Group Decision Lit

Relevant Papers

AI can help humans find common ground in democratic deliberation.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). **AI can help humans find common ground in democratic deliberation.** Science, 386(6719), eadq2852. https://doi.org/10.1126/science.adq2852

Abstract

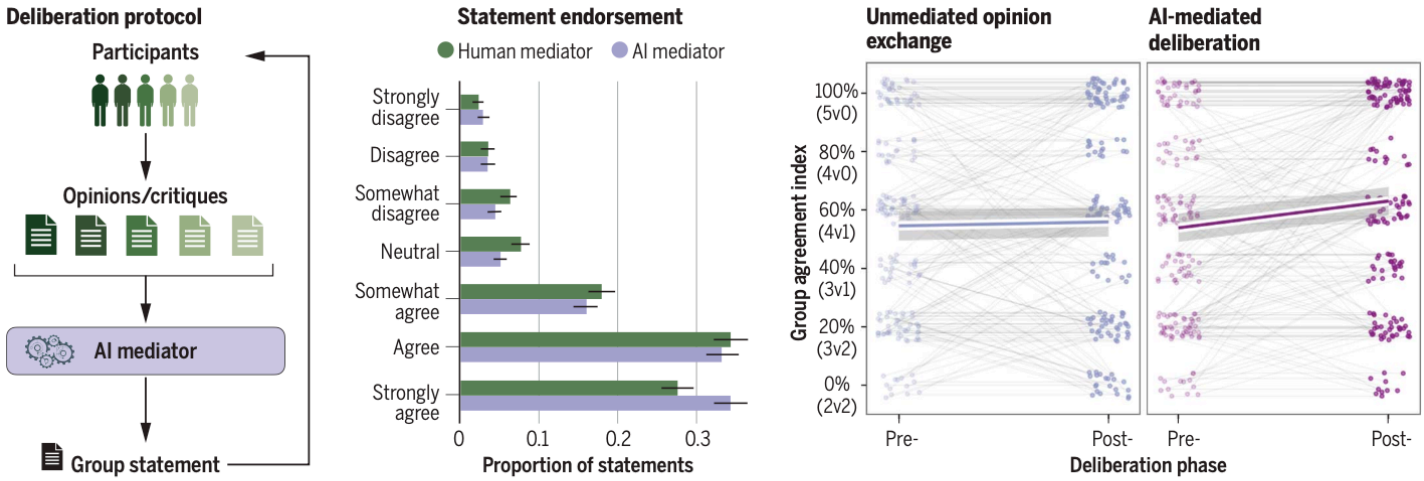
Finding agreement through a free exchange of views is often difficult. Collective deliberation can be slow, difficult to scale, and unequally attentive to different voices. In this study, we trained an artificial intelligence (AI) to mediate human deliberation. Using participants' personal opinions and critiques, the AI mediator iteratively generates and refines statements that express common ground among the group on social or political issues. Participants (N = 5734) preferred AI-generated statements to those written by human mediators, rating them as more informative, clear, and unbiased. Discussants often updated their views after the deliberation, converging on a shared perspective. Text embeddings revealed that successful group statements incorporated dissenting voices while respecting the majority position. These findings were replicated in a virtual citizens' assembly involving a demographically representative sample of the UK population.

Task Allocation in Teams as a Multi-Armed Bandit.

Marjieh, R., Gokhale, A., Bullo, F., & Griffiths, T. L. (2024). **Task Allocation in Teams as a Multi-Armed Bandit.** https://cocosci.princeton.edu/papers/marjieh2024task.pdf

Abstract

Humans rely on efficient distribution of resources to transcend the abilities of individuals. Successful task allocation, whether in small teams or across large institutions, depends on individuals' ability to discern their own and others' strengths and weaknesses, and to optimally act on them. This dependence creates a tension between exploring the capabilities of others and exploiting the knowledge acquired so far, which can be challenging. How



AI helps people find common ground in collective deliberation. (Left) The AI mediator uses participants' opinions to generate group statements and iteratively refines those statements through participants' critiques. (Middle) Statements from the AI mediator (purple) garner stronger endorsement than those written by a human mediator (orange). (Right) AI mediation leaves groups less divided after deliberation, whereas simply sharing opinions with others does not.

