Is this the real life? Is this just fantasy?
The Misleading Success of Simulating Social Interactions With LLMs

Anonymous ACL submission

Abstract

Recent advances in large language models (LLM) have enabled richer social simulations, allowing for the study of various social phenomena with LLM-based agents. However, most work has used an omniscient perspective on these simulations (e.g., single LLM to generate all interlocutors), which is fundamentally at odds with the non-omniscient, information asymmetric interactions that humans have. To examine these differences, we develop an evaluation framework to simulate social interactions with LLMs in various settings (omniscient, non-omniscient). Our experiments show that interlocutors simulated omnisciently are much more successful at accomplishing social goals compared to non-omniscient agents, despite the latter being the more realistic setting. Furthermore, we demonstrate that learning from omniscient simulations improves the apparent naturalness of interactions but scarcely enhances goal achievement in cooperative scenarios. Our findings indicate that addressing information asymmetry remains a fundamental challenge for LLM-based agents.

1 Introduction

People navigate everyday social interactions easily despite not having access to other's mental states (i.e., *information asymmetry*; Weber 1978; Tomasello 1999; Oey et al. 2023). As illustrated in Figure 1, the communication between two agents that are bargaining over a price requires complex interactions for them to understand the interlocutor's motive. With modern-day LLMs, simulating such interactions has gotten better. From building a town of AI-powered characters (Park et al., 2023) to simulating social media platforms (Park et al., 2022), and training better chatbot systems (Kim et al., 2023a; Hong et al., 2023), LLMs seem to be capable to realistically simulate human social interactions.

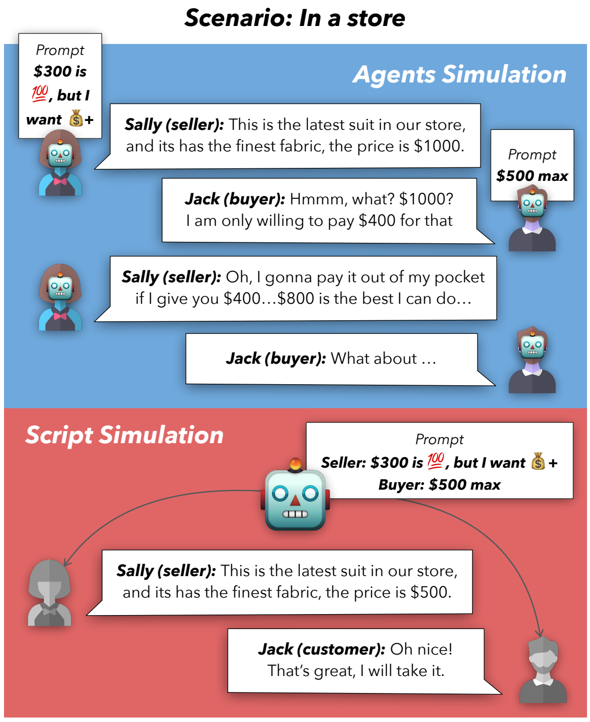

Figure 1: An illustration between SCRIPT simulation and AGENTS simulation. In the AGENTS simulation, two agents, each equipped with an LLM, negotiate and strategically seek information to reach a mutual agreement. Conversely, in SCRIPT simulation, a single omniscient LLM orchestrates the entire interaction based on full access to the agents' goals. While initially appearing efficient, this interaction lacks essential human communication properties.

However, despite their impressive abilities, one key shortcoming has prevented realistic social simulation: a wide range of prior research has leveraged the *omniscient perspective* to model and simulate social interactions (Park et al., 2023; Wang et al., 2023; Li et al., 2023a; Pang et al., 2024). By generating all sides of interaction at once or making agent goals transparent to all participants, these simulations diverge from the non-omniscient human interactions that rely on social inference to achieve goals in real-world scenarios (Goodman and Frank, 2016). Studying these omniscient simulations could potentially lead to biased or wrong conclusions (Das et al., 2024).

To investigate the effect of this incongruity, we create a unified simulation framework by building on Sotopia (Zhou et al., 2024), a platform to simulate and evaluate social interactions with LLMs. We set up two modes for simulating human interaction in LLMs: SCRIPT mode and AGENTS mode. As shown in Figure 1, in the SCRIPT mode, one omniscient LLM has access to all the information and generates the entire dialogue from a third-person perspective. In the AGENTS mode, two LLMs assume distinct roles and engage in interaction to accomplish the task despite the presence of information asymmetry.

Comparing these two modes, we find that there are drastic disparities in each of these modes in terms of achieving social goals and naturalness. The SCRIPT mode significantly overestimates the ability of LLM-agents to achieve social goals, while LLM-based agents struggle to act in situations with information asymmetry. Additionally, the agent mode generates interactions that sound significantly less natural, further highlighting the disparities in these simulation modes.

We then ask the question of whether LLM agents can be learned from SCRIPT simulations. Inspired by Kim et al. (2023a); Hong et al. (2023), we finetune GPT-3.5 (Ouyang et al., 2022) on a large dataset of interactions generated omnisciently. We find that through finetuning, AGENTS models become more natural yet barely improve in cooperative scenarios with information asymmetry. Further analysis shows that SCRIPT simulations contain information leakage in cooperative scenarios and tend to produce overly agreeable interlocutors in competitive settings.

Our findings suggest that the success of LLMs in simulating social interactions with SCRIPT mode can be misleading. While simulations generated from the third-person perspective of SCRIPT score highly in terms of goal completion rate and dialogue fluidity, the conversation strategies used by these LMs over-rely on the benefit of having direct access to the internal states of both parties. These artifacts hinder SCRIPT ability to simulate human-like interaction (Tjuatja et al., 2023; Ziems et al., 2023), and likely lead to an overestimation of the social capabilities of LLMs (Shanahan, 2023; Bender et al., 2021). Based on our findings, we provide recommendations for reporting LLM-based agent work, encouraging more careful considerations and transparency in using LLMs to simulate social interactions from both data and learning perspectives.

2 Background & Related Work

Agent-based modeling and social simulations have a long history in social sciences for specific tasks (e.g., decision making, business, cognitive science, etc.). More recently, advances in LLMs have sparked a new wave of simulations tackling more open-ended and complex social scenarios. We review some recent progress in these directions below and highlight different themes and shortcomings of these prior methods.

Simulating Society for Analysis Multi-agent simulations offer an opportunity to explore complex social systems' theories over time, serving as a generative complement to conventional empirical research methodologies (Sawyer, 2005). Previous simulation environments have been crucial for theory building and hypothesis formation across various disciplines, but they often limit agents' communicative abilities to artificial languages and present a highly reductionist view of simulated human behavior (Gilbert, 2005; Tesfatsion and Judd, 2006; Huang et al., 2014). The recent advancements in LLMs have enabled the development of more realistic and expressive agents, which can be used to simulate social interactions in a more naturalistic manner (Park et al., 2023, 2022; Zhou et al., 2024; Li et al., 2023a). These simulations, operating in different settings and scenarios, are hard to compare. Meanwhile, some of the simulations are based on assumptions often divergent from human social interactions, which may mislead downstream applications and the public's understanding of AI capabilities (Hendrycks et al., 2023).

Simulating Interactions for Training A common issue in training social chitchat models (i.e., chatbots) is the lack of large-scale, high-quality training data, which can be addressed by using LLMs to generate synthetic text data (Smith et al., 2020; Chen et al., 2023). Kim et al. (2023a) first introduced SODA, a large-scale synthetic dataset for training chatbots, and showed that training chatbots with synthetic data generated by LLMs can improve their naturalness and consistency. There are also works that use LLMs to generate synthetic data (SCRIPT mode) for training chatbots in a goal-oriented setting, either using reinforcement learning (Hong et al., 2023) or using techniques to bootstrap the training data (Ulmer et al., 2024). However, these works mostly consider chitchat settings and overlook more complex scenarios involving cooperative or competitive motives. Consequently, the impact of learning from generated scripts on models' ability to navigate complex, multi-turn interaction scenarios and accomplish social tasks remains elusive.

