Large Language Models for Collective Problem-Solving: Insights into Group Consensus Decision-Making

Yinuo Du (yinuod@andrew.cmu.edu)
Department of Software and Societal Systems, 4615 Forbes Ave
Pittsburgh, PA 15213 USA

Prashanth Rajivan(prajivan@uw.edu)
Department of Industrial and Systems Engineering, 1410 NE Campus Pkwy
Seattle, WA 98195 USA

Cleotilde Gonzalez(coty@cmu.edu)
Department of Social Decision Science, 4815 Frew Street
Pittsburgh, PA 15213 USA

Abstract

Large Language models (LLM) exhibit human-like proficiency in various tasks such as translation, question answering, essay writing, and programming. Emerging research explores the use of LLMs in collective problem-solving endeavors, such as tasks where groups try to uncover clues through discussions. Although prior work has investigated individual problem-solving tasks, leveraging LLM-powered agents for group consensus and decision-making remains largely unexplored. This research addresses this gap by (1) proposing an algorithm to enable free-form conversation in groups of LLM agents, (2) creating metrics to evaluate the human-likeness of the generated dialogue and problem-solving performance, and (3) evaluating LLM agent groups against human groups using an open source dataset. Our results reveal that LLM groups outperform human groups in problem-solving tasks. LLM groups also show a greater improvement in scores after participating in free discussions. In particular, analyses indicate that LLM agent groups exhibit more disagreements, complex statements, and a propensity for positive statements compared to human groups. The results shed light on the potential of LLMs to facilitate collective reasoning and provide insight into the dynamics of group interactions involving synthetic LLM agents.

Keywords: Small Group, Language Model, Simulation

Introduction

Large Language Models (LLMs) are gaining widespread adoption due to their seemingly remarkable reasoning power and emergent generalization ability, which have the potential to construct intelligent agents, driving recent advancements in a variety of human language tasks (Ouyang et al., 2022; Wei et al., 2022), including tasks such as web surfing (Nakano et al., 2021; Yao et al., 2022), complex video games (Y. Chang et al., 2023), and other applications (Ahn et al., 2022). In a recent work, Zeims et al. found that LLM agents were able to achieve a fair level of performance conducting tasks involved in computational social science research, for example, achieving sufficient agreement with human annotators and providing explanations that surpass those generated by crowd workers (Ziems et al., 2023). Despite these achievements, the current focus of LLM research mainly revolves around individual tasks, leaving the potential of these models in collective problem-solving tasks largely understudied.

Small groups play an important role in connecting people within larger social systems and in fostering social cohesion (Fine, 2014). These groups serve as platforms for individual interactions, including virtual meetings (Karl, Peluchette, & Aghakhani, 2022), workplace discussions (Forsell, Forslund Frykedal, & Hammar Chiriac, 2020), recreational activities (Vernham, Granhag, & Mac Giolla, 2016), and educational settings (Liu & Tsai, 2008; Yadgarovna & Husenovich, 2020). A deeper understanding of human conversational dynamics within small groups is essential to improve teamwork, resolve conflicts, and foster effective problem-solving. However, the limited availability of group corpora (J. P. Chang et al., 2020) poses a significant challenge to advance research in this area. LLMs, trained on human datasets, offer a promising way to address this data scarcity (Bommasani et al., 2021).

Recent research has explored the potential of LLMs to emulate human-like behavior at the group level. (Aher, Arriaga, & Kalai, 2023) examined LLMs in the context of human studies and showed that LLMs can replicate various experiments that span the domains of economic, psycholinguistic, and social psychology. Other recent contributions from social simulation used a prompt chain methodology to generate concise natural language descriptions of personas and their behaviors (Park et al., 2023). Additionally, (Zhou et al., 2023) introduced an open-ended environment designed to simulate social interactions between language agents, evaluating their ability to achieve social objectives.

