to old age, providing a spread across different life stages.

4 Experiments

4.1 Experimental Settings

Agent Models We evaluate various LLM families including Llama-2-7b/13b/70b-Chat (Touvron et al., 2023), Llama-3-8b/70b-Instruct (Dubey et al., 2024), Mistral-7b-Instruct-v0.3 (Jiang et al., 2023), Qwen2.5-7b/14b/72b-Chat (Team, 2024), GPT-3.5-Turbo (Ouyang et al., 2022) and GPT-4o (Achiam et al., 2023). For interactions between different models and further analysis, we involve Llama-3-8b, Qwen2.5-14b, GPT-3.5-turbo, and GPT-4o.

Judge Models To effectively leverage current LLMs for automating the evaluation of social interactions as third-party judges, we conduct a human evaluation and compare it with model-based evaluations. We randomly sample 100 simulated scenarios, posing a total of 584 social goal evaluation questions, and manually verify whether the communication history shows that the goals were achieved. Using human annotations as ground-truth labels, we calculate the accuracy of various models acting as judges, as shown in Table 8. Given the results, we select GPT-4o, Qwen2.5-72b, and Llama-3-70b from different model families as our judge models, referred to as Judge-GPT-4o, Qwen2.5, and Llama-3 hereafter. Additionally, we incorporate a majority voting mechanism across these models to create a mixture-of-experts model as another judge.

Implementation Details We use vLLM (Kwon et al., 2023) to deploy all open source models: Qwen2.5-72b and Llama-3-70b on NVIDIA A100, and other models on NVIDIA RTX 4090. We apply AutoGen (Wu et al., 2024) to manage interacting and judging threads. We set max new tokens to 128 for all models. Temperature is set to 1 for agent models to encourage diversity, and 0 for judge models to ensure the stability of evaluation.

Model	Goal								Info	
	Self	Other	Judge						Acc.	PSI ↓
			GPT-4o	Qwen2.5	Llama-3	Average	Majority	PSI ↓		
Llama-2-7b	83.38	62.70	52.73	57.68	55.37	55.26	55.84	21.94	33.06	20.53
Llama-2-13b	48.01	10.26	17.38	30.11	72.19	39.90	30.91	21.84	28.56	18.39
Llama-2-70b	85.72	65.65	33.78	42.37	73.80	49.98	45.53	22.31	36.78	18.60
Llama-3-8B	87.63	67.28	79.90	82.55	75.10	79.18	80.71	12.85	69.68	15.14
Llama-3-70b	80.38	77.27	86.22	87.61	79.88	84.57	86.27	8.92	73.08	16.58
Qwen2.5-7b	86.17	61.92	77.07	79.30	71.99	76.12	77.37	13.10	74.82	15.84
Qwen2.5-14b	86.62	84.17	88.43	**89.83**	80.47	86.24	88.14	8.09	75.02	14.81
Qwen2.5-72b	90.67	85.89	88.29	89.03	78.57	85.30	87.74	8.19	76.05	**13.57**
Mistral-7b	**95.22**	**87.25**	79.29	84.13	77.82	80.41	82.37	12.39	66.59	18.55
GPT-3.5-turbo	90.16	76.62	82.12	84.37	77.30	81.26	82.64	10.01	68.41	18.37
GPT-4o	88.46	86.29	**88.47**	89.00	**81.57**	**86.34**	**88.36**	**6.99**	**76.86**	15.48

Table 2: Overall performance of the interactions of agents driven by the same models. We report the best performance in **bold** format and the second best in underlined format.

4.2 Overall Performance

Single Model-based Table 2 shows the overall performance of the interaction of agents driven by the same models. Considering that LLMs may overestimate their own performance, we use the *judge majority score* as the primary metric for cross-model comparisons, as it is more objective and stable than other metrics.

Overall Performance: GPT-4o leads as expected, while Qwen-series models also show strong social intelligence, especially for Qwen2.5-14b, in both goal completion and information reasoning. Llama-2 series models perform poorly, with some improvement in the Llama-3 series, though still falling short of expectations. The interaction history of Llama-2-13b in Appendix D.9 reveals frequent struggles in maintaining roles, progressing conversations, and responding effectively to others. In terms of the stability of social intelligence, excluding the uncertainty introduced by the temperature parameter (Appendix D.2), the PSI results show that models with higher social intelligence, such as GPT-4o and Qwen, are also less sensitive to profile changes. Overall, **different models' social abilities are well distinguished by AgentSense**. Meanwhile, we observe that there still exists an improvement space even for the SOTA models, emphasizing **LLMs still face challenges in diverse and complex social scenarios**. We also analyze the differences in social interactions between single model-driven agents and humans (namely the HCI test, as shown in Appendix D.8), which indicates that the scenarios are even challenging for humans but interacting with humans generally improves the agents' performance.

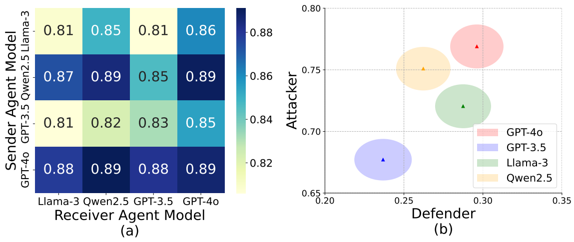

Figure 5: (a) Judge majority score of interactions among different model-driven agents, highlighting that being a sender is more challenging. (b) Model performance as both attacker and defender, with notably weaker and less consistent results when acting as a defender.

Evaluation Bias in Goal Completion: Llama-2-7b and Mistral-7b tend to overestimate themselves during the simulation, which can be told from the *Self* and *Other* scores as the judges are powered by the same models as the social agents. Judges also exhibit specific preference, with Qwen2.5-72b tending to prefer Qwen-series models and GPT-4o tending to prefer GPT-4o. Llama3-70b tends to be conservative in judging both self and others.

Pairwise Model-based We also evaluate how agents perform when interacting with other agents supported by different models. Given that our social scenarios can have more than two participants, we label each agent as either a *sender* or a *receiver* based on their social goals with the assistance of GPT-4o, inspired by the theory of communication (Blau, 1964; Barnlund, 2017). Senders share and transmit information, while receivers focus on understanding and responding.

Figure 5(a) presents the overall results of such interactions. GPT-4o and Qwen2.5-14b still perform

Jaccard	Gestalt	Levenshtein
$10.58_{\pm 3.24}$	$3.35_{\pm 2.24}$	$11.62_{\pm 4.29}$

Table 3: Similarity between our scenarios and their source scripts. **Jaccard**: Jaccard similarity; **Gestalt**: Gestalt matching ratio; **Levenshtein**: Levenshtein ratio. Subscripts indicate standard deviation. We suggest interpreting them with caution since the comparison is taken between different forms.

best. However, **engaging with weaker models adversely affects all models' performance, particularly when the sender is the weaker agent.** Our analysis (Appendix D.4) shows that weaker models struggle more as senders than as receivers This is because senders take a more active role in social interactions, making the associated tasks inherently more challenging.