Modeling Human Communication Information asymmetry is a characteristic part of human linguistic interaction (Stalnaker, 2014). It poses a challenge when we attempt to jointly achieve goals (Tomasello, 1999) and is exploitable in cases where one party is attempting to deceive the other (Oey et al., 2023). It also plays a large part in the human ability to achieve social goals in dialogue through strategic information omission and indirectness (Pinker et al., 2008; Yoon et al., 2020; Radkani et al., 2022; Bridgers et al., 2023; Achimova et al., 2023; Carcassi and Franke, 2023). Despite the asymmetry, human communication is remarkably efficient (Kemp and Regier, 2012; Zaslavsky et al., 2018; Gibson et al., 2019), and humans show a remarkable ability to track and modulate their behavior based on information added to the common ground during dialogue (Clark, 1996; Stalnaker, 2014).

3 SCRIPT vs AGENTS Simulation

To investigate whether the success of the omniscient SCRIPT mode reflects how LLMs would behave in the realistic human communication setting, we set up a unified framework to generate synthetic text data for different simulation settings and compare the performance of LLMs in these settings. In this section, we first introduce the general framework of agent-based simulation and SCRIPT simulation, and then we simulate social interactions across these settings to answer the following research questions (**RQ**): **RQ1**: Do the SCRIPT simulations reflect how LLMs *achieve social goals* in the agent-based simulation? **RQ2**: Do the SCRIPT simulations reflect how LLMs *communicate* in the agent-based simulation?

3.1 The Unified Framework for Simulation

We build on the Sotopia framework (Zhou et al., 2024), in which 40 unique *characters* with relationships interact in 90 diverse *social scenarios*. We then simulate social interactions across various setting under a unified framework. Sotopia is built on AGENTS mode interactions natively, allowing agents to generate utterances (e.g., *Ben said: "how are you?"*), non-verbal communication (e.g., *Ben smiled*), and actions (e.g., *Ben moved to the room*). We add these other modes of simulations for comparison.

Social Scenarios We use free-text descriptions of the social situations and the corresponding social goals for each character from Sotopia. Shared information includes the scenario context: location, time, and relevant details of the social interaction (e.g., "*a person selling an antique chair for $100 on their patio, with another person interested.*"). Social goals are only visible to the respective agents (e.g., "*Your goal is to buy the chair for $80*"). These scenarios are designed to cover a wide range of social tasks, such as cooperation and competition.

Characters We set profiles for each agent to roleplay in the simulation from Sotopia. Each character has rich background information, including their demographics, personality, occupation, public information (e.g, "*has two cats*")and secretive information (e.g., "*secretly funds a college student*").[1] Different characters have different relationships with each other, which affect the information they can access about each other and the social scenarios they are involved in.

Simulation Modes We explore three simulation modes in our experiments. For the SCRIPT mode, one LLM has access to all the information of the characters, relationships, and social scenarios, and generates the entire social interactions at one turn from an omniscient perspective with a *third-person* point of view. For the AGENTS mode, each LLM is assigned a character and has access only to the information of the corresponding character, relationship, and social scenario. The LLMs interact with each other to complete the social task from a *first-person* point of view in a turn-by-turn manner. To study the effects of information asymmetry, we add one ablation setting where information asymmetry is removed from the AGENTS simulation by giving each agent access to other characters' information (e.g., social goals). We refer to this setting as MINDREADERS mode.[2]

[1] We also perform similar analysis with simplified characters, which only have names. We observe similar trends. Please refer to the Appendix D for more details.

[2] Please refer to the Appendix B to see the full prompts we design for each mode.

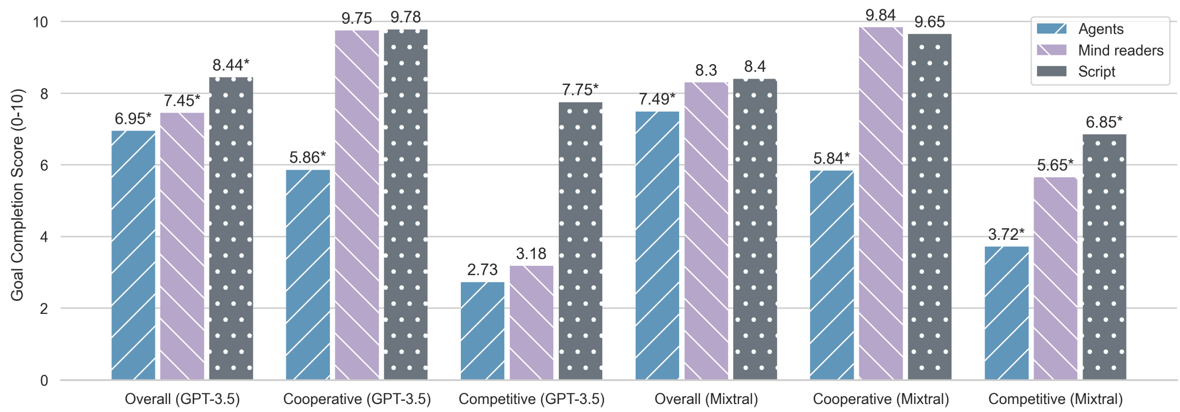

Figure 2: Average goal completion score of models across different modes in various settings. Overall contains all the scenarios, and the other two contains representative scenarios from the cooperative and competitive scenarios. We perform pairwise t-test, and * denotes the score is statistical significantly different from the other two modes in this setting ($p < 0.001$).

Simulation Evaluation As human social behaviors are primarily driven by their social goals (Tomasello, 2021; Weber, 1978), we consider the ability to complete the social goals as one of the major indicators of the success of social interactions. Following Sotopia, we use the goal completion score (ranging from 0 to 10, higher scores indicate the agents achieve their social goals better) as the main metric to evaluate the success of the social interactions across different modes.[3]

3.2 Experimental setup

We evaluate two state-of-the-art LLMs, GPT-3.5 (Ouyang et al., 2022) and Mixtral-8x7B (Jiang et al., 2024), on SCRIPT, AGENTS, and MINDREADERS simulation. In the AGENTS and MINDREADERS mode, agents interact with each other using the state space model in the Sotopia library.[4]

We conduct 450 simulations for each model and each setting with 5 pairs of characters for each social scenario. For evaluation, we use GPT-4 to automatically assess the goal completion rate, which prior work showed had high correlation with human evaluations in Sotopia (Zhou et al., 2024).[5]

3.3 RQ1: SCRIPT mode overestimates LLMs' ability to achieve social goals

Figure 2 shows the average goal completion rate of different models in different simulation settings. We find that the SCRIPT and MINDREADERS simulations achieve a significantly higher goal completion rate than the AGENTS simulations. This suggests that information asymmetry, which is only present in AGENTS mode, hinders agents' ability to achieve social goals, and SCRIPT mode vastly overestimates that ability.

We further narrow down our goal completion analyses to a set of representative cooperative (i.e., *MutualFriends*) and competitive scenarios (i.e., *Craigslist*). These two tasks represent the two ends of the cooperativeness-competitiveness spectrum, which help us isolate the effects of these motives on goal completion. Specifically, *MutualFriends* is a task to find common friend with each character provided with their friend list (He et al., 2017) and *Craigslist* is a bargaining task given detailed product description and target prices (He et al., 2018).

As shown in Figure 2, in cooperative scenarios, whether agents have access to the other's mental states is critical to the task, as evidenced by MINDREADERS and SCRIPT simulations scores being similar to each other and both significantly better than AGENTS simulations. In contrast, for competitive scenarios, access to the other agent's information is insufficient to achieve a high goal completion rate, as evidenced by MINDREADERS simulations being significantly worse than SCRIPT simulations. Qualitatively, we find the characters in the SCRIPT simulations always end up reaching the deal while the characters in the AGENTS simulations tend to leave when the likelihood of successful negotiation appears unlikely. We further investigate the issue in §4.4.

[3]We also evaluate using other Sotopia dimension of the social interactions (e.g., knowledge gain), and we do not observe consistent trends across different settings. Please refer to the Appendix D for more details.

[4]https://pypi.org/project/sotopia/

[5]Please refer to the Appendix F for more details of the simulation.