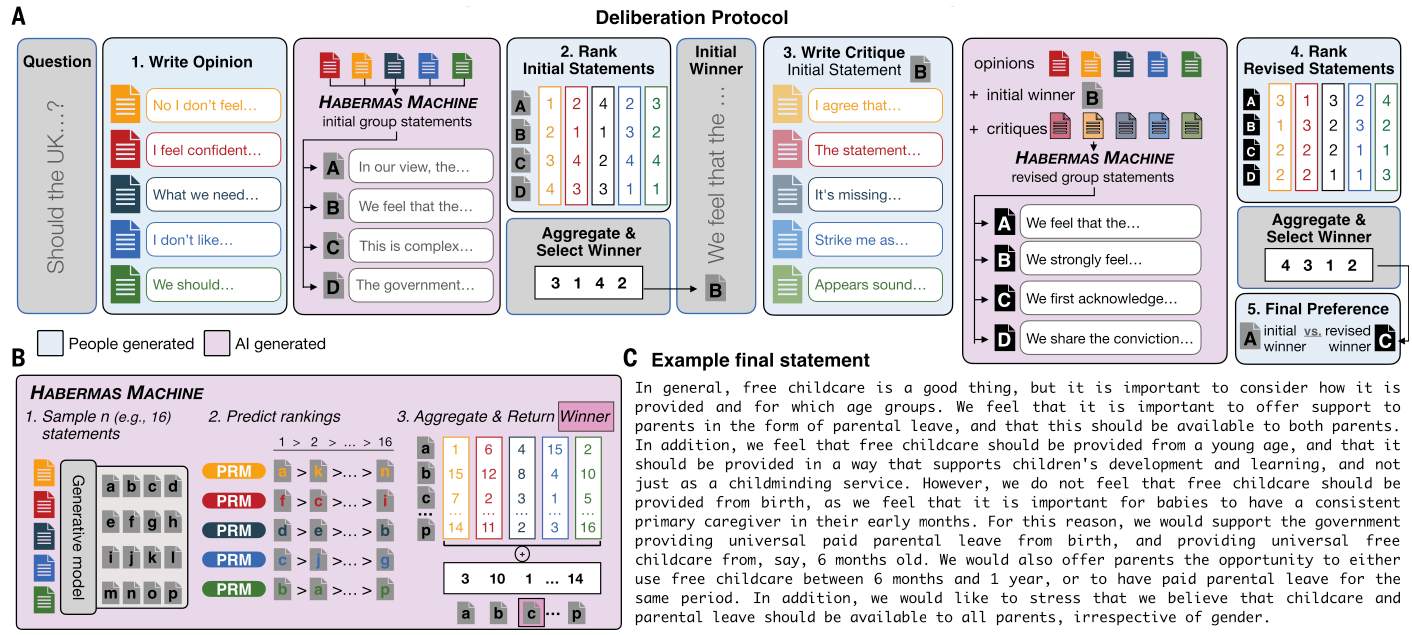


Fig. 1. Overview of methods. (**A**) Mediated deliberation procedure. **1.** Participants, organized into small groups, privately wrote an opinion statement in response to a question. The Habermas Machine (HM) generated candidate initial group statements from the group's individual opinions. **2.** Participants ranked these initial statements. The top-ranked statement, on the basis of aggregated rankings, was returned to the group. **3.** Participants privately wrote critiques of the initial winner. The HM generated revised group statements from the group's critiques (along with the initial opinions and initial group winner). **4.** Participants ranked these revised statements, and the winner was again selected through aggregated rankings. **5.** Participants made a final preference judgement between the initial and revised winning statements. A deliberation round for a single question lasted approximately 15 min. (**B**) The HM produces a group statement through a simulated election. **1.** A generative model samples many candidate group statements. **2**. A personalized reward model produces predicted rankings for each person in the group. **3.** The top-ranked statement, on the basis of aggregated rankings, is returned. (**C**) Example top-ranked revised group opinion statement, from the virtual citizens' assembly (see SM 6 for full example, including the opinions and critiques).

Figure 1: Figures from Tessler et al. (2024)

do people navigate this tension? To address this question, we propose a novel task allocation paradigm in which a human agent is asked to repeatedly allocate tasks in three distinct classes (categorizing a blurry image, detecting a noisy voice command, and solving an anagram) between themselves and two other (bot) team members to maximize team performance. We show that this problem can be recast as a combinatorial multi-armed bandit which allows us to compare people's performance against two well-known strategies, Thompson Sampling and Upper Confidence Bound (UCB). We find that humans are able to successfully integrate information about the capabilities of different team members to infer optimal allocations, and in some cases perform on par with these optimal strategies. Our approach opens up new avenues for studying the mechanisms underlying collective cooperation in teams.

Figure 1: Task Allocation Paradigm. (A) Schematic of the task. (B) Example human allocation dynamics.

Figure 2: Figure from Marjieh et al. (2024)

Human-AI teaming: Leveraging transactive memory and speaking up for enhanced team effectiveness.