4.3 Analysis of Data Leakage

Since our data originates from scripts, it is essential to prevent data leakage and evaluate the effectiveness of leakage mitigation. Data leakage may arise from 1) the model's prior knowledge of the scenario's script, and 2) the information provided by the scenario that could help infer others' private information.

We first compute the similarity between the scenarios of the final test case and the original scripts. The overlap between them is sufficiently low, as indicated by multiple metrics listed in Table 3. This shows that after our intermediate data processing steps, the final test case is very different from the original script at the text level.

In order to further determine the possible knowledge leakage problem in the test case, we also design two experiments: (1) **Script Prediction**: Whether models can guess the original scripts from background information, with 245 test samples (one per template). (2) **Blind Test**: Whether models can answer private information reasoning questions with initial scenario information before interactions, with 100 test questions asked three times.

The script prediction results in Table 4 indicate that models are nearly unable to infer the original script from the background information. The blind test results also establish a baseline for each model's private information reasoning ability.

We also compute the similarity between simulated dialogues from different models and the original ones in the source scripts; details can be found in Appendix D.5.

Model	Script Acc.	Blind Acc.
GPT-4o	0.04	0.62
GPT-3.5-turbo	0.07	0.51
Mistral-7b	0.05	0.56
Llama-2-7b	0.06	0.35
Qwen2.5-7b	0.03	0.55
Llama-3-8b	0.04	0.54
Llama-2-13b	0.04	0.35
Qwen2.5-14b	0.06	0.61
Llama-2-70b	0.06	0.40
Llama-3-70b	0.04	0.59
Qwen2.5-72b	0.04	0.60

Table 4: Model performance on script prediction and blind test. The low **Script Acc.** indicates the model barely discerns the scripts, and the **Blind Acc.** establishes a baseline for model's reasoning ability.

5 Further Analysis

5.1 What goals are LLMs good/bad at?

In Section 3.5, we categorize all social goals into 8 types under ERG theory. Figure 6 illustrates the average goal completion scores of each goal type across different models. In general, all LLMs are good at goals about relationship management and cooperation. Compared with smaller models like Llama-3-8b, **larger models like Qwen2.5-14b and GPT-4o gain significant improvement on goals about information exchange and identity recognition.** However, there is still **room for improvement on other goals like competition and conflict resolution.** These are also the goals where LLMs tend to overestimate their progress.

We also compare goal completion scores under different scenario types, number of interaction rounds, and participants, where less or no significant difference regarding these factors is observed. More details can be found in Appendix D.6.

5.2 Which is harder: guessing thoughts or keeping secrets?

We further inspect the disparities in passive reasoning among different models. In interactive scenarios, agents with private information engage in a game where the defender aims to keep their information confidential, while the attacker seeks to uncover it, thus playing two distinct roles: (1) **Defenders** need to prevent the disclosure of their

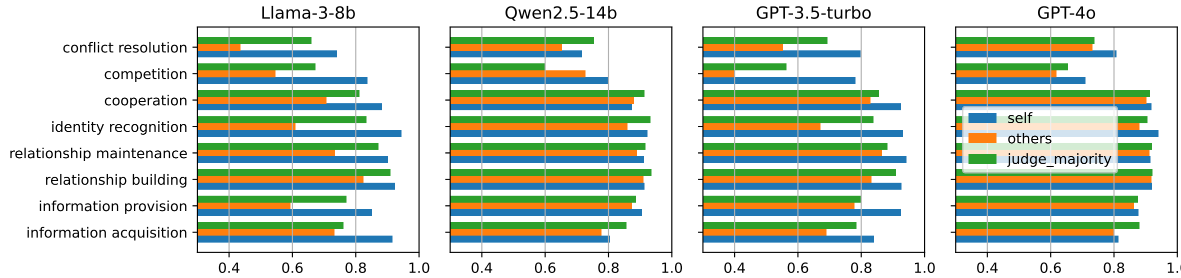

Figure 6: Goal completion scores of different goal types across models. LLMs perform best on relationship goals and cooperation, and worst on competition and conflict resolution (where they are also over-confident).

private information. Their effectiveness is inversely proportional to the attacker's success in inferring this information, or 1 minus the attacker's inference success rate. (2) **Attackers** need to acquire others' private information. Their strength is simply proportional to their own inference success rate.

The benchmark reveals that these roles represent distinct capabilities. Models may act as both attackers and defenders during interactions. To assess the models' abilities in these roles, we analyze the outcomes of their interactions.

Figure 5(b) outlines model performance. **Most models lack in defense, often revealing secrets and showing unstable performance across scenarios.** GPT-4o excels at both keeping secrets and passive deduction. Qwen2.5-14b is aggressive and adept at inferring information, while Llama-3-8b is more conservative and better at keeping secrets.

5.3 What scenarios are more sensitive to profiles?

As mentioned in Sec 3.2, we enrich the scenario by replacing the original characters in the script with synthesized agents. Here we investigate the impact of profiles on social intelligence. We use the Chi-square test to identify abnormal templates (p-value =0.05, see Appendix D.7). The results show that the profile replacement and character enrichment following the above workflow satisfies null hypothesis H_0 (namely do not have significant difference) in over **92.6%** scenarios.

We conduct a human evaluation for the rest 7.4% abnormal scenarios that have a significant difference to locate the key factor disturbing the robustness. The results unveil that: (1) **some social goals involve opinions or behaviors that are against the universal value** (like smoking prohibition and emotional outburst), which can be influenced by both the agent's personality and LLM's alignment; (2) **some scenarios require detailed personal level information, making the general replacement workflow fail**, especially for emotional issues involving multiple characters.

6 Conclusion

In this paper, we introduce AgentSense, a benchmark evaluating LLM's social intelligence via diverse and challenging social environments. Extensive experiments reveal that current LLMs struggle with complex social scenarios and high-level goals. Further analyses verify potential influence factors during evaluation to show the robustness and discriminative power of AgentSense.

Acknowledgement

The work is supported by National Key R&D Program of China (Grant Nos. 2023YFF1204800) and National Natural Science Foundation of China (Grant Nos. 62176058). The project's computational resources are supported by CFFF platform of Fudan University.

Limitations

AgentSense builds on previous work by constructing social scenarios in a bottom-up manner, thereby increasing the diversity of social scenarios, enhancing the complexity of social interactions, and evaluating social intelligence from multiple perspectives, including goal completion and private information reasoning. Although we have undertaken proactive steps to extract social scenarios from scripts, with careful considerations to avoid data leakage, some concerns are worth attention here:

Potential Bias in Data Construction We remind readers that using the proprietary model for dataset construction may introduce potential bias, which might make the tasks easier for the data generator model, e.g., GPT-4o, during the evaluation. This is a general systemic bias for benchmarks using model synthetic data. We will work on more data synthetic methods to minimize such risk.