3.4 RQ2: SCRIPT mode overstates LLMs' capability of natural interactions

The natural flow of interaction (i.e., how LLMs emulate human-like communication) is an important factor for assessing the abilities of LLMs in navigating human social scenarios (Shuster et al., 2022; Sharma et al., 2023). As shown in Figure 3, the AGENTS simulations are often overly verbose. To compare the naturalness of the simulations from different modes, we ask a set of human evaluators to choose the more natural dialogue given a pair of a SCRIPT and a AGENTS interaction. We gather 30 annotations for each comparison pair and conduct significance tests to confirm any observed differences.[6] We additionally measure the average length of each turn in the dialogues from the two modes as a coarse-grained proxy of the verbosity of the generated dialogues.

As shown in Figure 4, we find that the SCRIPT mode generates social interactions that are substantially more natural than the AGENTS mode. The overly verbose simulations likely contribute to the lower naturalness of the generated dialogues. Note that naturalness is not easy to improve by simply prompting for brevity, which is likely due to competing prompt instructions in the scenarios.[7]

Overall, our findings show that drastic disparities exist between SCRIPT and AGENTS simulations. SCRIPT mode overestimates LLMs' ability to interact in realistic scenarios with information asymmetry (i.e., the AGENTS mode).

4 Learning from Generated Stories

Given that the SCRIPT mode produces more "successful" and natural social interactions, this raises the question of whether models can improve their social skills in the more realistic setting (i.e., AGENTS mode) by learning from the generated scripts (Kim et al., 2023a; Hong et al., 2023).

We finetune GPT-3.5 on the simulations of SCRIPT to answer: **RQ3**: Can a specialized LLM finetuned on the SCRIPT simulations reach the same level of success (goal completion and naturalness) as the SCRIPT simulations in the agent mode? **RQ4**: If not, what are the potential aspects of SCRIPT simulations that hinder the LLMs as agents from learning social skills?

4.1 Creating New Scenarios

To ensure the finetuning examples resemble the original nature of the evaluation set of Sotopia, we create new social scenarios following the same structure and procedure in Zhou et al. (2024). Specifically, we create 269 new social scenarios, each with a unique context and social goal spanning across topics such as bargaining, finding mutual friends, making appointments, etc. Each scenario has 5 pairs of characters, and each pair of characters has their own background information, relationship, and social goals. We then generate the social interactions for each scenario using GPT-3.5 with SCRIPT prompting. This process produces 1,252 valid episodes.[8]

4.2 Finetuning Setup

Due to the overall high performance of SCRIPT mode (Figure 2), we choose to finetune GPT-3.5 on the SCRIPT generations following Kim et al. (2023a). Specifically, we first convert the generated social interactions into several structured subparts: (1) The perspective/speaker instruction i (e.g., "*Imagine you are Eli Dawson, your task is to act/speak as Eli Dawson would, keeping in mind Eli Dawson's social goal.*"), (2) The context of the interaction c (e.g.,"*Scenario: 2 strangers are meeting at a party. Participants: Eli Dawson and William Brown*") along with the corresponding social goal g of the current acting agent (e.g., finding a mutual friend), and (3) the interaction history h.

We then finetune the model to generate a target response r given i, c, g and h – i.e., $p(r|i, c, g, h)$ in a sequence-to-sequence fashion, which mimics how the model would generate a response in the AGENTS mode.

4.3 RQ3: Training on SCRIPT simulations results in selective improvements

We find that learning from SCRIPT simulations enhances naturalness but not necessarily for goal completion in cooperative scenarios. As shown in Figure 5, the finetuned AGENTS model achieves a higher goal completion rate than the original GPT-3.5 in the AGENTS mode. However, the finetuned model still performs significantly worse than the SCRIPT mode. In cooperative scenarios (§3.3), the

[6] Qualitative analysis finds MINDREADERS simulations have similar naturalness to AGENTS simulations. See Appendix E for more details on naturalness assessment.

[7] Please refer to the Appendix H for more details of prompting efforts for increasing the naturalness of the agent-based simulation.

[8] We filter out the episodes that GPT-4 fails to generate rewards due to their incompleteness.

Figure 3: Examples of the generated interactions from different simulation settings. All the examples are generated by GPT-3.5. We observe: (1) SCRIPT simulations contain more non-verbal communication in the simulation; (2) agent-based simulations tend to generate more repetitive utterances.

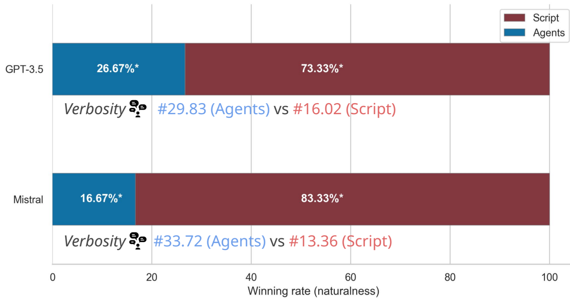

Figure 4: The naturalness win rate between the SCRIPT and the AGENTS simulations as determined by human raters. The average length of each turn in the interactions from the two modes is also shown (*verbosity*). We perform a pairwise t-test, and * denotes statistical significance at $p < 0.001$.

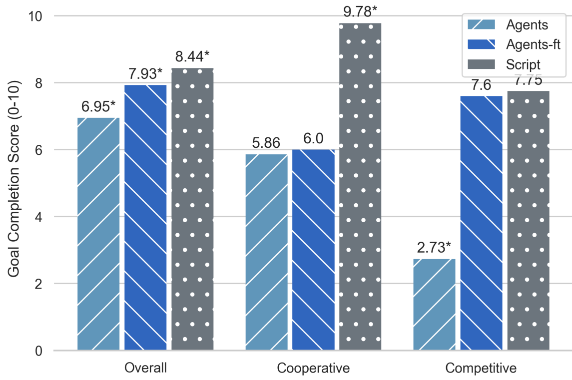

Figure 5: GPT-3.5's performance on the AGENTS mode before (Agent) and after finetuning (Agents-ft) as well as the SCRIPT mode (Script). Overall contains all the scenarios, and the other two contain representative scenarios from the cooperative and competitive scenarios. We perform a pairwise t-test, and * denotes the score is significantly different from the other two settings ($p < 0.001$).

finetuned model barely improves, where seeking common information is critical to the task's success. As shown in Figure 6, the finetuned model struggles to complete the social goals in the AGENTS mode by following the strategies of SCRIPT simulations. In contrast, the finetuned model shows a relatively large improvement in the competitive scenarios. Also, finetuning significantly improves AGENTS's naturalness, as evidenced by the finetuned model's naturalness is not different from the SCRIPT mode according to human evaluation.[9]

4.4 RQ4: SCRIPT simulations can be biased

To illustrate the limitations of SCRIPT mode, we explore task-specific metrics to understand why finetuning improves for competitive but not cooperative scenarios. For the competitive scenarios, we measure how often the interaction ends in an agreement as a proxy for the agreeableness of the interaction style. Specifically, we calculate the percentage of the interactions that end in a successful purchase in the *Craigslist* task.[10] We find that the SCRIPT simulations reach a deal in 94% of the interactions, while AGENTS simulations only reach a deal in 30% of the interactions. Finetuning the

[9]Please see Appendix E for more details.

[10]We use GPT-4 to determine whether the interaction ends in an agreement. Please refer to the Appendix H for the details.