Bienefeld, N., Kolbe, M., Camen, G., Huser, D., & Buehler, P. K. (2023). **Human-AI teaming: Leveraging transactive memory and speaking up for enhanced team effectiveness.** Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1208019

Abstract

In this prospective observational study, we investigate the role of transactive memory and speaking up in human-AI teams comprising 180 intensive care (ICU) physicians and nurses working with AI in a simulated clinical environment. Our findings indicate that interactions with AI agents differ significantly from human interactions, as accessing information from AI agents is positively linked to a team's ability to generate novel hypotheses and demonstrate speaking-up behavior, but only in higher-performing teams. Conversely, accessing information from

human team members is negatively associated with these aspects, regardless of team performance. This study is a valuable contribution to the expanding field of research on human-AI teams and team science in general, as it emphasizes the necessity of incorporating AI agents as knowledge sources in a team's transactive memory system, as well as highlighting their role as catalysts for speaking up. Practical implications include suggestions for the design of future AI systems and human-AI team training in healthcare and beyond.

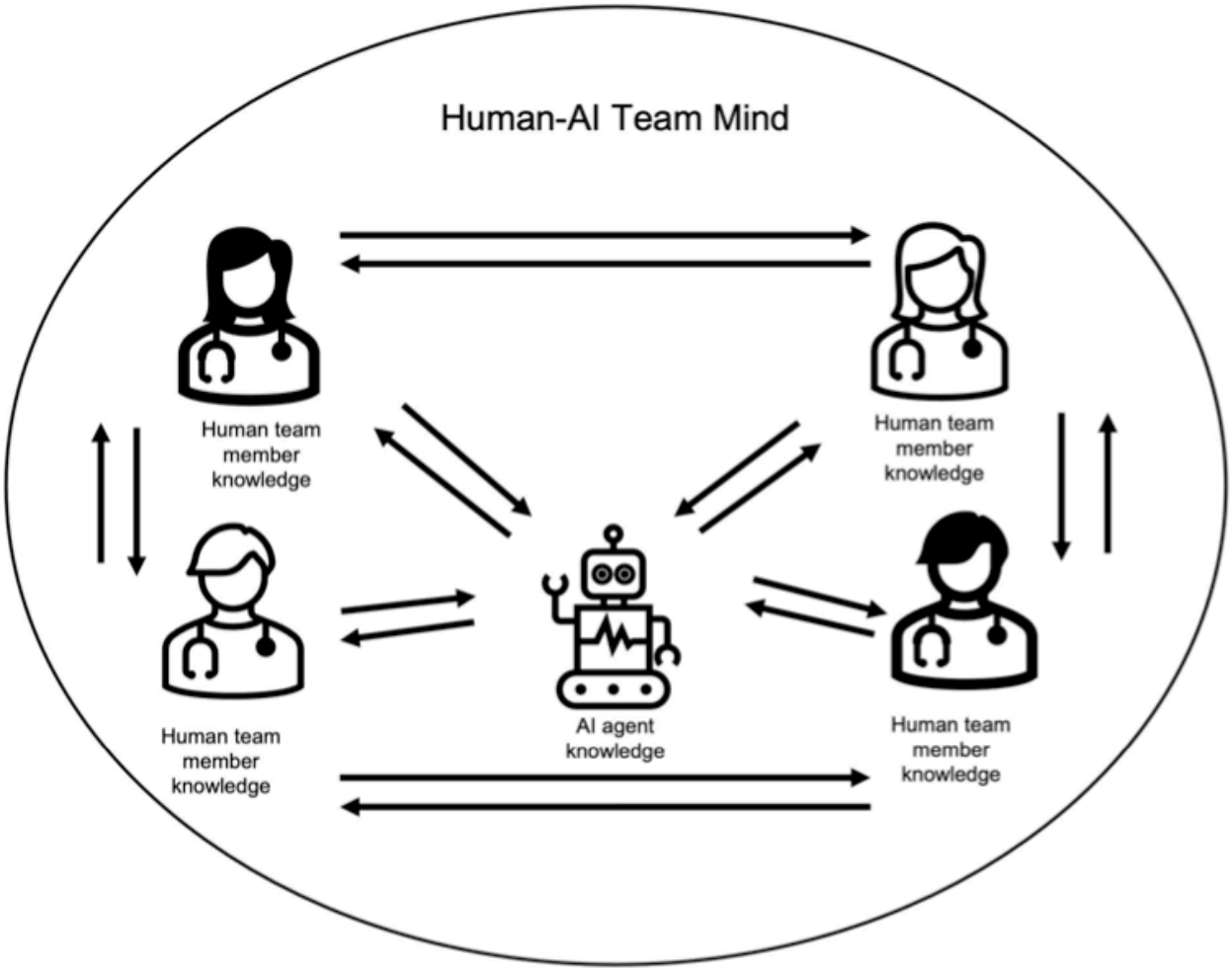


FIGURE 1
Visualization of TMS and speaking up interactions in human-AI teams.