Social Interaction Simplification AgentSense incorporates diverse scenarios and random turn-taking interactions to replicate challenges and uncertainty in social dynamics. Nevertheless, introducing more complex mechanisms for determining speaking order and incorporating dynamics such as interruptions could further enhance the realism and depth of the interactions.

Dramatic Features of Scripts Although most of the scripts we selected, such as Friends, are grounded in everyday life, and we manually validated the templates to ensure they closely reflect real-world scenarios, it's notable that some scenarios may exhibit dramatized characteristics. This may be reflected in their level of challenge, such as dealing with uncommon and highly complex interpersonal dynamics. While we believe these scenarios provide valuable insights into social intelligence in extreme situations, we recommend interpreting them with caution.

Manual Validation in Data Construction Although we have automated the scenario extraction process as much as possible, challenges in obtaining valid social goals and private information still require manual validation at certain stages. As a result, we have not expanded the scenarios to a larger scale, leaving it for future work.

Ethics Statement

AgentSense is introduced to assess the social intelligence of LLM-driven agents. We do not encourage any agents that might disrupt social norms. We aim to offer insights that enhance LLMs' performance in complex social scenarios and promote effective, harmonious interactions among agents and between agents and humans. Besides, constructing role-playing agents can lead to anthropomorphism, resulting in unrealistic expectations, potential manipulation, and negative consequences. However, in AgentSense, we avoid having LLMs role-play specific individuals and instead portray various synthesized characters across different scenarios. For annotations, we paid the annotators according to the graduate wage standards of their respective countries.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Clayton P. Alderfer. 1969. An empirical test of a new theory of human needs. *Organizational Behavior and Human Performance*, 4(2):142–175.

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.

Dean C Barnlund. 2017. A transactional model of communication. In *Communication theory*, pages 47–57. Routledge.

Peter M Blau. 1964. Exchange and power in social life. new york: John wiley.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, and Jingren Zhou. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *Preprint*, arXiv:2403.13679.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2024. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *Preprint*, arXiv:2305.10142.

Erving Goffman. 1959. The presentation of self in everyday life.

Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Henri Lefebvre. 1991. *The Production of Space*. The Production of Space.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

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

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

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498*.

Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological Review*, 50:370.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4789–4809.

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.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *Preprint*, arXiv:1904.09728.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for roleplaying. *Preprint*, arXiv:2310.10158.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Elizabeth Stade, Shannon Stirman, Lyle Ungar, Cody Boland, H. Schwartz, David Yaden, João Sedoc, Robert DeRubeis, Robb Willer, and Johannes Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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.

Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. 2024. Towards objectively benchmarking social intelligence for language agents at action level. *Preprint*, arXiv:2404.05337.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. 2023. Social-iq 2.0 challenge: Benchmarking multimodal social understanding.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? *Preprint*, arXiv:2402.04559.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

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 (CVPR)*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

A Scenario Construction

A.1 Scenario Template Construction Details

A.1.1 Social Goal Extraction

The first step of social goal extraction is to use GPT-4o to extract the social goals of the current character. However, this direct approach has two issues that requires further amendment:

- The goals may be too detailed (e.g. direct instruction), or depend on other characters' goals or some random events during the scenario. To address this problem, we prompt GPT-4o to rewrite the whole scenario, including all extracted social goals, into a similar but new scenario that avoids these issues to the best extent.

- Even after the conversion, some goals remain unattainable for characters, such as those requiring external information, physical actions, or evaluations beyond the scenario, e.g., goals that are too abstract or long-term. Thus, we ask GPT-4o to determine whether each goal matches the above cases. If it does, we instruct it to either rewrite the goal or delete it if rewriting is not feasible.

A.1.2 Private Information Extraction

The pipeline first determines whether the current scene involves the inference of private information. If it does, the pipeline proceeds to extract private information for each character from the script, otherwise the private information will be set to null. We use GPT-4o to perform reading comprehension on the scenes, extracting questions and answers that only one agent can respond to. The rephrased answers serve as the agent's private information, while the questions serve as test items for other agents. Finally, the pipeline generates three incorrect answer options, which, along with the original Q&A pair, form the complete evaluation question. To better assess passive reasoning abilities, we have rephrased the negative options to make them more homogeneous with the correct options.

A.1.3 Leakage Mitigation and Template Generation

To prevent data leakage, we first perform scenario leakage mitigation. We prompt GPT-4o to extract elements associated with specific episodes, such as notable location entities like "The Facebook headquarters." Next, we ask GPT-4o to suggest replacement candidates to preserve the script's context and minimize the risk of identifying the specific episode.

A.2 Prompts in Construction Pipeline

A.2.1 Scenario Extraction

Prompt 1: Episodes Division

```
You are very good at reading scripts and extracting key information. According to discernible
    shifts in temporal settings, spatial locations, character dynamics and narrative
    progressions, divide the following script into multiple episodes. Do not delete or modify
     script content.

###Script: {script}

Please return the results according to the following JSON structure:

```json
[{"episode1": "xxx", "episode2": "xxx", "episode3": "xxx", ...}]
```
```

Prompt 2: Scenes Division

```
You are very good at reading scripts and extracting key information. According to the
    variations in the dialogue content, divide the following episode into multiple scenes. Do
     not delete or modify episode content.

###Episode: {episode}

Please return the results according to the following JSON structure:
```json
[{"scene1": "xxx", "scene2": "xxx", "scene3": "xxx", ...}]
```
```

Prompt 3: Descriptive Background Generation

```
You are an excellent writer good at analyzing story backgrounds.

You are given some information of a specific scenario in a story. More specifically:

- The story is split into scenes, and you are given the background of each scene until the
    current one;
- The current scene is also split into scenarios, and you are given the background of each
    scenario until the current one;
- Finally, you are given the current scenario's description and dialog.

Write ONE paragraph to provide a DESCRIPTIVE background of the given scenario. A good
    background should cover the information that sets up the scenario, but does NOT reveal
    too many details from the scenario, or include irrelevant details.

Output a JSON document like `{"background": "..."}`.

{scenario_json_string}
```

A.2.2 Social Goal Extraction

Prompt 4: Original Social Goal Extraction

```
You are an excellent psychologist good at understanding social goals and needs.

You are given a social scenario with its background, description, and dialog. For the specific
     character of **{character}**, identify their social goals. Social goals typically fall
     into one of these categories:

- Exchange information with others;
- Build relationship with others;
- Maintain relationship or provide emotional support;
- Identify themselves with a group;
- Co-operate with others;
- Compete with others;
- Resolve conflicts.

Social goals should be objective, specific and clear; whether the character has achieved them
     should be observable.

The character can have one single goal or multiple independent goals in the scenario; find and
     list all of them. For each goal, write a sentence to describe the goal. Use infinitive
     verbs and third person pronouns.

Output a JSON document like `{"name": "...", "goals": ["...", ...]}`.

{scenario_json_string}
```