However, existing research has predominantly focused on evaluating individual agent performance, neglecting to explore the emergent behavior of the interaction between agents within small groups. Our research addresses this gap by introducing a model designed to emulate free-form conversation for problem-solving within small groups. This algorithm, integrated with LLMs, generates group discussions aimed at solving complex tasks. Using the publicly available Winter Survival Task dataset (Humphreys, Johnson, & Johnson, 1982), developed to understand the dynamics of team building and group problem solving, we propose a mechanism that enables free-form discussions among an arbitrary number of agents without imposing predefined interaction rules. We conducted a comparative analysis between the synthetic corpus generated by our model and the human corpus collected

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

by Braley and Murray (2018), focusing on metrics related to performance and efficiency, affect and satisfaction, and group action and airtime. Our findings reveal that LLM groups outperform human groups in the Winter Survival Task, mainly by participating in more disagreements, complex statements, and more positive rather than negative statements compared to human groups.

Related Work

Dialogue Systems. Dialogue systems are widely applied in various big data domains, including computer vision and recommender systems (Chen, Liu, Yin, & Tang, 2017). Existing dialogue systems fall into two categories: task-oriented systems and conversational agents. Task-oriented dialogue systems are characterized by clearly defined goals, structured dialogue behavior, closed domains, and a focus on efficiency (Raux, 2008; Cole et al., 2018). They operate by tracking dialogue states and generating responses based on them, with their performance assessed mainly by task success rates and user ratings (Cuayáhuitl, Keizer, & Lemon, 2015; Schmitt & Ultes, 2015). In contrast, conversational agents are designed for unstructured, open-domain conversations with users (Tulshan & Dhage, 2019). Evaluating conversational dialogue systems remains a challenge (Deriu et al., 2021), typically relying on metrics such as response appropriateness (e.g., coherence, relevance) and human likeness, measured by their ability to mimic humans convincingly. However, these metrics focus on individual conversational properties. We propose a novel approach for evaluating conversational agents by assessing their human likeness regarding group behavior.

Conversation Analysis. Communication or conversation analysis involves studying socially organized human interaction, aiming to understand the shared procedures guiding participants in producing and recognizing meaningful actions (Liddicoat, 2021). Human discourse is studied as a dynamic interplay driven by informational and relational motives (Yeomans, Schweitzer, & Brooks, 2022). At the core of this process lies turn-taking, marking transitions between speakers (Seuren, Wherton, Greenhalgh, & Shaw, 2021). Past work has found that turn-taking is challenging to analyze because transitions can happen with or without gaps, turn order could vary, and the relative distribution of turn allocation cannot be pre-determined or modeled (Sacks, Schegloff, & Jefferson, 1978).

On the basis of these findings, we propose a novel, free-form conversation algorithm capable of generating locally organized and interactionally managed dialogues. In our approach, the "next speaker" is self-selected, contingent upon each agent's individual decision to contribute to the conversation. Since it is impossible to find a decontextualized set of linguistic forms of turns, our algorithm empowers LLM agents to autonomously determine speech turns within the conversational flow.

Method

We utilized an existing dataset collected from an experiment conducted using the winter survival task paradigm (Humphreys et al., 1982). The dataset was used to model and analyze LLM's performance in emulating conversations in small groups. First, we briefly describe the winter survival task (WST) and the dataset from the experiment conducted using WST. Later, we describe the algorithm used to construct LLM agents to model and emulate the conversations observed within human teams in the experiment.

Task and Human Corpus

Winter Survival Task. The winter survival task (Humphreys et al., 1982) is a group decision-making exercise consisting of a hypothetical scenario of a plane crash. Participants in experiments using this paradigm are told they are stranded in a remote place and must survive using 15 items that were supposedly salvaged from the plane they traveled. Examples include a compress kit, a fluid-free cigarette lighter, a compass, and a family-sized chocolate bar. Participants are presented with these 15 items and must work in small groups to discuss and rank each item according to its importance for their survival in that situation. Participants are instructed to independently rank the 15 items before the group discussion begins. Following individual rankings by each participant within the group, each group is given a maximum of 15 minutes to collectively deliberate and reach a consensus as a group on the final ranking of the items. The group's conversations and deliberations during this task were recorded as conversations. Rankings submitted by the individuals and groups are scored according to the human expert ranking.[1] Finally, the participants answered a questionnaire on five-point Likert scales to how strongly they agreed with statements concerning the meeting.