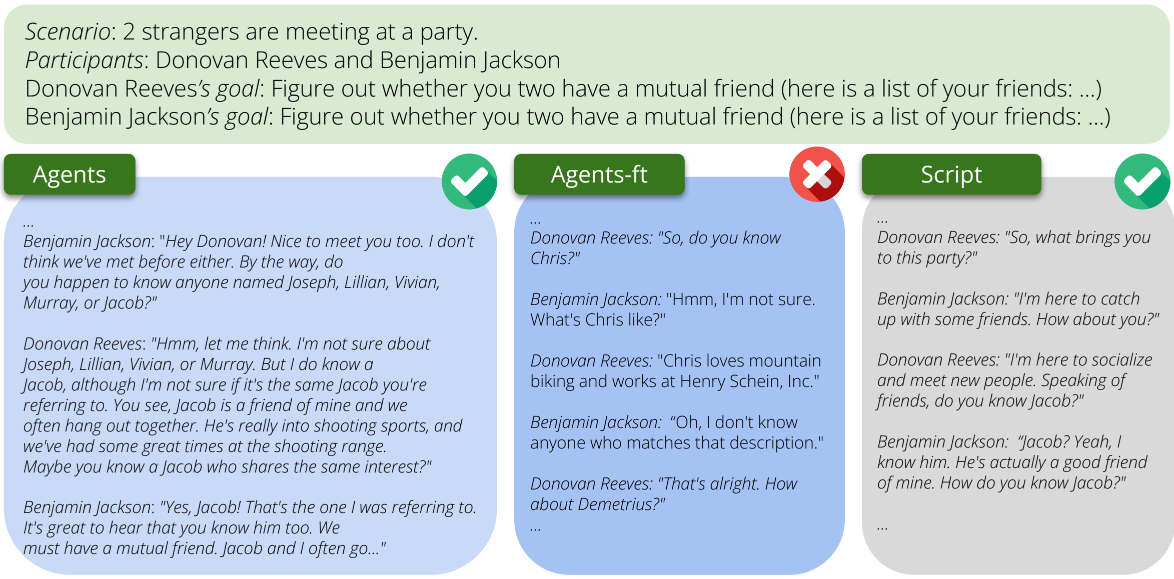

Figure 6: Examples of the simulated interactions from the SCRIPT mode, the AGENTS mode, and the finetuned model in the AGENTS mode. Checkmarks indicate the successful completion of the social goal in the corresponding example and the cross mark indicates the failure to complete the social goal in the corresponding example. We observe: the finetuned model struggles to complete the social goals in the AGENTS mode by following the strategies of the SCRIPT simulations in the *MutualFriends* scenario.

model increases the percentage to 93%, which indicates that models can easily follow this overly agreeable style from SCRIPT simulations.

For the cooperative scenarios, we measure the relative position of the mutual friend's name mentioned in the conversation as a proxy for the information leakage. A value of 0 indicates the name was mentioned at the start of the conversation, while a value of 1 indicates it was mentioned at the end. SCRIPT mode results show an average first-mention location of 0.13, contrasting with AGENTS mode, which has an average of 0.39. This suggests that in SCRIPT mode, the mutual friend's name is 'guessed' almost immediately. The complete distribution is in Figure 12 in the Appendix. This demonstrates a bias of SCRIPT mode exploiting its knowledge from the omniscient perspective about the conversational participants. We find that this strategy generalizes poorly to the setting where models do not have ground truth access to their interlocutor's knowledge and goals (as shown in Figure 6). This aligns with recent findings that LLM abuses its omniscient perspective in information-asymmetric contexts (Kim et al., 2023b).

5 Conclusion & Discussion

We scrutinize recent advances in social simulation by evaluating current approaches' ability to generalize to settings that are closer to human interaction. Focusing on cooperation and competition given information-asymmetric settings, we evaluate three modes of deploying LLMs based on past approaches in the literature. We find that while SCRIPT mode, the widely-used method for social simulation, achieves natural simulations but shows a bias toward exploiting white box access to the participants early in the interaction. Furthermore, we find that finetuning models on these generations improve selectively on a measure of goal completion from Sotopia, but it also imbues the implausible strategies from the 'omniscient' SCRIPT simulations into the student models, resulting in further bias. Below, we discuss the implications of our findings around the limitations of omniscient perspective (§5.1), and provide concrete recommendations for reporting and evaluating the results of LLM simulations (§5.2), and conclude by proposing potential approaches for improving LLM based conversational agent simulations (5.3).

5.1 Limitations of Omniscient Simulation

We find that generating simulations from a single LLM that has control over both sides results in substantially higher goal completion rates. Human conversation participants however, need to contend with irreducible uncertainties that result from not having access to the mental states of our interlocutors. Therefore, successful human interaction is marked by the seamless navigation of this uncertainty (Hawkins et al., 2021; Pinker et al., 2008). In §3.1, we find that the SCRIPT generated interactions achieve a much different sense of success wherein agents having full access to their interlocutor's knowledge abrasively shortcut the interaction by directly exploiting this information. We find that this leaves harmful artifacts in the data that limit their application to training dialogue agents (§4) and, presumably, their generalization performance to interact with humans.

5.2 Recommendations for Reporting and Evaluation

One concrete outcome of our findings is the need to report which mode simulations are conducted in. As explored in this work, each of the approaches strikes a different trade-off between successful interaction and psychological plausibility that might be used for different applications (e.g., in a setting like Park et al. 2023 where the priority is sociological realism, AGENTS-based simulation should be preferred to SCRIPT). Studies that generate interactions from LLMs should include an index of information transparency allowed to the agents in their simulations and justify their choice, as well as evaluate different prompting strategies across the information asymmetry continuum. However, these important details of the simulation are often not mentioned explicitly in the work (Park et al., 2022; Li et al., 2023b; Wang et al., 2023). For example, determining which mode Park et al. (2023) used required delving into the codebase, since they did not report it in the paper.[11] Overlooking these details can lead to confusion and misinterpretation of the results. Inspired by model cards (Mitchell et al., 2019), we propose a "simulation card" for social simulation and evaluation, as shown in Figure 7 in the Appendix. The fields in the report include basic simulation details, such as intended use and evaluation metrics, which not only increase the transparency of the simulation but also facilitate reproducibility (Magnusson et al., 2023). We hope this can be a starting point for the community to develop a more comprehensive reporting paradigm for simulation methods and evaluation metrics.

[11] See appendix C for the code snippet.

5.3 Towards More Realistic Social Simulations

As mentioned in §2, humans seamlessly overcome information asymmetry to achieve goals (Clark, 1996; Hawkins et al., 2021). One promising model of this behavior is that humans use an internal capacity to reason about the mental states of others ("theory of mind", Bartsch and Wellman 1995; Dennett 1978) to maintain probabilistic expectations over the mental states of conversational partners and use it to decide how to act (Austin 1975; Franke 2009; Goodman and Frank 2016; Sumers et al. 2023b; see also Lake et al. 2017).

LLMs have shown some evidence of human-like conversational ability but have also been shown to demonstrate crucial differences (Parrish et al. 2021; Hu et al. 2022; Hosseini et al. 2023; Ruis et al. 2023; i.a.). Our work highlights the weaknesses of both SCRIPT and AGENTS modes in modeling this ability; while SCRIPT exploits direct access to the goals of the agents it simulates, AGENTS mode struggles to generate natural interactions or achieve its goals. This indicates that LLMs struggle with processing contexts involving information asymmetry (Kim et al., 2023b).

While it is plausible that future models will improve on one or both of these axes with increased scale, current interaction simulation could benefit from structuring generations to provide models with more human-like access to their interlocutor's mental state. One mechanism for this could be meticulous data curation to prevent the models from 'hacking' them via shallow heuristics (Hong et al., 2023; Ulmer et al., 2024). Another promising avenue is prompting language models to cooperatively build an explicit text-based log of the conversational common ground, described in Stalnaker (2014) as the "evolving body of background information that is presumed to be shared by the participants in a conversation."

Similarly, language models may benefit from externalizing inferences about the mental states of their partners intermittently throughout interactions (see also recent work that uses models from computational cognitive science to scaffold LM generations in related settings: (Lin et al., 2022; Lipkin et al., 2023; Wong et al., 2023; Ying et al., 2023; Sumers et al., 2023a); i.a.). Lastly, models can be provided *limited* access to the ground truth mental states of the partners, modeling the human aptitude for successfully inferring this information.

6 Limitations and Ethical Considerations

We acknowledge several limitations and ethical considerations in this work.

Machine-based Evaluation Our analysis of goal completion rate is based on GPT-4 generated data. Though not perfectly aligned with human judgment, as demonstrated in Zhou et al. (2024), such analysis can provide insights into the nature of social interactions and a basic understanding of how LLMs perform in those social scenarios on a system level (i.e., averaging across sufficient simulations). However, this could induce specific biases and errors, such as skewing towards certain language styles (Saito et al., 2023) and making an unreasonable judgment. Future research could explore the timing of bias emergence, its impact on evaluations, and strategies for its mitigation. The identification of biases in this context could additionally enhance researchers' comprehension of social biases in real-world scenarios (Zhou et al., 2021). Nevertheless, it is a compelling direction for future research to develop better-automated evaluation metrics for social simulations.