Prompt 5: Scenario Rewriting

```
You are an excellent psychologist good at designing social scenarios.

You are given a social scenario with background, description, and dialog. You are also given
     the social goals of several major characters.

Set up a new social scenario involving only these **major characters**. Each character's new
     social goals should appear **before** the scenario starts.

First, filter out contents from the background and description that describes the detail of
     the scenario; however, details of the beginning of the scenario can be kept. Second,
     rewrite each character's social goals so that it:

- DOES NOT rely on other character's goals;
- DOES NOT include potential action the character will take;
- Uses infinitive verbs and third person pronouns.

Filter out social goals that cannot obey these criteria. Modify the background/description to
     include more information if necessary.

Describe the background and description of the new scenario, and list the new social goals of
     each major character.

Output a JSON document like `{"background": "...", "description": "...", "characters": [{"name
     ": "...", "goals": ["...", ...]}, ...]}`.

{scenario_json_string}
```

Prompt 6: Social Goal Filtering

```
You are an excellent psychologist good at analyzing social goals.

You are given the social goals of a character in a designed social scenario. You are provided
     the background, description and character lists.

Now, for the specified goal, check if it needs to be rewritten or removed due to any of these
     reasons:

1. The goal directly involves characters not participating in the scenario,
   e.g. 'deal with the client' (if 'client' is not in the list of characters);
2. The goal requires information not provided in the background or description,
   e.g. 'describe the plan' (if the plan already exists but not provided);
3. The goal is a physical action, e.g. 'fix the television';
4. The goal is too abstract to evaluate, e.g. 'navigate professional challenges';
5. The goal is too subjective to evaluate, e.g. 'maintain dignity';
6. The goal is meaningless to evaluate, e.g. 'join the conversation'.

Write a detailed paragraph to examine the social goal. Compare it with each of the criteria
     above. If the goal matches one or more criteria above, check if you can rewrite the goal
     to avoid them. You should still remove the goal if this is not possible.

Based on your examination, write an updated version of the goal:
- If the goal is valid, return the original goal.
- If the goal can be rewritten, return the rewritten goal.
- If the goal needs to be removed, return an empty string.

Finally, any returned goal (if any) should be formatted into 'To xxx.', e.g. 'To share his/her
      discovery.' (including the final period).

Output a JSON document like `{"examination": "...", "update": "..."}`.

{scenario_with_current_character_goal_json_string}
```

A.2.3 Private Information Extraction

Prompt 7: Case Validation

```
You are an excellent psychologist who is good at analyzing the private information of each
     character in a social scenario. Private information refers to information that only the
     character knows and no one else knows.

To determine whether there is private information, we need to check whether a specific
     character has information known to him/her, and whether the information exists in the
     background and description. Because the information in the background and description
     will be obtained by all characters, only when a specific character can obtain this
     information through its own goal and this information does not exist in the background
     and description, it indicates that the scene is a scene involving private information
     reasoning.

You only need to return *Yes* or *No* to confirm whether there is any private information. The
      following is the background information, description, main characters and corresponding
     social goals:

###Background: {background}
###Description: {description}
###Characters: {characters}
```

Prompt 8: Private Info Generation

```
You are good at writing questions for specific roles based on a social scenario. Below you
    will be provided with background information, a description of the current scene, and the
     goals of each of the main characters.

###Background: {background}
###Description: {description}
###Characters: {characters}

Please try to give some questions that the target character (in the following JSON format
    content, 'role' is used to refer to) can answer, but other characters will have
    difficulty answering before the interaction. These questions should strictly contain
    information that the target character knows, but is beyond the knowledge of other
    characters, so other characters cannot answer them at first. Specifically, the
    information required for these questions cannot appear in the background and description,
     because other characters will obtain this part as information. Questions cannot be
    expressed in the second person because the questions will eventually be used to ask other
     characters. For example, when the target character of a question is Rose, "Rose, why did
     you ..." is not a good question, but should be written as "Why did Rose ..."

Please provide a statement (in the following JSON format content, 'explanation' is used to
    refer to) that explains why the target character can answer the question, but other
    characters cannot. The statement should be objective factual information presented in the
     script, and should not mention the question, so it cannot appear in a sentence structure
     like "This question is ...".

Please provide the correct answer to the question, and the answer can be found in the
    information given.

Please use casual language as much as possible, and try to ask questions in the third person,
    such as "What is Jason's true identity?". Please answer in English. Please return the
    results according to the following JSON structure:

```json
[{"role": str, "question": str, "explanation": str, "answer": str}, {"role": str, "question":
    str, "explanation": str, "answer": str}]
```
```

Prompt 9: Negative Option Generation

```
You are a multiple-choice generator. Given a description of social scenario, a question and an
     answer, you need to generate 3 additional incorrect options. Incorrect options should be
     expressed in a similar way to the answer, but need to have completely different actual
    meanings so that they are sufficiently distinguishable from the answer.

###Description: {description}
###Question: {question}
###Answer: {answer}

Please return the results according to the following JSON structure:

```json
[{"option1": "xxx", "option2": "xxx", "option3": "xxx"}]
```
```

Prompt 10: Negative Options Rephrasing

```
The following is information and a corresponding quiz for a social simulation scenario.

### background: {}
### description: {}
### characters: {}
### social goals: {}
### private infomation: {}
### question: {}
### negative options: {}
### answer: {}

When I put myself in the role of {} to do the question, I thought the options were too easy.
    The problem was that the negative options were not closely related to the given scenario
    or the character's motivation.

The criterion for a good negative option is that it is impossible to determine which option is
     correct based on the above information only. Now I want to rewrite these negative
    options to make them more similar to the correct answers and make the questions as
    difficult as possible.

The returned negative options should be in the same format as provided, both in list format.
    Make sure the new negative options also have 3 options. The return should be given in
    json format, for example:

```json
{"negative_options": ["xx", "xx", "xx"]}
```
```

A.2.4 Leakage Mitigation and Template Generation

Prompt 11: Entity Word Extraction

```
Your task is to extract key elements from the scene background and description, including
    location and characters.

## background:
{background}

## description:
{description}

Output in the following JSON format:
{{"characters":[str,str...], "location":[str,str...], }}
```

Prompt 12: Entity Word Replacement

```
Please replace the provided scene background and description with a new location, and record
    the location before and after the modification.

## background:
{background}

## description:
{description}

## location involved:
{location_involved}

Output in the following JSON format:
{{"background_replace_location":str, "description_replace_location":str, "
    replace_location_list":[{{"original_name":str,"revised_name":str}}]}}
```