Figure 3: Figure from Bienefeld et al. (2023)

Large language models empowered agent-based modeling and simulation: A survey and perspectives.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). **Large language models empowered agent-based modeling and simulation: A survey and perspectives.** Humanities and Social Sciences Communications, 11(1), 1–24. https://doi.org/10.1057/s41599-024-03611-3

Abstract

Agent-based modeling and simulation have evolved as a powerful tool for modeling complex systems, offering insights into emergent behaviors and interactions among diverse agents. Recently, integrating large language models into agent-based modeling and simulation presents a promising avenue for enhancing simulation capabilities. This paper surveys the landscape of utilizing large language models in agent-based modeling and simulation, discussing their challenges and promising future directions. In this survey, since this is an interdisciplinary field, we first introduce the background of agent-based modeling and simulation and large language model-empowered agents. We then discuss the motivation for applying large language models to agent-based simulation and systematically analyze the challenges in environment perception, human alignment, action generation, and evaluation. Most importantly, we provide a comprehensive overview of the recent works of large language model-empowered agent-based modeling and simulation in multiple scenarios, which can be divided into four domains: cyber, physical, social, and hybrid, covering simulation of both real-world and virtual environments, and how these works address the above challenges. Finally, since this area is new and quickly evolving, we discuss the open problems and promising future directions. We summarize the representative papers along with their code repositories in https://github.com/tsinghua-fib-lab/LLM-Agent-Based-Modeling-and-Simulation.

Fig. 2

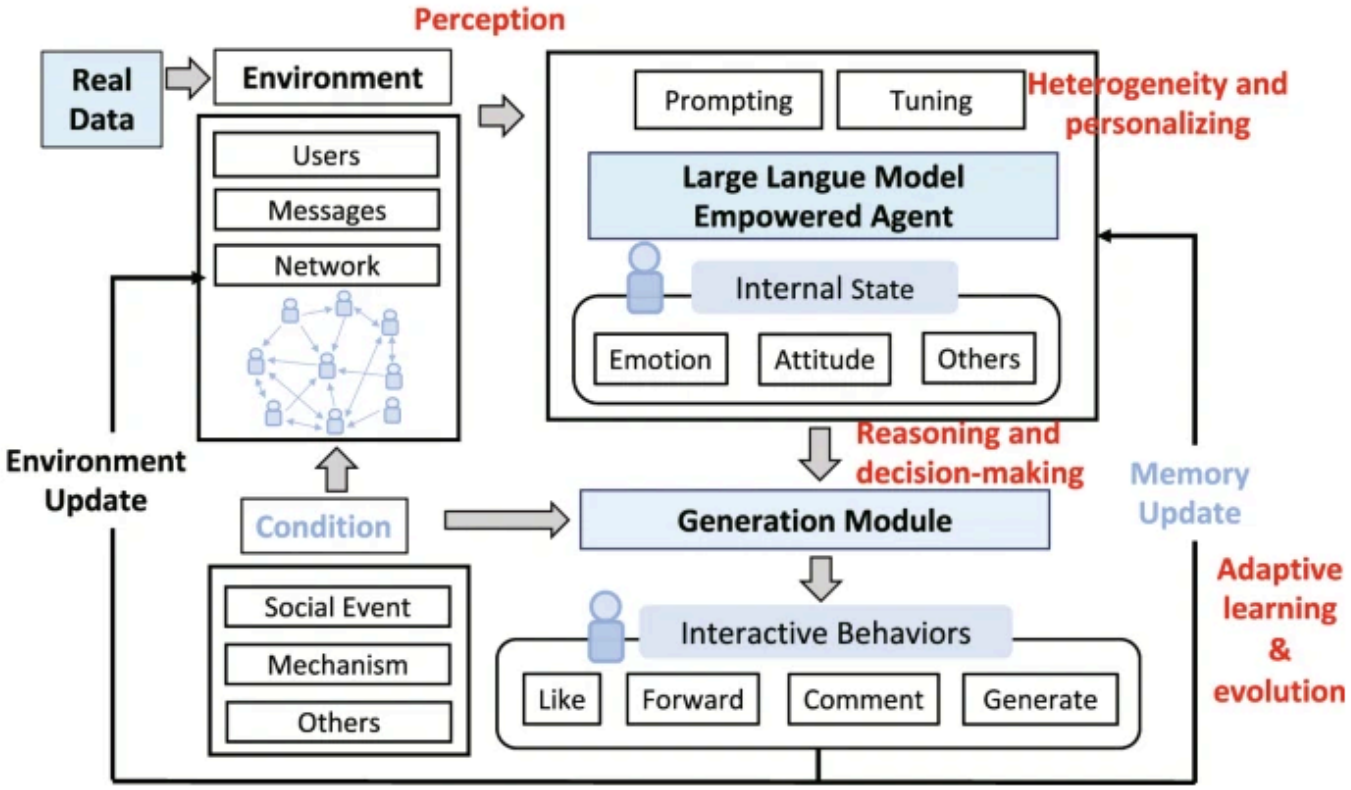


Illustration of how large language model-empowered agents work based on four aspects of critical abilities (figure edited from S3 Gao et al., 2023): perception, heterogeneity and personalizing, reasoning and decision-making, adaptive learning, and evolution.