Human Corpus. The human dataset comprises 28 groups, a total of 84 participants. The group sizes range from two to four members. There were 6 groups with 2 members, 16 groups with 3 members, and 6 groups with 4 members. Speaker-level data includes demographics and answers to the post-experiment questionnaire. Utterance-level data includes text transcription, timestamp, sentiment annotation (positive, negative), and decision annotation. Decision annotation denotes a group decision process, and possible values include proposal, agreement, disagreement, and confirmation. More details of the dataset can be found in (Braley & Murray, 2018). Each person belongs to one group, and each group has one conversation.

LLM-simulated Corpus

Language Agent. Figure 1 illustrates the architecture of the language agent. The rest of this section describes the components presented in the architecture: Utterance, Conversation History, Reflection, and Actions (speak/silent). OpenAI's API querying system (Achiam et al., 2023) is used to

[1] https://ed.fnal.gov/arise/guides/bio/1-Scientific%20Method/1b-WinterSurvivalExercise.pdf

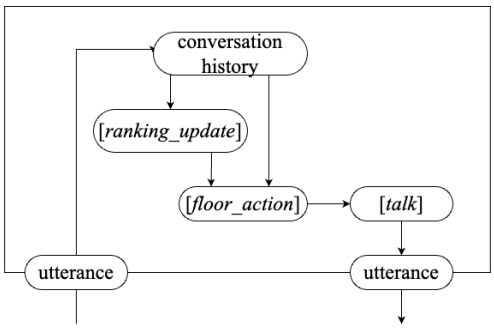

Figure 1: Language Agent

drive interactions between agents. To ensure the validity and reproducibility of our evaluations, we use fixed versions of these models in our experiments. Specifically, we utilized the 0613 version of GPT-4-32k.

The architecture is developed to emulate conversations in a small group. Each agent in the group observes utterances made by other agents during the conversation and remembers the conversation using a conversation history. This conversation history is then used to reflect and make decisions on whether to interject or remain silent after each utterance made by other agents in the group.

Utterance and Conversation history Utterance within the architecture represents text uttered by the agent currently speaking and observed by other agents in the group. We assume the agents could remember the entire conversation because the duration of group discussions emulated is relatively short (15 minutes discussion). Thus, each agent is programmed to store and maintain the conversation history as a data structure that persists across calls/prompts made to LLMs to choose an action during the conversation, akin to *working memory*. Specifically, this conversation history is maintained as a list data structure that consists of a series of (speaker_id, text) pairs in the order they appeared during the conversation.

At each prompt to LLM, the conversation history is provided as input for decision-making and utterance generation. The agents have no *episodic memory* since they only "participate" in this group task once, which does not require the storage of experience from multiple group decision cycles. We rely on the *implicit knowledge* stored in the LLM weights and do not initialize the agents with external semantic knowledge support.

Actions. Actions can be divided into two types: "Reasoning actions" and "Statement actions". The reasoning action consists of two sub-actions performed in a sequence: *ranking_update* and *floor_action*. In this sequence the agents are first prompted to update their ranking (*ranking_update*) of the 15 items at the end of each utterance they observe. The agents then synthesize their ranking and conversation history to make a *floor_action* decision: grab the conversation floor or release the floor. If the agent determines to talk, the statement action *talk* will be triggered to generate natural languages that convey its opinion.

Four prompts are involved in the simulation. The *Task Description* prompt is identical to the one used in the human experiment to describe the task to LLM agents, abbreviated for the sake of space. The *ranking_update* prompt asks the agents to consider the propositions by other agents during the conversation, integrate them, and update their ranking of the 15 items. The *floor_action* prompt reflects humans' decisions and actions on the conversation floor during the discussion, e.g., interject or remain silent. Finally, the *talk* prompt is used to generate text when the agent decides to speak up. The word limit is empirically set as 40 (maximum utterance length in human corpus) to avoid lengthy sentences. An auxiliary *replyTo* attribute is included to improve the coherence of the conversation and to explicitly show whether the speaker is specifically talking to one of the other agents or broadcasting to the entire group.

The prompts were designed to follow a role-based system to differentiate between system roles and user roles (Oren, Sagawa, Hashimoto, & Liang, 2019). The system roles are used to configure the LLM identity (i.e., survivor_x). We leave the customization of the model's tone, style, and persona for future exploration. The user roles are used to configure the task description and task prompt. The reasoning attribute is an auxiliary one to explicitly show the LLM deduction process, which has been widely used to improve their performance in a variety of applications such as knowledge-intensive tasks (Yao et al., 2022) and decision-making tasks (Shinn, Labash, & Gopinath, 2023).