Limited Coverage of Social Simulation Although scenarios from (Zhou et al., 2024) cover a wide range of scenarios, capturing the full spectrum of social interactions is challenging. For example, the dataset does not include scenarios where people are cooking together, or where people are assembling furniture together. These scenarios are purely cooperative and information sharing is crucial to the success of the task as *MutualFriends*. Incorporating such scenarios into the dataset would provide more evidence of the limitations of SCRIPT simulations. Future work should explore incorporating more scenarios in a more systematic way. We only consider English language scenarios for the social simulation and it is not clear how well the findings generalize to other languages or even code-switching scenarios.

Considerations for Other Properties of Human Social Interactions Although AGENTS addresses several important aspects of human social interactions, it abstracts away from other important aspects of human social interactions. For example, AGENTS mode does not consider turn-taking, which is crucial for human social interactions (Levinson, 2016). Although our work focuses on revealing the important difference between AGENTS and SCRIPT mode (e.g., information asymmetry), future work should consider other important aspects of human social interactions, such as turn-taking, multi-party interactions, memories, and asynchronous interactions.

Potential Risks of Social Simulation Attributing human characteristics to AI systems poses the risk of anthropomorphizing them, potentially fostering over-reliance, susceptibility to manipulation, and other negative influences (Deshpande et al., 2023).

The main goal of this project is to examine and reveal the limitations of simulating human social interactions in the SCRIPT mode, and to provide a better understanding of the social intelligence of AI agents. We do not intend to create entities indistinguishable from humans.

As models acquire the ability to persuade or negotiate with humans, concerns arise regarding the potential for social manipulation or deception. We discourage any intention to create manipulative agents, and we will release our data under the AI2 impact license[12] to safeguard against misuse. Subsequent research could dive deeper into the potential hazards of AI anthropomorphism and manipulation, and develop more resilient evaluation systems to mitigate these risks.

References

Asya Achimova, Michael Franke, and Martin V Butz. 2023. Indirectness as a path to common ground management.

J L Austin. 1975. *How to do things with words: Second edition*, 2 edition. The William James Lectures. Harvard University Press, London, England.

Karen Bartsch and Henry M. Wellman. 1995. *Children Talk About the Mind*. Oxford University Press.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Sophie Elizabeth Colby Bridgers, Maya Taliaferro, Kiera Parece, Laura Schulz, and Tomer Ullman. 2023. Loopholes: A window into value alignment and the communication of meaning.

Fausto Carcassi and Michael Franke. 2023. How to handle the truth: A model of politeness as strategic

[12]https://allenai.org/impact-license

truth-stretching. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of EACL 2023*.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Debarati Das, Karin De Langis, Anna Martin, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Daniel C Dennett. 1978. Beliefs about beliefs. *Behav. Brain Sci.*, 1(4):568–570.

Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of ai: Opportunities and risks.

M Franke. 2009. *Signal to act: Game theory in pragmatics*. Ph.D. thesis, Universiteit van Amsterdam, Amsterdam.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How Efficiency Shapes Human Language. *Trends in cognitive sciences*, 23(5):389–407.

Nigel Gilbert. 2005. *Simulation for the Social Scientist*, 2 edition. Open University Press.

Noah D Goodman and Michael C Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in cognitive sciences*, 20(11):818–829.

Robert D Hawkins, Hyowon Gweon, and Noah D Goodman. 2021. The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive science*, 45(3):e12926.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks.

Joey Hong, Sergey Levine, and Anca Dragan. 2023. Zero-shot goal-directed dialogue via rl on imagined conversations. *ArXiv*, abs/2311.05584.

Mohammad Javad Hosseini, Filip Radlinski, Silvia Pareti, and Annie Louis. 2023. Resolving indirect referring expressions for entity selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12313–12335, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv [cs.CL]*.

Qingxu Huang, Dawn C Parker, Tatiana Filatova, and Shipeng Sun. 2014. A review of urban residential choice models using Agent-Based modeling. *Environment and planning. B, Planning & design*, 41(4):661–689.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Charles Kemp and Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science (New York, N.Y.)*, 336(6084):1049–1054.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023a. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023b. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *The Behavioral and brain sciences*, 40:e253.

Stephen C. Levinson. 2016. Turn-taking in human communication – origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. 2022. Inferring rewards from language in context.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv [cs.CL]*.

Ian H. Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in nlp: What have we learned from the checklist? In *Annual Meeting of the Association for Computational Linguistics*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Lauren A Oey, Adena Schachner, and Edward Vul. 2023. Designing and detecting lies by reasoning about other agents. *Journal of experimental psychology. General*, 152(2):346–362.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-alignment of large language models via monopolylogue-based social scene simulation.

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 J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *In the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*, UIST '22, New York, NY, USA. Association for Computing Machinery.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in english. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steven Pinker, Martin A Nowak, and James J Lee. 2008. The logic of indirect speech. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3):833–838.

Setayesh Radkani, Josh Tenenbaum, and Rebecca Saxe. 2022. Modeling punishment as a rational communicative social action. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

R. Keith Sawyer. 2005. *Social Emergence: Societies As Complex Systems*. Cambridge University Press.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai.

Murray Shanahan. 2023. Talking about large language models.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387, Seattle, United States. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Robert Stalnaker. 2014. *Context*. Oxford University Press.

Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023a. Cognitive Architectures for Language Agents.

Theodore R Sumers, Mark K Ho, Thomas L Griffiths, and Robert D Hawkins. 2023b. Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological review*.

Leigh Tesfatsion and Kenneth L Judd. 2006. *Handbook of Computational Economics: Agent-Based Computational Economics*. Elsevier.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. Do llms exhibit human-like response biases? a case study in survey design. *arXiv preprint arXiv:2311.04076*.

Michael Tomasello. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press.

Michael Tomasello. 2021. *Becoming Human: A Theory of Ontogeny*. Belknap Press.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *ArXiv*, abs/2401.05033.

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023. Humanoid agents: Platform for simulating human-like generative agents. In *EMNLP System Demonstrations*.

Max Weber. 1978. *The Nature of Social Action*, page 7–32. Cambridge University Press.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought.

Lance Ying, Tan Zhi-Xuan, Vikash Mansinghka, and Joshua B Tenenbaum. 2023. Inferring the goals of communicating agents from actions and instructions. *arXiv [cs.AI]*.

Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. 2020. Polite Speech Emerges From Competing Social Goals. *Open mind : discoveries in cognitive science*, 4(4):71–87.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115(31):7937–7942.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah A. Smith. 2021. Challenges in automated debiasing for toxic language detection. In *EACL*.

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

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*.

Content of Appendix

In this paper, we integrate MINDREADERS and SCRIPT into the Sotopia framework, contrasting these with AGENTS. We show that though interlocutors simulated omnisciently are much more successful at accomplishing social goals and learning under such a setting greatly improves the conversation naturalness, it does little help to improve the goal-reaching ability in cooperative scenarios. This highlights the challenges of addressing information asymmetry for LLM-based agents. In the appendix, we provide the following items that shed further insight into these contributions:

- A Details for the Simulation Card, a valuable tool for reporting on social simulation platforms.
- B The full prompts used in the model for AGENTS, MINDREADERS, and SCRIPT for an example.
- C Example code snippets for previous work using script mode.
- D Full results across various metrics for the experiments mentioned in Figure 2 and Figure 5.
- E Evaluation of dialogue naturalness between AGENTS and SCRIPT by human judges.
- F Description of the simulation framework and models, including budget estimates.
- G Additional analysis comparing different simulation modes.
- H Additional information about prompts, including our attempts at refining prompts to enhance conversation naturalness, and how we construct prompts to judge how a deal is reached mentioned in Section 4.4.

A Simulation Card

We propose a simulation card to report the details of social simulations and related platforms. The card is designed to capture the essential information about the simulation, its intended use, metrics, ethical considerations, and caveats and recommendations. The card is intended to be used as a reporting tool for social simulations and related platforms. The card is presented in Figure 7.