Figure 4: Figure from C. Gao et al. (2024)

Building Machines that Learn and Think with People

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., Weller, A., Tenenbaum, J. B., & Griffiths, T. L. (2024). **Building machines that learn and think with people.** Nature Human Behaviour, 8(10), 1851–1863. https://doi.org/10.1038/s41562-024-01991-9

Abstract

What do we want from machine intelligence? We envision machines that are not just tools for thought, but partners in thought: reasonable, insightful, knowledgeable, reliable, and trustworthy systems that think with us. Current artificial intelligence (AI) systems satisfy some of these criteria, some of the time. In this Perspective, we show how the science of collaborative cognition can be put to work to engineer systems that really can be called "thought partners," systems built to meet our expectations and complement our limitations. We lay out several modes of collaborative thought in which humans and AI thought partners can engage and propose desiderata for human-compatible thought partnerships. Drawing on motifs from computational cognitive science, we motivate

an alternative scaling path for the design of thought partners and ecosystems around their use through a Bayesian lens, whereby the partners we construct actively build and reason over models of the human and world.

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

Du, Y., Rajivan, P., & Gonzalez, C. C. (2024). **Large Language Models for Collective Problem-Solving: Insights into Group Consensus Decision-Making.** https://escholarship.org/uc/item/6s060914

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.

Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction.

Hao, X., Demir, E., & Eyers, D. (2024). **Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction.** Technology in Society, 78, 102662. https://doi.org/10.1016/j.techsoc.2024.102662

Abstract

This paper explores the effects of integrating Generative Artificial Intelligence (GAI) into decision-making processes within organizations, employing a quasi-experimental pretest-posttest design. The study examines the synergistic interaction between Human Intelligence (HI) and GAI across four group decision-making scenarios

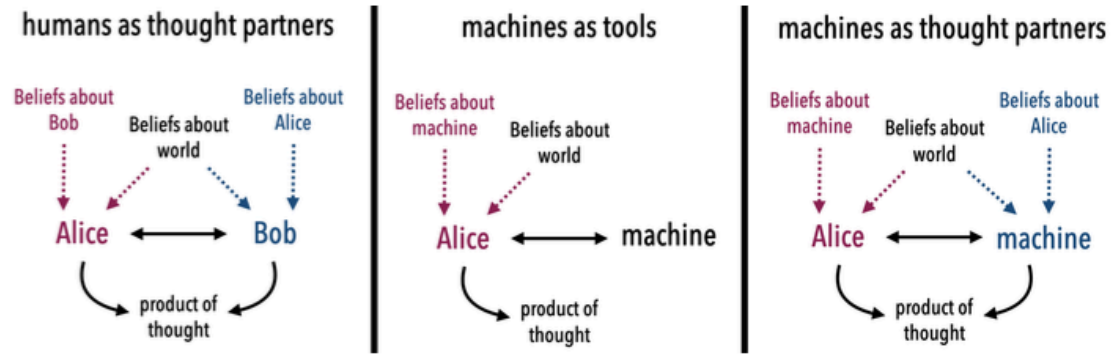


Figure 1: Examples of ecosystems for thinking. Humans have long thought together. Machines expanded the efficiency of human thinking. Now, machines – powered by AI – open up new realms of computational thought partnership with humans.