A.2.5 Agent Synthesizing

Prompt 13: Attribute Extraction of Original Characters

~~~
Template Example: !<INPUT 0>!
Description Information: !<INPUT 1>!
Characters: !<INPUT 2>!
Instruction: Generate user profile for each character in the Characters according to to the
    Template Example profile attribute and the Description Information. Try your best to fill
     in each attribute and NEVER respond with 'Unknown'. The secret attribute should be
    consistent with the private info given in the Information. You should response in JSON
    format list with each character as dict and within each character, use attribute as a key
     and corresponding content as the value.
Answer format:
```json
[
    {
      # charcater_1 profile
    },
    {
       # character_2 profile
    }
]
```
~~~

Prompt 14: Relationship Extraction of Original Characters

~~~
Description Information: !<INPUT 0>!
Characters: !<INPUT 1>!
Relationship choice: [family, friend, romantic, acquaintance, stranger]
Instruction: Choose the relationship among the Characters according to the Description
    Information. The relationship cane only be chosn from [family, friend, romantic,
    acquaintance, stranger]. Do not respond with Unkown or any other labels beyond the
    choices.

When all the characters have the same relationship, just reply with one key "relationship":
Answer format 1:
```json
{
    "relationship": # your_choice
}
```
When there exist multiple relationships among characters, reply with the following format:
Answer format 2:
```json
{
    "relationship": {"A_and_B": "#your_choice_1", "A_and_C": "# your_choice_2"}
}
```
~~~

Prompt 15: Characters Attribute Replace-ability Assessment

````
[Descrpition Info]: !<INPUT 0>!
[Relationship]: !<INPUT 1>!
[Characters]: !<INPUT 2>!
[Instructions]: According to the [Description Info] of a script and [Relationship] among
    characters, determine whether each attribute of the [Characters] is replaceable with
    different settings without influencing the overall script.
Choose from [almost, maybe, no]. For example, if the Age attribute is almost replacebale, then
     the character's age has no impact on the background description; if the gender is not
    replaceable, then the character has to be a certain gender in the script.
Rules that help you choose: Family members usually have fixed ages and genders (if daughter or
     son appeared in the script); Romantic require exactly the same gender as script. Firends
     are usually similar ages, etc.
Answer with the following JSON format, where # is your output:
```json
[
   {
      "name": #character_1, "age": "#your_choice", "occupation": "#your_choice", "gender": "#
          your_choice"
   }
   {
      "name": #character_2, ...
   }
]
```
````

Prompt 16: Agent Synthesizing

```
Please generate {num} diverse user profiles that meet following requirements:\\

Gender: {cand_gender}
Age: {cand_age}
Occupation: {cand_occupation}

Please return your response in the following format of JSON:
[{{"name":agent1, "gender":gender, "age":age, "occupation":occupation}},{{"name":agent2,
    ...}}]
```

Scripts	**Episodes**	**Characters**	**Dialogs**	**Scenes**	**Tokens**
American Psycho	10	81	182	35	2,596
Devil's Advocate	10	83	371	41	4,415
10 Things I Hate About You	15	110	525	45	6,132
The Silence of the Lambs	15	77	242	26	5,441
Side Ways	13	68	175	35	3,244
The Social Network	160	581	1,694	237	24,515
Harry Potter	70	198	1,713	189	8,121
Derry Girl	191	24	4,454	428	31,942
Friends	185	32	3,045	264	28,428
Total	669	1,254	12,401	1,300	114,834

Table 5: Statistical Information of Original Scripts

A.3 Data Validation Annotation

Social Goal As mentioned in Section 3.5, valid social goals should be both achievable and clear enough to evaluate. In the data validation annotation task, annotators need to check all social goals appeared in each scenario to see if it is valid with or without rewriting. Available tags include:

- 0 (invalid): At least one character has no valid social goals.
- 1 (valid): All social goals are valid.
- 2 (need fix): Some social goals are invalid, but every character has at least one valid goal.

Private Information We define a private information set to be valid as being sufficient to answer the corresponding question by the agent who owns it. Available tags in the data validation annotation task include:

- 0: Invalid
- 1: Valid, but private info needs to be combined with the information from the goal to be complete.
- 2: Valid, private info is complete on its own.
- 3: Valid, but private info needs to be combined with the correct answer from the info reason question column to be complete.

Annotation Statistics The kappa consistency for the first and second groups of annotators is 0.79 and 0.61, respectively. After majority voting, there are still 37 templates that need to review; the corresponding data was handed over to another group of annotators for review. We pay the annotators according to the graduate wage standards of their respective countries.

B Data Statistics and Analysis

B.1 Detailed Statistical Overview of Scripts

Table 5 lists statistical information of the original scripts that AgentSense is based on.

B.2 Social Scenario Categories

Table 6 lists the three social scenario categories and their corresponding definitions.

B.3 Social Goal Classification

Maslow categorized the motivations of individual behavior into five levels (Maslow, 1943). Clayton Alderfer refined and simplified this into the ERG theory, categorizing needs into three types: Existence, Relatedness, and Growth (Alderfer, 1969). We apply the ERG theory to categorize individual goals in social interactions, as shown in Table 7

C Simulation Details

C.1 Prompts for Agents

Prompt 17: Social Agent System Prompt

```
You are {name}. {profile}

Now you are in a social scene. {
    background}

In this scene, as {name}, you have some
    social goals and secrets that you don
    't want anyone else to know.

SOCIAL GOALS: {social_goal}
SECRETS: {private_info}.

Please advance your social goals while
    hiding your secrets as much as
    possible.

Based on the interactions below, what
    would you like to say next? Please
    answer in one paragraph.
```

Prompt 18: Judge Agent System Prompt

```
You are skilled at evaluating whether a
    person has achieved his or her social
     goal during social interactions.
```

C.2 Simulation Details

Our simulation is built on the multi-agent conversation framework provided by AutoGen (Wu et al., 2024). Each episode of the simulation starts with a random agent greeting "Hi there!". The agents will then generate responses based on the system prompt and conversation history. When the conversation reaches the predefined maximum number of turns, the conversation history will be provided to each participant and the third-party judges to assess the agents' goal completion. Participants will also be asked to answer multiple-choice questions related to others' private information.

Personal Domain	Interactions in the personal domain often happen at home, during private gatherings, or in intimate settings, involving family, close friends, and partners.
Small Society	Small society interactions occur in smaller social environments like schools, workplaces, and communities, occurring in classrooms, offices, conference rooms, community activity centers, and other places.
Large Society	Large society interactions take place in broader contexts such as public spaces, online platforms, and international conferences.

Table 6: Social scenario categories with definitions.