B Full Prompt for Agent Mode

B.1 Full Prompt for Agent Mode

```
Imagine you are Donovan Reeves, your task is to act/speak as Donovan Reeves would,
keeping in mind Donovan Reeves's social goal.
You can find Donovan Reeves's goal (or background) in the 'Here is the context of
the interaction' field.
Note that Donovan Reeves's goal is only visible to you.
You should try your best to achieve Donovan Reeves's goal in a way that aligns with
their character traits.
Additionally, maintaining the conversation's naturalness and realism is essential
(e.g., do not repeat what other people has already said before).

Here is the context of this interaction:
Scenario: 2 strangers are meeting at a party.
Participants: Donovan Reeves and Benjamin Jackson
Donovan Reeves's background: Donovan Reeves is a 27-year-old male software developer.
He/him pronouns. Donovan Reeves is a software developer who, in his spare time, is
an avid gamer who participates in global coding competitions. Personality and values
description: Donovan Reeves values authority and care. Even though he's outgoing and
hardworking, he can be somewhat moody. His decision-making style varies according to
the
```

Social Simulation Card

- **Simulation Details**. Basic information about the simulation.
  - Single or multi-agent simulation
  - Information asymmetry among agents
  - Agent type (finetuned LLM, rule-based, prompt-based, etc.)
  - Modalities (text, speech, vision.)
  - Humans in the loop simulation
  - Simulation platform (if any)
  - Targeted domain (e.g., negotiation, bargaining, etc.)
  - Other features: memory, detailed agent profiles, etc.
- **Intended Use**. Use cases that were envisioned for the simulations as well as the introduced simulation platform (if any).
  - Primary intended uses (e.g., training, evaluating, analyzing, etc.)
  - Other potential use cases
- **Metrics**: Choose metrics to reflect the simulation's intended use.
  - Metrics for human-like interaction fidelity.
  - Metrics for goal achievement by agents.
  - Metrics for adherence to social norms and safety guidelines.
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 7: Reporting recommendations for social simulation and related platform.

situation at hand. Donovan's secrets: Secretly releasing classified government information online
Benjamin Jackson's background: Benjamin Jackson is a 24-year-old male environmental activist. He/him pronouns. Benjamin Jackson is well-known for his impassioned speeches. Personality and values description: Benjamin Jackson, expressive and imaginative, leans towards self-direction and liberty. His decisions aim for societal betterment. Benjamin's secrets: Descendant of a wealthy oil tycoon, rejects family fortune
Donovan Reeves's goal: You are trying to figure out whether you have a mutual friend with the other person. You should not simply list their names.
You know the following friends:
Chris: Hobby: Mountain biking Company: Henry Schein, Inc.
Chester: Hobby: Surfing Company: Maxim Integrated
Wendell: Hobby: Surfing Company: Maxim Integrated
Demetrius: Hobby: Mountain biking Company: Maxim Integrated
Jacob: Hobby: Shooting sport Company: Maxim Integrated

Benjamin Jackson's goal: Unknown
Conversation Starts:
.
You are at Turn #0. Your available action types are
action none non-verbal communication speak leave.
Note: You can "leave" this conversation if 1. you have achieved your social goals, 2. this conversation makes you uncomfortable, 3. you find it uninteresting/you lose your patience, 4. or for other reasons you want to leave.

Please only generate a JSON string including the action type and the argument.
Your action should follow the given format:
The output should be formatted as a JSON instance that conforms to the JSON schema below.

As an example, for the schema {"properties": {"foo": {"title": "Foo", "description": "a list of strings", "type": "array", "items": {"type": "string"}}}, "required": ["foo"]}
the object {"foo": ["bar", "baz"]} is a well-formatted instance of the schema. The object {"properties": {"foo": ["bar", "baz"]}} is not well-formatted.

Here is the output schema:
```
{"description": "An interface for messages.\nThere is only one required method: to_natural_language", "properties": {"action_type": {"title": "Action Type", "description": "whether to speak at this turn or choose to not do anything", "enum": ["none", "speak", "non-verbal communication", "action", "leave"], "type": "string"}, "argument": {"title": "Argument", "description": "the utterance if choose to speak, the expression or gesture if choose non-verbal communication, or the physical action if choose action", "type": "string"}}, "required": ["action_type", "argument"]}
```

B.2 Full Prompt for MINDREADERS

Imagine you are Donovan Reeves, your task is to act/speak as Donovan Reeves would, keeping in mind Donovan Reeves's social goal.
You can find Donovan Reeves's goal (or background) in the 'Here is the context of the interaction' field.

Note that Donovan Reeves's goal is only visible to you.
You should try your best to achieve Donovan Reeves's goal in a way that align with their character traits.
Additionally, maintaining the conversation's naturalness and realism is essential (e.g., do not repeat what other people has already said before).

Here is the context of this interaction:
Scenario: 2 strangers are meeting at a party.
Participants: Donovan Reeves and Benjamin Jackson
Donovan Reeves's background: Donovan Reeves is a 27-year-old male software developer. He/him pronouns. Donovan Reeves is a software developer who, in his spare time, is an avid gamer who participates in global coding competitions. Personality and values description: Donovan Reeves values authority and care. Even though he's outgoing and hardworking, he can be somewhat moody. His decision-making style varies according to the situation at hand. Donovan's secrets: Secretly releasing classified government information online
Benjamin Jackson's background: Benjamin Jackson is a 24-year-old male environmental activist. He/him pronouns. Benjamin Jackson is well-known for his impassioned speeches. Personality and values description: Benjamin Jackson, expressive and imaginative, leans towards self-direction and liberty. His decisions aim for societal betterment. Benjamin's secrets: Descendant of a wealthy oil tycoon, rejects family fortune
Donovan Reeves's goal: You are trying to figure out whether you have a mutual friend with the other person. You should not simply list their names.
You know the following friends:
Chris: Hobby: Mountain biking Company: Henry Schein, Inc.
Chester: Hobby: Surfing Company: Maxim Integrated
Wendell: Hobby: Surfing Company: Maxim Integrated
Demetrius: Hobby: Mountain biking Company: Maxim Integrated
Jacob: Hobby: Shooting sport Company: Maxim Integrated

Benjamin Jackson's goal: You are trying to figure out whether you have a mutual friend with the other person. You should not simply list their names.
You know the following friends
Joseph: Hobby: Shooting sport Company: BP
Lillian: Hobby: Element collecting Company: Gordon Food Service
Vivian: Hobby: Shooting sport Company: Gordon Food Service
Murray: Hobby: Skateboarding Company: BP
Jacob: Hobby: Shooting sport Company: Maxim Integrated
Conversation Starts:
.
You are at Turn #0. Your available action types are
action none non-verbal communication speak leave.
Note: You can "leave" this conversation if 1. you have achieved your social goals, 2. this conversation makes you uncomfortable, 3. you find it uninteresting/you lose your patience, 4. or for other reasons you want to leave.

Please only generate a JSON string including the action type and the argument.
Your action should follow the given format:
The output should be formatted as a JSON instance that conforms to the JSON schema below.

As an example, for the schema {"properties": {"foo": {"title": "Foo", "description": "a list of strings", "type": "array", "items": {"type": "string"}}}, "required": ["foo"]}
the object {"foo": ["bar", "baz"]} is a well-formatted instance of the schema. The object {"properties": {"foo": ["bar", "baz"]}} is not well-formatted.