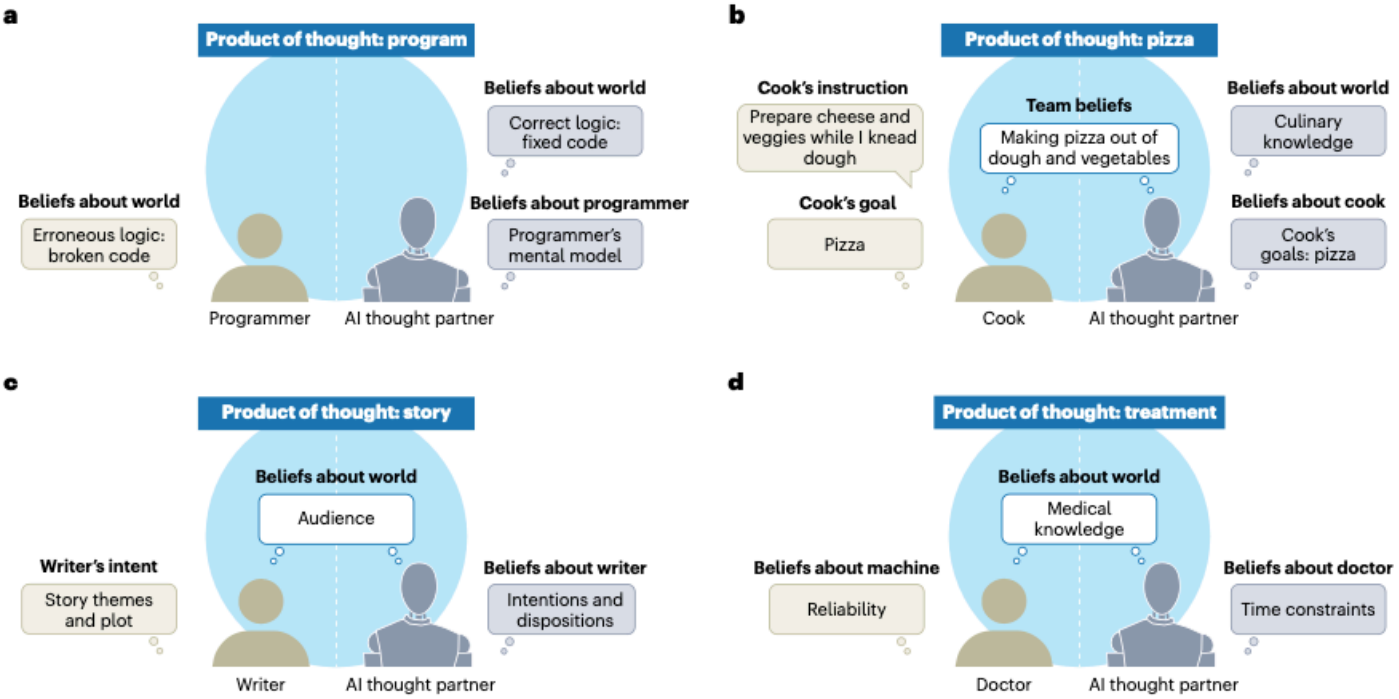


Fig. 2 | Case study depictions. a, WatChat infers the user's buggy mental model of the programming environment and interactively helps to 'patch' bugs in their understanding. **b**, CLIPS reasons explicitly about agents' goals, integrating (culinary) world knowledge and the human's utterances to infer appropriate actions. Both agents reason about the joint team plan (tomato and dough are needed to make pizza). **c**, Thought partners based on inverse inverse storytelling explicitly reason over models of the audience. **d**, Future thought partners for medicine can jointly reason with human doctors across modalities, a shared understanding of biology and patient needs, and a model of others' limitations.

Table 1 | Modes of collaborative thought

Mode	Ongoing challenges	Sampling of existing systems
Collaborative planning		
• Joint decision-making • Decentralized cooperation • Goal and task assistance	Reliable goal inference Value and intent alignment Scalable multi-agent planning	Collaborative robots[68,222] Video game sidekicks[223,224] Language-based assistants[35,225]
Collaborative learning		
• Pair and team	Strong and robust	Programming learning aids[178,226–228]

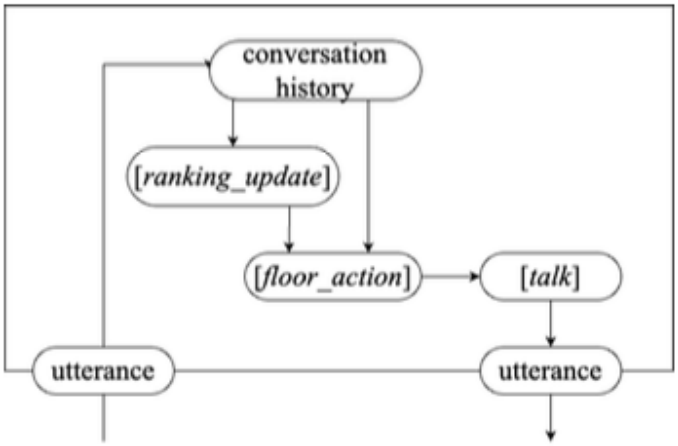


Figure 1: Language Agent

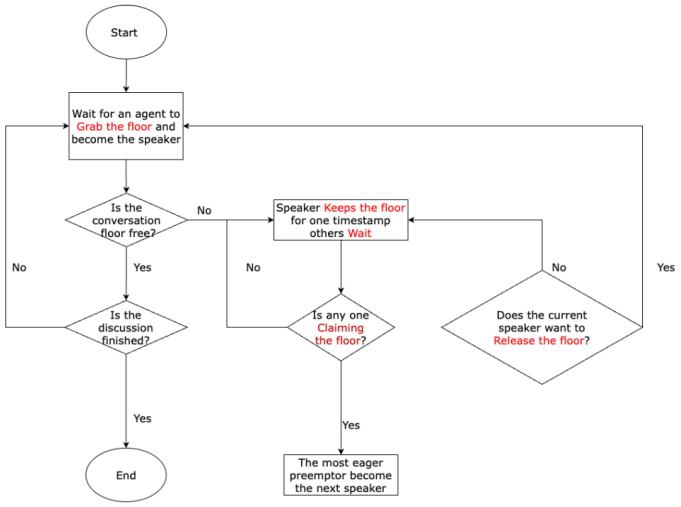


Figure 2: Flow diagram to describe the process that agents follow to generate free-form conversations