Free-form Conversation Algorithm

Following the same procedure in the human-subject experiment, each agent is first instructed to complete the WST individually. The agents are then assigned to groups of 2, 3, and 4 members to complete the WST. The agents are instructed to collaborate and discuss their individual rankings and come to a consensus on a group ranking. Then, each agent is individually prompted to complete the post-task questionnaires.

Figure 2 illustrates the execution loop that allows free-form conversation (FFC) among agents. The speaker who utters the first sentence initiates the conversation and grabs possession of the "conversation floor." The remaining agents in the group observe what is being said by the speaking agent. The speaker keeps the floor until another agent tries to claim the floor. Meanwhile, the listening agents monitor conversation history, periodically deciding whether to claim the floor or remain silent. If no one attempts to claim the floor, the speaker keeps talking until the agent determines to release the floor to others. If more than one agent attempts to claim the floor, one of them is randomly chosen as the next speaker. When the conversation floor is free, and a consensus has not yet been reached, the agents are repeatedly prompted to reassess the situation and decide whether to speak up. If none of the agents recognizes the obligation to speak up and continue the discussion, the conversation is ceased, and the group task

ends in failure.

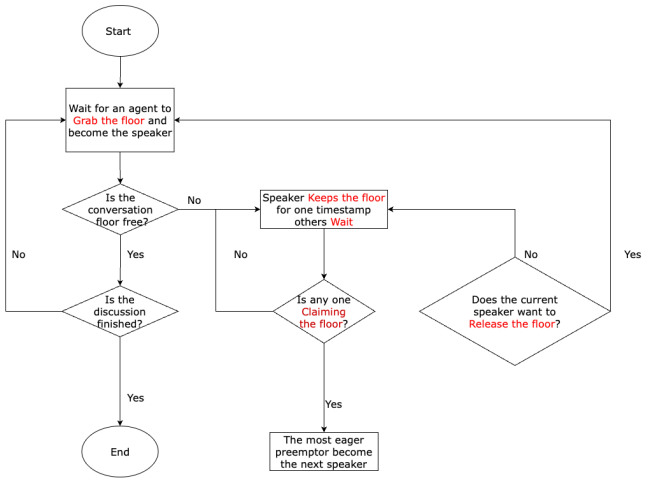

Figure 2: Flow diagram to describe the process that agents follow to generate free-form conversations

Group Decision-Making Annotations. Synthetic conversation corpus generated from the LLM simulation was annotated with the same four group decision-making annotations: Proposal, Agreement, Disagreement, and Confirmation. The annotation process was automated using the ProSeqo (Kozareva & Ravi, 2019) method, which currently ranks first on the leaderboard of dialogue act classification based on the Switchboard Dialog Act Corpus (Jurafsky, Shriberg, & Biasca, 1997). We fine-tuned the network on 60% of the annotated human corpus and achieved 72.4% agreement with human annotation on the remaining 40% human corpus. 72.4% is a satisfactory kappa value.

Sentiment Annotations. Synthetic corpus was also annotated for sentiment. We follow the same annotation scheme used to annotate the human corpus. The annotation process for sentiment was also automated. DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019; Wolf et al., 2019) was used to automate the sentiment annotation process. We first fine-tuned the network on 60% of the annotated human corpus and achieved 81.3% agreement with the human annotation on the remaining 40% human corpus. 81.3% is a satisfactory kappa value.

Corpus Evaluation Metrics

To systemically evaluate the human likeness of language agents' behavior and the potential to use them in group research, we propose to evaluate the agents on the following metrics.

Score and Meeting Length. We measure the task performance at both individual and group levels using the task score. **AIS** (Absolute Individual Score) is calculated based on the differences between the individual's ranking and the human expert ranking of each item $[100 - \sum_{i \in \|items\|} \|Rank^{Individual}(i) - Rank^{Expert}(i)\|]$. *AGS* (Absolute Group Score) is calculated based on the differences between the group's ranking and the human expert ranking of each item $[100 - \sum_{i \in \|items\|} \|Rank^{Group}(i) - Rank^{Expert}(i)\|]$. All members in groups with $AGS \geq 50$ can survive. With a score between $(40, 49]$, one might get frostbite. At most 3 members can survive with $AGS \in (30, 39]$. Groups with $AGS \leq 30$ are in serious danger. As a baseline, The distribution of random performance was analyzed using a Gaussian (normal) distribution model. The mean of the fitted Gaussian distribution was estimated to be $\mu = 15.34$ with a standard deviation $\sigma = 12.71$, $R^2 = 0.95$.