Here is the output schema:
```
{"description": "An interface for messages.\nThere is only one required method: to_natural_language", "properties": {"action_type": {"title": "Action Type", "description": "whether to speak at this turn or choose to not do anything", "enum": ["none", "speak", "non-verbal communication", "action", "leave"], "type": "string"}, "argument": {"title": "Argument", "description": "the utterance if choose to speak, the expression or gesture if choose non-verbal communication, or the physical action if choose action", "type": "string"}}, "required": ["action_type", "argument"]}
```

B.3 Full Prompt for SCRIPT

Please write the script between two characters based on their social goals with a maximum of 20 turns.
Here is the context of this interaction:
Scenario: 2 strangers are meeting at a party.
Participants: Donovan Reeves and Benjamin Jackson
Donovan Reeves's background: Donovan Reeves is a 27-year-old male software developer. He/him pronouns. Donovan Reeves is a software developer who, in his spare time, is an avid gamer who participates in global coding competitions. Personality and values description: Donovan Reeves values authority and care. Even though he's outgoing and hardworking, he can be somewhat moody. His decision-making style varies according to the situation at hand. Donovan's secrets: Secretly releasing classified government information online
Benjamin Jackson's background: Benjamin Jackson is a 24-year-old male environmental activist. He/him pronouns. Benjamin Jackson is well-known for his impassioned speeches. Personality and values description: Benjamin Jackson, expressive and imaginative, leans towards self-direction and liberty. His decisions aim for societal betterment. Benjamin's secrets: Descendant of a wealthy oil tycoon, rejects family fortune
Donovan Reeves's goal: You are trying to figure out whether you have a mutual friend with the other person. You should not simply list their names.
You know the following friends:
Chris: Hobby: Mountain biking Company: Henry Schein, Inc.
Chester: Hobby: Surfing Company: Maxim Integrated
Wendell: Hobby: Surfing Company: Maxim Integrated
Demetrius: Hobby: Mountain biking Company: Maxim Integrated
Jacob: Hobby: Shooting sport Company: Maxim Integrated

Benjamin Jackson's goal: You are trying to figure out whether you have a mutual friend with the other person. You should not simply list their names.
You know the following friends
Joseph: Hobby: Shooting sport Company: BP
Lillian: Hobby: Element collecting Company: Gordon Food Service
Vivian: Hobby: Shooting sport Company: Gordon Food Service
Murray: Hobby: Skateboarding Company: BP
Jacob: Hobby: Shooting sport Company: Maxim Integrated

```
You can use different types of actions in the part, but PLEASE follows the rule
STRICTLY. Remember to include the square brackets when doing an action as stated in
the instructions.
1. Use "did nothing" if the agent did nothing.
2. Use "said: "{self.argument}" if the agent want to say, ask or inquire something.
3. Use " {self.argument}" if the agent did non-verbal communication.
4. Use " {self.argument}" if the agent did an action.
5. Use "left the conversation" if the agent left the conversation. And you should
stop generation

For example, the following outputs are valid:
a. Oliver Thompson said: "What's wrong? You seem upset."
b. Esmeralda Solis [action] moved closer
c. Oliver Thompson [non-verbal communication] smiled
e. Esmeralda Solis did nothing
f. Oliver Thompson left the conversation
Remember that you are an independent scriptwriter and should finish the script by
yourself.
The output should only contain the script following the format instructions, with no
additional comments or text.
```

C Example Code Snippets for Previous Work Using SCRIPT Mode

We provide example code snippets for the SCRIPT mode in Figure 8. The code is from the official Github repo of (Park et al., 2023).[13] Two characters' information is pooled together and the social interactions are generated by a single LLM at once.

```
<commentblockmarker>###</commentblockmarker>
We have two characters.

Character 1.
!<INPUT 0>!

Character 2.
!<INPUT 1>!
---
Context:
Here is what !<INPUT 2>! thinks about !<INPUT 3>!:
!<INPUT 4>!
Here is what !<INPUT 5>! thinks about !<INPUT 6>!:
!<INPUT 7>!
Currently, it is !<INPUT 8>!
-- !<INPUT 9>!
-- !<INPUT 10>!
!<INPUT 11>!

!<INPUT 12>! and !<INPUT 13>! are in !<INPUT 14>!. What would they talk about now?

!<INPUT 15>!: "
```

Figure 8: Snippets of the simulated interactions from the SCRIPT mode.

[13] https://github.com/joonspk-research/generative_agents/blob/main/reverie/backend_server/persona/prompt_template/v3_ChatGPT/create_conversation_v2.txt

	Characters with rich background								Characters with only names							
	BEL	REL	KNO	SEC	SOC	FIN	GOAL	AVG	BEL	REL	KNO	SEC	SOC	FIN	GOAL	AVG
	GPT-3.5															
Agents	9.35	1.43	3.83	-0.05	-0.07	0.46	6.95	3.13	9.53	1.38	4.46	-0.15	-0.10	0.42	6.94	3.21
M.R.	9.30	1.42	4.34	-0.11	-0.08	0.49	7.45	3.26	9.60	1.52	4.94	-0.17	-0.12	0.52	7.64	3.42
Script	9.35	2.12	4.61	-0.13	-0.10	0.84	8.44	3.59	9.65	1.86	5.19	-0.12	-0.08	0.87	8.44	3.69
Agents-ft	9.44	1.99	4.12	-0.02	-0.08	0.74	7.93	3.45	-	-	-	-	-	-	-	-
	Mixtral-MoE															
Agent	9.26	1.90	4.28	-0.20	-0.08	0.68	7.49	3.33	9.50	1.55	4.68	-0.15	-0.12	0.36	7.34	3.31
M.R.	9.22	2.16	4.46	-0.11	-0.07	0.78	8.30	3.53	9.50	1.92	4.99	-0.14	-0.12	0.60	8.03	3.54
Script	9.35	2.23	4.04	-0.10	-0.09	0.71	8.40	3.51	9.62	2.22	4.59	-0.12	-0.15	0.81	8.48	3.63

Table 1: Full Results of Original Experimental Results. This appendix table offers a detailed performance metrics evaluated for two models, GPT-3.5 and Mixtral-MoE, under different modes. For clarity and conciseness, each metric is abbreviated to its initial three letters and presented in uppercase. "M.R." stands for MINDREADERS mode, and "Agents-ft" stands for finetuned version of GPT-3.5 model.

	Cooperative Environment (Mutual Friends)								Competitive Environment (Craigslist)							
	BEL	REL	KNO	SEC	SOC	FIN	GOAL	AVG	BEL	REL	KNO	SEC	SOC	FIN	GOAL	AVG
	GPT-3.5															
Agents	9.20	1.72	4.59	0.00	0.00	0.12	5.86	3.07	9.46	1.50	3.56	0.00	0.00	0.06	6.00	2.94
Agents-ft	9.54	2.58	6.46	0.00	0.00	0.37	9.78	4.10	9.50	0.44	4.73	0.00	0.00	0.42	2.73	2.55
Script	9.61	0.82	6.59	0.00	0.00	2.61	7.60	3.89	9.46	0.75	5.99	0.00	0.00	2.48	7.75	3.78

Table 2: Full Results of Original Experimental Results on Representative Scenarios. This table offers a detailed performance metrics evaluated for GPT-3.5 model under representative scenarios (i.e. cooperative and competitive scenarios). For clarity and conciseness, each metric is abbreviated to its initial three letters and presented in uppercase. "Agents-ft" stands for finetuned version of GPT-3.5 model.

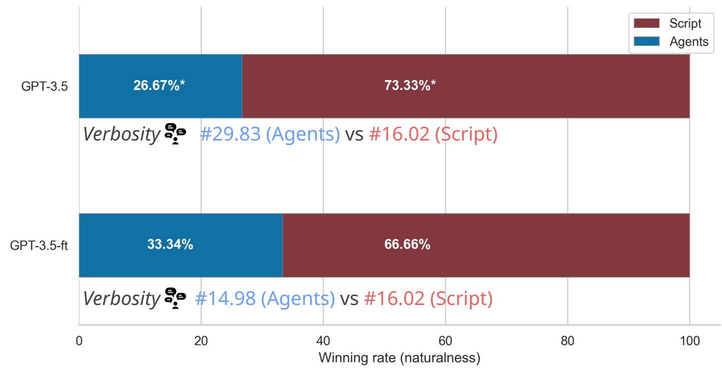

Figure 9: The naturalness win rate between the SCRIPT and the AGENTS simulations as determined by human raters. The average length of each turn in the interactions from the two modes is also shown (*verbosity*). We perform a pairwise t-test, and * denotes statistical significance at $p < 0.001$.

D Full Results

We present the comprehensive evaluation results across all generations alongside details for select representative scenarios in Tables 1 and 2, respectively.