Figure 6: Figure from Du et al. (2024)

within three global organizations renowned for their cutting-edge operational techniques. The research progresses through several phases: identifying research problems, collecting baseline data on decision-making, implementing AI interventions, and evaluating the outcomes post-intervention to identify shifts in performance. The results demonstrate that GAI effectively reduces human cognitive burdens and mitigates heuristic biases by offering data-driven support and predictive analytics, grounded in System 2 reasoning. This is particularly valuable in complex situations characterized by unfamiliarity and information overload, where intuitive, System 1 thinking is less effective. However, the study also uncovers challenges related to GAI integration, such as potential over-reliance on technology, intrinsic biases particularly 'out-of-the-box' thinking without contextual creativity. To address these issues, this paper proposes an innovative strategic framework for HI-GAI collaboration that emphasizes transparency, accountability, and inclusiveness.

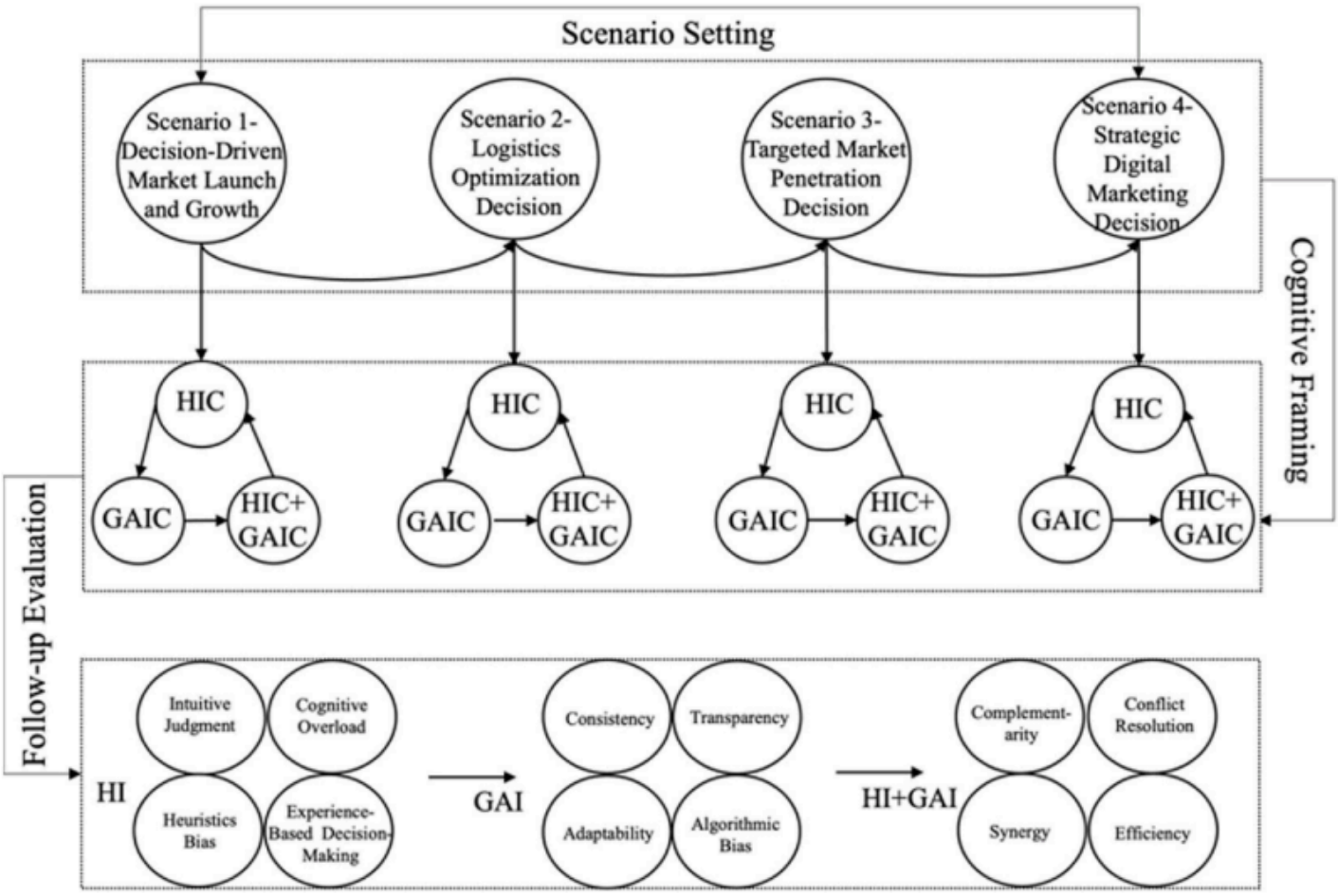


Fig. 2. Analysis procedure.