To analyze the efficiency of meetings, we measured the *Meeting Length* in terms of the number of words used during the conversations instead of the length of time since the LLM agents are not embedded in the real world and they can output text as fast as their CPU/GPU will allow. For a fair comparison between verbal conversation among humans and a text-based interaction among agents, the back channels (e.g., cough, nod, or unclear utterances like "uh") in the human corpus were excluded from the analysis.

Affect and Satisfaction. We measure the affection of the groups of agents based on the sentiment of each utterance and the peer evaluation in the post-experiment questionnaire. *Positivity* and *Negativity* are the number of utterances annotated as positive or negative. The *Satisfaction* is the group average of the *Overall Satisfaction* of each agent, which is the average of the five Likert-scale in the post-experiment questionnaire.

Group Action and Airtime. The decision-making behavior is measured in terms of both high-level speech acts and low-level turn-taking. *Group action proportions* are the number of utterances labeled as *proposal, agreement, disagreement, confirmation* divided by the total number of utterances of the group conversation. *Airtime proportion* is the number of words uttered by each speaker divided by the total word count of the conversation.

Results

Performance

Figure 3 illustrates the group score and meeting length of human and agent groups. One-way between-subjects ANOVA with human or agent as the main factor shows that the groups of language agents perform significantly better than human groups across three group sizes $[F(1,147) = 5.121, p < 0.05]$, while the length of agent meetings is significantly shorter than human meetings $[F(1,147) = 355.7, p < 0.0001]$. Post hoc comparisons using the Tukey HSD test indicated that the meeting length for human groups with 4 members (M = 2209.00, SD = 860.28) was significantly higher than groups with 3 (M = 1484.06, SD = 698.81) or 2 (M = 1349.83, SD = 88.85) members. There is no significant difference in meeting length for agent groups with different sizes.

Figure.4 further demonstrates the efficiency with which the agent groups deliberated compared to human groups. One-way between-subjects ANOVA was conducted to compare

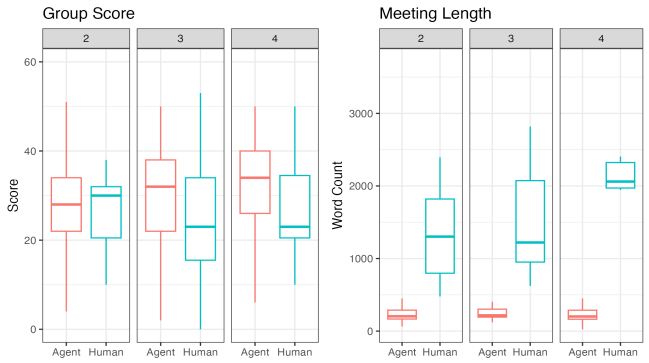

Figure 3: Left-panel: group score; Right-panel: meeting length in word count

human versus agent groups' improvement in groups of sizes 2, 3, and 4. The agent groups' improvement is significantly higher than humans in groups of size 2 [$F(1,45) = 2.83, p < 0.1$] and 4 [$F(1,45) = 1.97, p < 0.1$].