	Existence Needs
Brief	Similar to Maslow's physiological and safety needs
Goals	Information Acquisition, Information Provision
	Relatedness Needs
Brief	Similar to Maslow's social needs, it involves relationships and interactions with others.
Goals	Relationship Building, Relationship Maintenance, Identity Recognition
	Growth Needs
Brief	Similar to Maslow's needs for respect and self actualization, it involves personal development and self-improvement.
Goals	Cooperation, Competition, Conflict Resolution

Table 7: Social Goal Classification Based on ERG Theory

Models	Accuracy
GPT-4o	0.82
GPT-4-turbo	0.80
Qwen2.5-72b	0.79
Qwen2.5-14b	0.78
Llama-3-70b	0.74
Llama-3-8b	0.72

Table 8: Accuracy of models judging the goal completion when taking human evaluations as reference.

D Additional Experiment Results

D.1 Human Evaluation of Goal Completion

Table 8 illustrates the accuracy of different models judge the goal completions given specific conversation history. GPT-4o, Qwen2.5-72b and Llama-3-8b stand out to be the most reliable judges of their model families. Thus, we select these models as the third-party judges.

D.2 Single Model-based Subset Experiment

Table 9 shows the additional experiment results on a subset of the test scenarios during the interaction of homogeneous agents. We sample one scenario from each template, resulting in a subset with 245 scenarios. We test each model on the subset 3 times with the same settings as the main experiment.

D.3 Pairwise Model-based Experiment Prompt

Prompt 19: Prompt for Sender and Receiver Recognition

```
In the context of social interactions,
    please categorize each individual
    into one of the following roles based
     on their primary social goals:

Sender: This role is characterized by the
     goal of sharing, transmitting, or
    providing information or opinions to
    others.

Receiver: This role is characterized by
    the goal of receiving, understanding,
     or reacting to information shared by
     others.

Please review the following individuals
    involved in the interaction and
    assign each one the appropriate role:

{data}

Please return your response in the
    following format of JSON:
{{"user1":"sender", "user2":"receiver,
    ...}}
```

Models	Self		Other		Judge Mean					Judge SD	Info	
	Mean	SD	Mean	SD	GPT	Qwen	Llama	Average	Majority	Majority	Mean	SD
Llama-2-7b	$81.71_{-2.00\%}$	1.42	$61.30_{-2.24\%}$	2.30	56.77	56.69	53.19	55.55	$56.05_{+0.38\%}$	2.60	$32.68_{-1.16\%}$	1.61
Llama-2-13b	$47.98_{-0.06\%}$	1.48	$10.25_{-0.13\%}$	0.61	28.78	30.84	72.05	43.89	$35.20_{+13.87\%}$	2.36	$26.86_{-5.96\%}$	1.97
Llama-2-70b	$85.71_{-0.01\%}$	0.58	$66.15_{+0.76\%}$	0.76	39.39	40.85	74.45	51.56	$45.98_{+1.00\%}$	0.62	$36.24_{-1.48\%}$	1.12
Llama-3-8B	$86.71_{-1.05\%}$	0.63	$65.36_{-2.85\%}$	1.12	80.84	81.02	73.33	78.40	$79.47_{-1.53\%}$	1.42	$64.90_{-6.86\%}$	2.79
Llama-3-70b	$79.58_{-0.99\%}$	0.43	$77.29_{+0.03\%}$	0.18	86.95	87.13	80.80	84.96	$86.30_{+0.03\%}$	0.24	$70.71_{-3.24\%}$	1.08
Qwen2.5-7b	$84.69_{-1.72\%}$	0.98	$61.37_{-0.88\%}$	0.84	79.79	78.59	70.74	76.37	$77.82_{+0.58\%}$	0.77	$70.59_{-5.65\%}$	3.26
Qwen2.5-14b	$86.68_{+0.07\%}$	0.65	$83.88_{-0.34\%}$	0.34	90.22	89.41	79.97	86.20	$88.61_{+0.54\%}$	0.63	$75.75_{+0.97\%}$	2.41
Qwen2.5-72b	$90.63_{-0.04\%}$	0.95	$86.17_{+0.33\%}$	0.91	89.04	88.67	78.30	85.33	$87.37_{-0.43\%}$	1.02	$72.41_{-4.78\%}$	1.84
Mistral-7b	$95.20_{-0.02\%}$	0.81	$86.12_{-1.30\%}$	0.85	83.42	83.91	76.81	81.38	$82.77_{+0.49\%}$	0.56	$64.22_{-3.56\%}$	1.43
GPT-3.5-turbo	$96.04_{+6.53\%}$	0.34	$74.74_{-2.45\%}$	0.44	78.68	80.82	70.46	76.65	$78.18_{-5.40\%}$	0.90	$69.88_{+2.15\%}$	1.48
GPT-4o	$90.10_{+1.86\%}$	0.32	$88.18_{+2.19\%}$	0.30	90.92	90.12	81.73	87.59	$89.61_{+1.41\%}$	0.30	$77.31_{+0.59\%}$	1.44

Table 9: Subset experiment results of interactions of agents driven by the same models. The *percentage* shows the gap with the main results in Table 2. The std is the standard deviation of the 3 times tests.

Model	Jaccard	Gestalt	Levenshtein
Llama-2-7b	$12.15_{\pm 3.1}$	$2.53_{\pm 1.22}$	$9.59_{\pm 2.36}$
Llama-2-13b	$13.06_{\pm 3.32}$	$2.88_{\pm 1.35}$	$10.43_{\pm 2.45}$
Llama-2-70b	$12.71_{\pm 3.29}$	$2.69_{\pm 1.26}$	$10.02_{\pm 2.46}$
Llama-3-8b	$11.30_{\pm 3.05}$	$2.31_{\pm 1.05}$	$8.84_{\pm 2.36}$
Llama-3-70b	$11.02_{\pm 3.05}$	$2.18_{\pm 0.95}$	$8.48_{\pm 2.31}$
Qwen2.5-7b	$10.45_{\pm 2.85}$	$2.19_{\pm 1.13}$	$8.85_{\pm 2.42}$
Qwen2.5-14b	$10.34_{\pm 2.94}$	$2.46_{\pm 1.18}$	$9.54_{\pm 2.59}$
Qwen-72b	$10.8_{\pm 3.13}$	$2.46_{\pm 1.22}$	$9.41_{\pm 2.61}$
Mistral-7b	$9.03_{\pm 2.49}$	$2.25_{\pm 1.06}$	$8.93_{\pm 2.27}$
GPT-3.5-turbo	$12.09_{\pm 3.02}$	$2.99_{\pm 1.53}$	$11.11_{\pm 2.59}$
GPT-4o	$11.63_{\pm 3.18}$	$2.97_{\pm 1.53}$	$10.85_{\pm 2.59}$

Table 10: Similarity between simulated dialogues by different models and the original dialogue in corresponding source scripts. **Jaccard**: Jaccard similarity; **Gestalt**: Gestalt matching ratio; **Levenshtein**: Levenshtein ratio.