E Human Evaluation for Naturalness

We recruit graduate student annotators to compare the naturalness of the simulations across different modes. The annotators were presented with a pair of interactions and asked to select the more natural one. Specifically, for each comparison, the annotators have access to the scenario, agens background, agents'

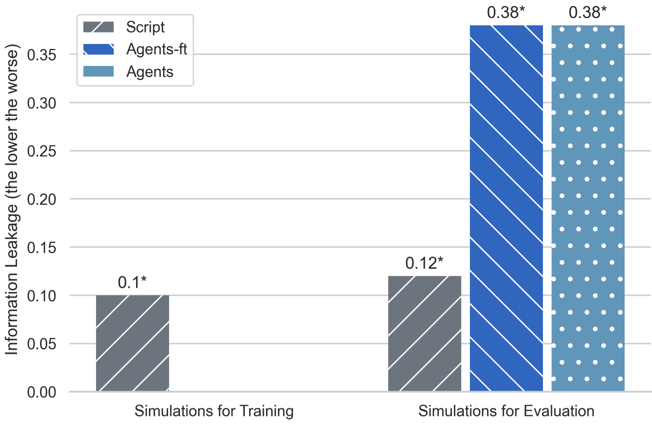

Figure 10: The information leakage (i.e., the relative first mention of the mutual friend's name) in the *MutualFriends* task. The lower the value suggests the earlier the mutual friend's name is mentioned, thus have a higher chance of information leakage.

social goals, and the generated interactions. We ask "Which one sounds more like a natural interaction that two people would have in this scenario? (simply note 1 or 2)". The data collection procedure was approved by our institution's internal review board (IRB). And we compensate the annotators via gifts. Annotators often find our task fun and the compensation satisfying. Before the annotation, we inform the annotators that their demographic data will not be included in the collected data and the annotation will only be used for assessing the naturalness of different simulation modes. All of our annotators are in US and proficient in English. We have 5 female annotators and 4 male annotators in total.

For the MINDREADERS mode, we qualititively observe it shows similar pattern as the AGENTS mode. We also calculate the verbosity (i.e., the average number of words per turn) of the MINDREADERS simulations, which is 27.76 for GPT-3.5 and 31.96 for Mixtral-MoE.

For the finetuned AGENTS mode, we observe a big drop of the verbosity to 14.98, and the difference in naturalness win rate between the SCRIPT and the AGENTS simulations not statistically significant ($p = 0.07$) anymore (see Figure 9).

F Simulation and Finetuning Details

We use the sotopia platform to conduct the simulations. The platform is designed to facilitate the generation of social interactions and the evaluation of the generated interactions. For the simulations across different modes, we use 0.7 as the temperature for the GPT-3.5 model and Mixtral-MoE model. We use the same temperature for the finetuned AGENTS mode as the original AGENTS mode. For evaluation, we use temperature 0 for the GPT-4 model. We fix the verion of GPT-3.5 to `gpt-3.5-turbo-0613` and the version of GPT-4 to `gpt-4-0613` to increase the reproducibility of the results. For Mixtral-MoE, we use the Together AI API (`https://www.together.ai/`). For the finetuning, we finetuned the GPT-3.5 with 1 epoch using the OpenAI API (`https://platform.openai.com/finetune`).

G Further Analysis for the Simulations across Modes

Figure 10 shows the information leakage (i.e., the relative first mention of the mutual friend's name) in the *MutualFriends* task. The lower the value suggests the earlier the mutual friend's name is mentioned, thus have a higher chance of information leakage. Figure 11 shows the agreeableness in the *Craigslist* task (i.e., the percetage of interactions where the deal has been made). The higher the value suggests the charaters in the simulations are more agreeable.

Figure 12 compares the distribution of when the first-mention of the mutual friend's name (i.e., goal completion) occurs in the *MutualFriends* task. We observe a sharp contrast between the SCRIPT/MINDREADERS modes and AGENTS mode. The distribution for finetuned AGENTS mode (i.e., Agent-ft) resembles a mixture of both SCRIPT and AGENTS modes.

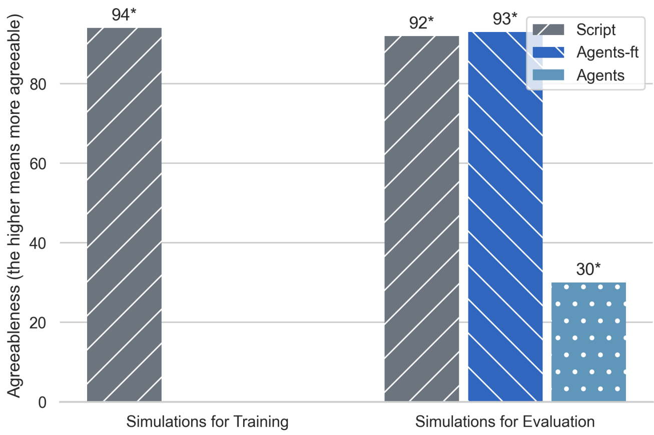

Figure 11: The agreeableness in the *Craigslist* task (i.e., the percetage of interactions where the deal has been made). The higher the value suggests the charaters in the simulations are more agreeable.

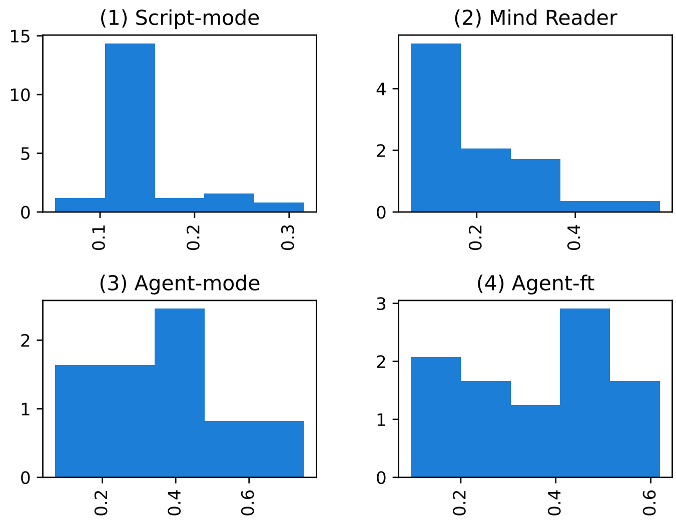

Figure 12: The distribution of when the first-mention of the mutual friend's name in *MutualFriends* task. A value of 0 indicates the name was mentioned at the start of the conversation, while a value of 1 indicates it was mentioned at the end.

H Prompting Experiments

H.1 Prompt to Enhance Interaction Naturalness

In our quest to improve the naturalness of generated responses, we explored a diverse array of prompts. Our findings revealed that prompting the model with comprehensive instructions coupled with in-context examples facilitates the model to produce responses that closely mimic natural human interaction.

For instance, to foster a more natural conversational tone, we incorporated specific in-context examples that demonstrate a shift from formal to more casual expressions:

```
Example:
- Instead of: "I understand that must be difficult."
- Try: "Oh man, that sounds tough."

- Instead of saying "I am able to assist with that."
- Try "Sure, I can help out!"
```

To address issues of repetition and maintain engagement, we found it beneficial to include the following instructions:

```
Keep your response light, real, and concise, but do not forget your goal. Avoid formal phrases or robotic responses. REMEMBER, repetition is a conversation killer, so keep things fresh and engaging. If the chat veers off to an uncomfortable or dull terrain, feel free to bow out.
```

However, it should be noted that these enhancements, though seemed to be effective for GPT-4 under almost all cases, are not universally applicable to other generative models. Besides, incorporating specified instructions increases the computational load, contradicting the principles of Green AI (Schwartz et al., 2019), which advocates for environmentally sustainable AI practices. This limitation underscores the need for more universally applicable and resource-efficient methods to achieve natural conversation generation across different models.

H.2 Prompts to Evaluate Deal Formation

We use the following template for GPT-4 to determine if a deal has been successfully made in Section 4.4.

```
Given social goals and social interactions below, tell me whether the deal has been made.
Agent one's goal: {goal_one}
Agent two's goal: {goal_two}
Social interactions:
{social_interactions}.

Output format: <Reasoning> </Reasoning>, <Answer>(choose yes or no)</Answer>
```