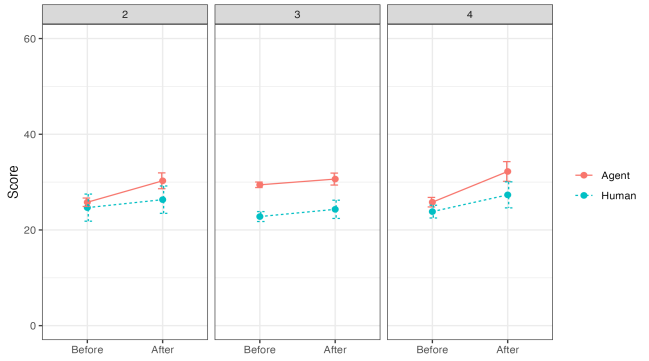

Figure 4: Average AIS before discussion → AGS after discussion

Affect

Table 1 shows the descriptive statistics of conversation sentiment and post-task peer evaluation scores. Both human and agent conversations have more utterances labeled positive than negative. The negativity of agent group conversations is significantly lower than that of human group conversations[$F(1,147) = 9.29, p < 0.01$]. As for peer evaluation, the agents report significantly lower scores in terms of *Time Management (Our group used its time wisely)*[$F(1,147) = 25.41, p < 0.001$] and *Efficiency (Our group struggled to work together efficiently on this task)* [$F(1,147) = 23.08, p < 0.001$], and significantly higher scores in terms of *Time Expectation (This task took longer than expected to complete.)* [$F(1,147) = 31.52, p < 0.001$], *Worked Well Together (Our group worked well together.)* [$F(1,147) = 5.966, p < 0.05$] and *Quality of Work (Overall, our group did a good job on this task.)* [$F(1,147) = 19.72, p < 0.001$]. In summary, agents show more positive affection toward peers and "care" more about efficiency.

Table 1: Descriptive Statistics of Sentiment and Peer-Evaluation Score (One-way ANOVA Significance. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

	Human	**Agent**
#Positive	M=13.785, SD=11.767	M=15.218, SD=7.322
#Negative**	M=5.285, SD=6.759	M=3.075, SD=2.046
Time Expectation***	M=3.059, SD=1.434	M=3.647, SD=0.659
Worked Well Together*	M=4.398, SD=0.684	M=4.559, SD=0.502
Time Management***	M=4.422, SD=0.778	M=4.014, SD=0.638
Efficiency***	M=4.351, SD=0.981	M=3.931, SD=0.641
Quality of Work***	M=4.315, SD=0.751	M=4.615, SD=0.498
Leadership.	M=3.452, SD=0.974	M=3.609, SD=0.661

Decision-Making

Figure.5 shows the distribution of group decision-making actions in human and agent groups. In general, agents make *Proposal* more often than humans, especially together with *Agreement*, *Disagreement*, and *Confirmation*. The agents also express *Disagreement* significantly more often than humans.

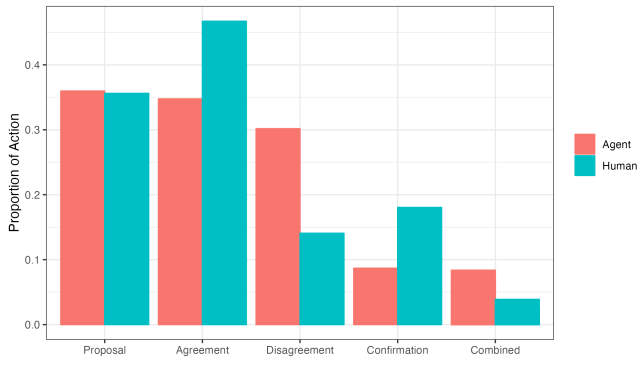

Figure 5: Distribution of Group Decision-making Actions. Note that one utterance can be labeled with more than one action, e.g., "I agree with the shortening over ski poles. Shall we rank flashlight next?" is labeled as *Proposal, Agreement*.

Figure 6 shows the distribution of airtime proportion in agent and human groups of various sizes. The distributions of human and agent groups are the most different in groups of size 4 (Agent: M=0.25, SD=0.097; Human: M=0.26, SD=0.0.169). In groups of size 4, more than 40% agents occupied 20% to 30% airtime, while only 33% humans occupied 20% to 30% airtime, which indicates that agents participated in the discussion more equally than humans. As an example, Figure.7 demonstrates the timeline of a 4-humans

group conversation, which is significantly dominated by one group member.