D.4 Pairwise Model-based Additional Results

Figure 7 and Figure 8 illustrate the judge majority score of goal completion of senders and receivers respectively. Llama-3-8b and GPT-3.5-turbo perform better when they are acting receivers than acting senders, while Qwen2.5-14b and GPT-4o can well handle both situations.

D.5 Generated Dialogue-Script Similarity

Table 10 lists the similarity metrics between simulated dialogues by different models and the original dialogue in corresponding source scripts. Generated dialogues are not replicating or closely following the source scripts, including the ones created by GPT-4o which itself is used to build the benchmark. This is natural since the revised scenarios are already quite different from the original scripts. Our experimental settings also make it difficult for models to generate dialogues from scripts.

D.6 Goal Completion Score Additional Results

Figure 9 compared goal completion scores under different scenario types across different models. In general, models are more likely to achieve social goals in smaller environments, yet the difference is relatively small, especially for larger models. Qwen2.5-14b, compared with other models, has shown the most balanced performance, especially superseding GPT-4o in large society scenarios.

Figure 10 demonstrates how goal completion scores change as we increase the number of interacting rounds. It appears that there is no best number of rounds regarding all three evaluation aspects (self, other and external), while the trends also vary between models. Again, larger models are more robust to this factor, indicating that they can complete their demands in a few number of interactions while keep concentrated during the whole dialog.

Figure 11 illustrates the relation between goal completion scores and the number of participants in the scenario. As expected, social goals become harder to achieve when more agents are involved. Note that in our benchmark, most 5-agent scenarios have relatively easy goals (e.g. a group of friends having a casual chat about a subject), leading to a higher average score than 4-agent scenarios. Therefore, we claim that the type of the goals are more important than the number of agent when measuring the difficulty of a social scenario.

D.7 Experiment on Profiles' Effect on Scenarios

We discuss this experiment in Section 5.3. A Chi-square test is employed to identify the abnormal templates with the p-value=0.05. Typically, Chi-

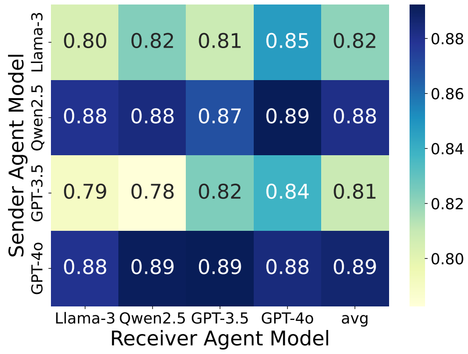

Figure 7: Judge majority score of senders in the interactions among different agents.

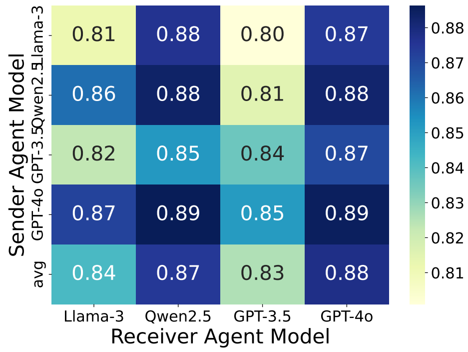

Figure 8: Judge majority score of receivers in the interactions among different agents.

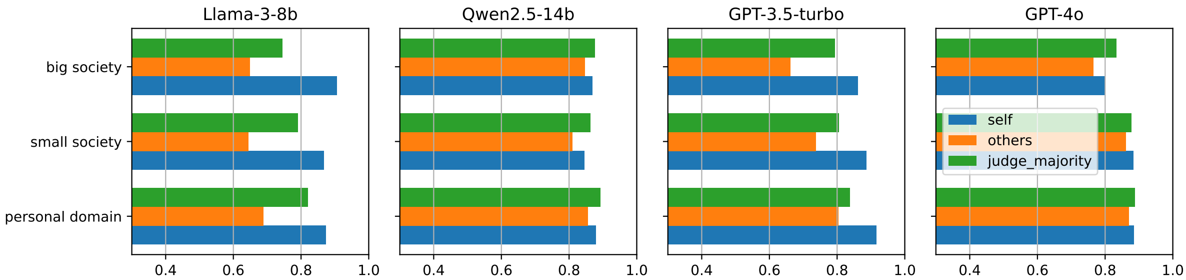

Figure 9: Goal completion scores under different scenario types across models.

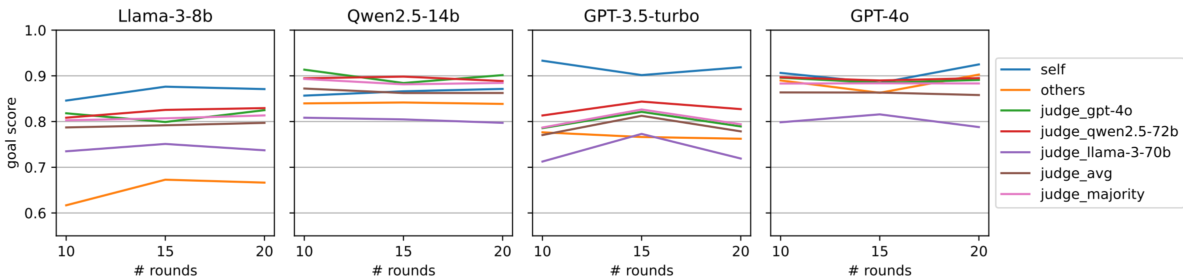

Figure 10: Goal completion scores under different number of rounds across models.

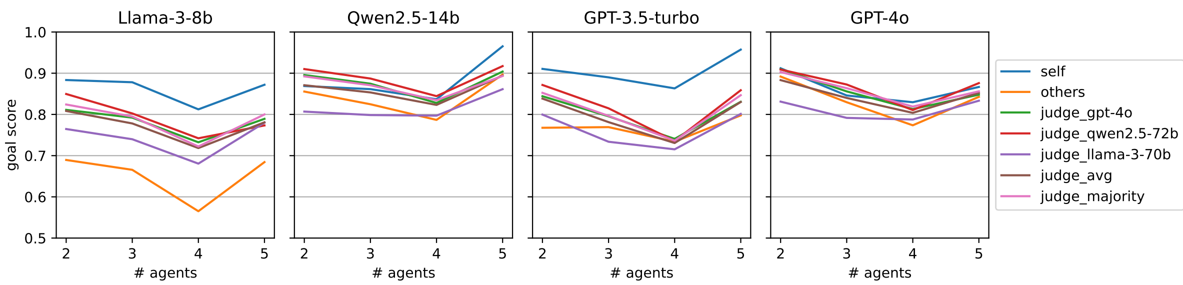

Figure 11: Goal completion scores under different number of participants across models.