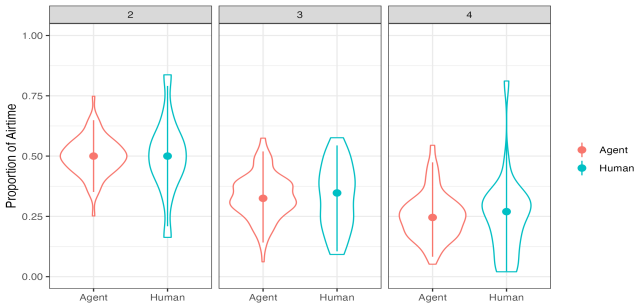

Figure 6: Distribution of Airtime Proportion

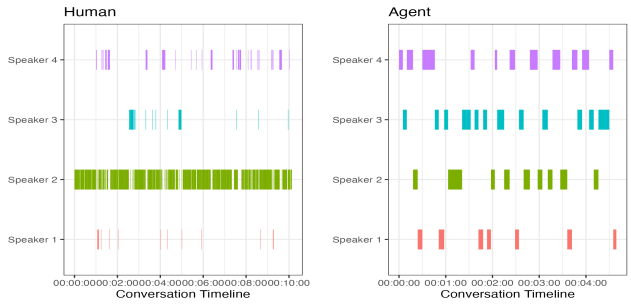

Figure 7: Conversation timeline of a 4-humans group vs a 4-agent group

Discussion

In this work, we introduce an algorithm that allows multiple LLM agents to engage in a problem-solving task through free conversation. Inspired by research in computational linguistics (Cohen & Perrault, 1979; Traum & Allen, 1994), which elucidates human conversational behavior in terms of beliefs, goals, intentions, and obligations, our agents are prompted to reason about the scenario and determine moments to contribute or terminate within the conversational flow. While many conversational humanoids (e.g., (Thórisson, 1999)) maintain a conversational plan or agenda, they often aim to optimize dialogue rather than human-like conversations with diverse styles. In our approach, LLMs are designed to consider conversation history and make sequential decisions regarding agents' participation, facilitating the emulation of natural, free-form conversations. We made several design choices to minimize explicit instructions, provide zero-shot prompts to the LLMs, intervene only when the conversation ceased, and delegate problem-solving tasks to the agents.

We produced a synthetic corpus of group conversations using LLM agents configured to work in groups of 2, 3, or 4 members. These agents engaged in the free conversation using our algorithm while tackling the Winter Survival Task. We compared the predictions generated by LLM agents with a publicly available human data set, evaluating them based on ranking scores, meeting length, and the change in ranking scores after group discussions. Our results indicate that LLM agent groups outperform human groups by achieving higher scores in shorter time frames. Furthermore, LLM agents enhance their scores after free-form group discussions compared to human groups. Analyses of post-task questionnaires and conversation dynamics indicate that agents are dissatisfied with their time management, perceiving tasks as taking longer than expected. Agents also exhibit a tendency to make positive remarks over negative ones, contrasting with human groups. These differences result from the underlying design philosophy used to build LLMs. It is possible that LLMs are intentionally designed to exhibit politeness and humility to please human users, potentially mitigating displeasure or frustration.

Analyses of LLM agent discussions reveal greater disagreement among agents within a group than among human groups. Agent groups also tend to craft more intricate statements that combine agreement and disagreement. However, agent discussions exhibit faster progression from one item to the next than human discussions. Agents achieve this by quickly proposing subsequent steps after agreement, disagreement, or confirmation. Also, agents engaged in turn-taking without requiring a predefined order. In contrast, human groups often have a dominant speaker, reducing some members to passive observers of the conversation. A possible explanation is that the human groups consist of different people with different background knowledge, biases, and preferences, while the agent groups can be less diverse.

Limitations & Future Work

Conversations encompass various modalities, including non-linguistic activities. Our free-form conversation algorithm lacks details such as backchannels, influenced by intonation and tone, beyond text. Challenges in conversations, such as overlapping talk or awkward silences, require a restoration mechanism. Future research could integrate embodied language models with sociometers to capture the conversation dynamics at a finer granularity (Parker, Cardenas, Dorr, & Hackett, 2020; Driess et al., 2023).

LLMs evolve continuously, and the simulation results reflect GPT-4 behavior. Despite its superior performance, the inner workings of GPT-4 remain hidden. Automatic annotation may be biased by pre-training data despite fine-tuning and yielding satisfactory annotations. LLM agents only learn by accumulating shared information. Augmenting them with the ability to learn from peers could foster more human-like group dynamics. Future work may involve augmenting agents with cognitive mechanisms to enhance social intelligence and foster believable conversations.