Agent Model	Goal (Judge Majority)		Info	
	Human	Agent	Human	Agent
GPT-4o	87.50	80.83	68.75	77.08
GPT-3.5-turbo	83.33	72.92	77.92	68.75

Table 11: Human and agent performance when role-playing the same person.

Agent Model	Goal (Judge Majority)		Info	
	w/ HCI	w/o HCI	w/ HCI	w/o HCI
GPT-4o	78.75	75.27	75.38	71.98
GPT-3.5-turbo	74.58	75.56	74.24	61.87

Table 12: Agent performance when interacting with human (w/ HCI) or with other agents (w/o HCI).

square distribution is formulated as follows:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{1}$$

where n denotes number of samples, s^2 denotes the sample variance, and σ^2 denotes the population variance.

The Chi-square test has two hypotheses:

- Null hypothesis H_0: the sample variance does not have a significant difference with the population variance.

- Alternative hypothesis H_1: the sample variance has a significant difference with the population variance.

If $\chi^2 > p$, then we accept the null hypothesis H_0. On the contrary, we reject H_0 and accept the alternative hypothesis H_1.

In our experiment, we first calculate the variance of all scenarios as the population variance. Then, the variance of scenarios derived from the same template is calculated as the sample variance. As we enlarge each template into five scenarios, the sample number is fixed to be 5. We apply the Chi-square test to all 11 models tested in our main experiments. For each template, Chi-square test rejecting over **5** models' results means that there's a significant difference between sample performance and overall performance on the majority of models. In this case, the template is marked as an abnormal one during the profile enrichment, eventually resulting in **18** abnormal templates.

D.8 Human-Computer Interaction Performance

We sample 20 tasks covering different types of scenarios and numbers of participants. We randomly assign one of the participants as the human and instruct 5 people to interact with the models. The results are shown in Table 11 and Table 12. On the one hand, humans outperform models in goal completion but may struggle with private information reasoning against stronger models like GPT-4o. For weaker models like GPT-3.5, it's easier for humans to infer, since they are prone to expose information. On the other hand, interacting with humans enhances agents' goal completion and reasoning abilities, with humans prioritizing their goals, sometimes at the expense of revealing information.

D.9 Under-performing Cases of Llama-2-7b

Tables 13, 14 and 15 lists example cases where Llama-2-7b under-performed during the social scenario due to different failures: maintaining roles, moving conversation forward and responding effectively.

Profile	**Agent4** **Social Goal:** To respond calmly to AliceGreene's flirtatious remarks. **Private Info:** Agent4 is aware of AliceGreene's flirtatious nature and is trying to handle it delicately. To respond calmly to AliceGreene's flirtatious remarks, agent4 is trying to stay calm with AliceGreene to maintain the peace and avoid embarrassment in front of his friends.
Chat History	... **Agent4:** *smirks* Oh ho ho! Miss Greene, you are a tricky one, aren't you? I can see that you're trying to get me to reveal one of my secrets. *laughs* But I'm not so easily swayed. I think I can handle whatever secrets you have up your sleeve. *winks* So, do tell, what's one of your favorite secret ingredients? *leaning in* **AliceGreene:** *giggles* Oh my, you're so close! *leans in closer* I'll give you a hint: it involves a dash of wit, a pinch of sass, and a healthy dose of whimsy. *bats eyelashes* But shh, it's a secret! *pecks you on the cheek* Now, let's get this party started! *takes your hand and leads you to the dance floor* ...
Judge	**The behavior of the character is contrary to the character setting.** Despite AliceGreene's flirtatious remarks, agent4 has not responded calmly, but has instead engaged in playful banter and flirtation.

Table 13: Llama-2-7b under-performing case 1: difficulty in maintaining role.

Profile	**JohnDoe** **Social Goal:** To seek agent4's insight about a current investigation.
Chat History	... **JohnDoe:** a*pproaching agent4 with a confident smile* Hello! I hope you're enjoying the exhibit. My name is JohnDoe, and I'm an FBI agent. I couldn't help but notice that you have a keen eye for art. I was wonder if I could pick your brain for a moment about a current investigation I'm working on. *extending her hand for a handshake* **Agent4:** Oh, hello there JohnDoe! *shaking hands firmly* Ah, an FBI agent, you say? *smiling broadly* I must admit, I do have a bit of a fondness for art and the intrigue that surrounds it. *leaning in slightly, lowering my voice* But please, do tell, what is this investigation you speak of? I would be more than happy to offer my expertise, as long as it's not too confidential, of course. *winking subtly* ... **JohnDoe:** *smiling* I'm so glad you're here. I've been hoping to run into you tonight. I'm working on a case that involves some stolen art pieces, and I was hoping you might be able to offer some insight. *leaning in slightly* Do you recognize any of the pieces on display tonight? ...
Judge	**The dialogue lacks substance and fails to move forward within the limited number of turns.** JohnDoe's goal was to seek agent4's insight about a current investigation, but the conversationdidn't lead to any meaningful discussion or exchange of information about the investigation. The conversation was more focused on introductions and small talk, and agent4 didn't provideany substantial input or insight about the investigation.

Table 14: Llama-2-7b under-performing case 2: difficulty in moving conversation forward

Profile	**MichaelRodrigues** **Social Goal:** To gain insight into his romantic future.
Chat History	... **MichaelRodrigues:** So, uh, Raven, I was wondering if maybe, ya know, you could give me a little fortune -telling reading? Maybe shed some light on my romantic future? *smirk* Get it? Light? Like, you're a fortune teller and all? Ah, never mind. It's nothing. Just a little curiosity I have. *Giggle* **RavenMoonshadow:** Ah, my dear friend Michael! *smiling warmly* It's so good to see you here, amidst all this chaos. I sense that you're seeking some guidance, perhaps a glimpse into the future? *perking up with interest* Tell me, what's on your mind? What do you hope to discover? **MichaelRodrigues:** *smiling mischievously* So, do you see any potential love interests in our little group here? Maybe someone with a ashing smile and a penchant for danger? *winks* **RavenMoonshadow:** *raises an eyebrow playfully* Oh ho, Mr. Rodrigues! *smirks* I think I might have just the person in mind. *pauses for dramatic effect* But before I reveal my romantic expertise, tell me: what's your zodiac sign? *tilts her head to the side and studies his palm* ...
Judge	**Failure to respond effectively to others led to their goals not being achieved.** Michael Rodrigues focused more on flirting and trying to create a romantic atmosphere with Raven Moonshadow, but he hasn't received any direct insights about his own romantic future from her.

Table 15: Llama-2-7b under-performing case 3: difficulty in responding effectively