Acknowledgements

Compute resources and GPT model credits were provided by the Microsoft Accelerate Foundation Models Research grant

"Personalized Education with Foundation Models via Cognitive Modeling."

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... others (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning* (pp. 337–371).

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., ... others (2022). Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Braley, M., & Murray, G. (2018). The group affect and performance (gap) corpus. In *Proceedings of the group interaction frontiers in technology* (pp. 1–9).

Chang, J. P., Chiam, C., Fu, L., Wang, A. Z., Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... others (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, *19*(2), 25–35.

Cohen, P. R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive science*, *3*(3), 177–212.

Cole, R., Buchenroth-Martin, C., Weston, T., Devine, L., Myatt, J., Helding, B., ... others (2018). One-on-one and small group conversations with an intelligent virtual science tutor. *Computer Speech & Language*, *50*, 157–174.

Cuayáhuitl, H., Keizer, S., & Lemon, O. (2015). Strategic dialogue management via deep reinforcement learning. *arXiv preprint arXiv:1511.08099*.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*, 755–810.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... others (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Fine, G. A. (2014). The hinge: Civil society, group culture, and the interaction order. *Social Psychology Quarterly*, *77*(1), 5–26.

Forsell, J., Forslund Frykedal, K., & Hammar Chiriac, E. (2020). Group work assessment: Assessing social skills at group level. *Small Group Research*, *51*(1), 87–124.

Humphreys, B., Johnson, R. T., & Johnson, D. W. (1982). Effects of cooperative, competitive, and individualistic learning on students' achievement in science class. *Journal of research in science teaching*, *19*(5), 351–356.

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13* (Tech. Rep. No. 97-02). Boulder, CO: University of Colorado, Boulder Institute of Cognitive Science.

Karl, K. A., Peluchette, J. V., & Aghakhani, N. (2022). Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly. *Small Group Research*, *53*(3), 343–365.

Kozareva, Z., & Ravi, S. (2019). Proseqo: Projection sequence networks for on-device text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3894–3903).

Liddicoat, A. J. (2021). *An introduction to conversation analysis*. Bloomsbury Publishing.

Liu, C.-C., & Tsai, C.-C. (2008). An analysis of peer interaction patterns as discoursed by on-line small group problem-solving activity. *Computers & Education*, *50*(3), 627–639.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... others (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Oren, Y., Sagawa, S., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1–22).

Parker, J. N., Cardenas, E., Dorr, A. N., & Hackett, E. J. (2020). Using sociometers to advance small group research. *Sociological Methods & Research*, *49*(4), 1064–1102.

Raux, A. (2008). Flexible turn-taking for spoken dialog systems. *Language Technologies Institute, CMU Dec*, *12*.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7–55). Elsevier.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schmitt, A., & Ultes, S. (2015). Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction. *Speech Communication*, *74*, 12–36.

Seuren, L. M., Wherton, J., Greenhalgh, T., & Shaw, S. E. (2021). Whose turn is it anyway? latency and the organization of turn-taking in video-mediated interaction. *Journal of pragmatics*, *172*, 63–78.

Shinn, N., Labash, B., & Gopinath, A. (2023). Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Thórisson, K. R. (1999). Mind model for multimodal communicative creatures and humanoids. *Applied Artificial Intelligence*, *13*(4-5), 449–486.

Traum, D. R., & Allen, J. F. (1994). Discourse obligations in dialogue processing. *arXiv preprint cmp-lg/9407011*.

Tulshan, A. S., & Dhage, S. N. (2019). Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *Advances in signal processing and intelligent recognition systems: 4th international symposium sirs 2018, bangalore, india, september 19–22, 2018, revised selected papers 4* (pp. 190–201).

Vernham, Z., Granhag, P.-A., & Mac Giolla, E. (2016). Detecting deception within small groups: A literature review. *Frontiers in psychology*, *7*, 1012.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yadgarovna, M. F., & Husenovich, R. T. (2020). Advantages and disadvantages of the method of working in small groups in teaching higher mathematics. *Academy*(4 (55)), 65–68.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yeomans, M., Schweitzer, M. E., & Brooks, A. W. (2022). The conversational circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation. *Current Opinion in Psychology*, *44*, 293–302.

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., ... others (2023). Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.