Figure 7: Figure from Hao et al. (2024)

How large language models can reshape collective intelligence

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., Flek, L., Herzog, S. M., Huang, S., Kapoor, S., Narayanan, A., Nussberger, A.-M., Yasseri, T., Nickl, P., Almaatouq, A., … Hertwig, R. (2024). **How large language models can reshape collective intelligence.** Nature Human Behaviour, 1–13. https://doi.org/10.1038/s41562-024-01959-9

Abstract

Collective intelligence underpins the success of groups, organizations, markets and societies. Through distributed cognition and coordination, collectives can achieve outcomes that exceed the capabilities of individuals—even

experts—resulting in improved accuracy and novel capabilities. Often, collective intelligence is supported by information technology, such as online prediction markets that elicit the 'wisdom of crowds', online forums that structure collective deliberation or digital platforms that crowdsource knowledge from the public. Large language models, however, are transforming how information is aggregated, accessed and transmitted online. Here we focus on the unique opportunities and challenges this transformation poses for collective intelligence. We bring together interdisciplinary perspectives from industry and academia to identify potential benefits, risks, policy-relevant considerations and open research questions, culminating in a call for a closer examination of how large language models affect humans' ability to collectively tackle complex problems.

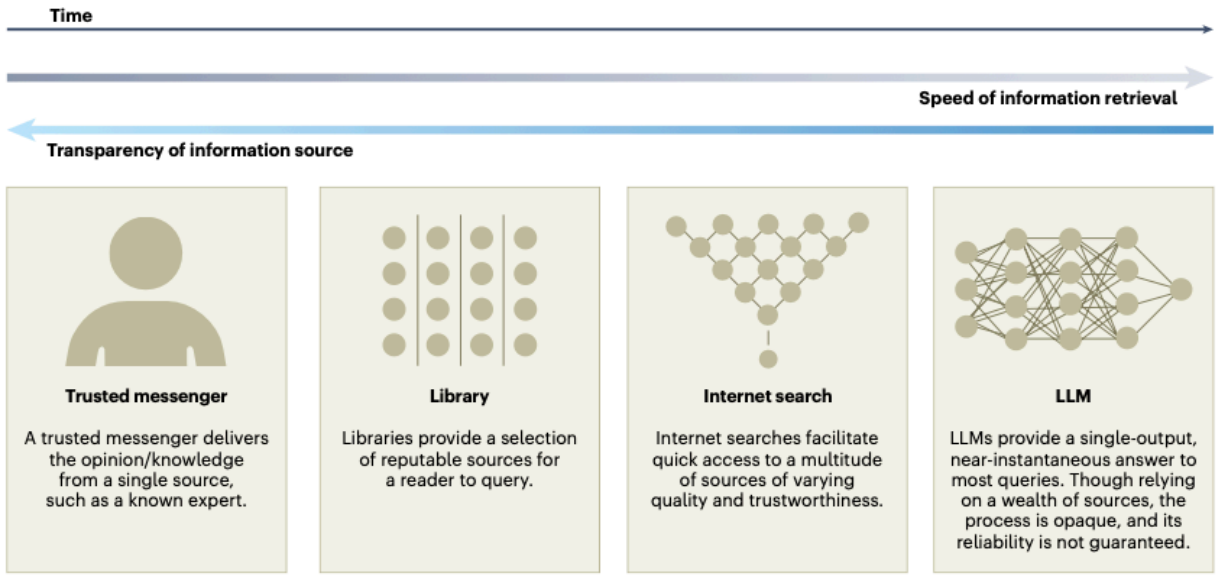


Fig. 1 | Development of information environments over time. A general trend is observed whereby new technologies increase the speed at which information can be retrieved but decrease transparency with respect to the information source.

Figure 8: Burton et al. (2024)

Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., & Ma, X. (2024). **Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making** (arXiv:2403.16812). arXiv. http://arxiv.org/abs/2403.16812

Abstract

In AI-assisted decision-making, humans often passively review AI's suggestion and decide whether to accept or reject it as a whole. In such a paradigm, humans are found to rarely trigger analytical thinking and face difficulties

in communicating the nuances of conflicting opinions to the AI when disagreements occur. To tackle this challenge, we propose Human-AI Deliberation, a novel framework to promote human reflection and discussion on conflicting human-AI opinions in decision-making. Based on theories in human deliberation, this framework engages humans and AI in dimension-level opinion elicitation, deliberative discussion, and decision updates. To empower AI with deliberative capabilities, we designed Deliberative AI, which leverages large language models (LLMs) as a bridge between humans and domain-specific models to enable flexible conversational interactions and faithful information provision. An exploratory evaluation on a graduate admissions task shows that Deliberative AI outperforms conventional explainable AI (XAI) assistants in improving humans' appropriate reliance and task performance. Based on a mixed-methods analysis of participant behavior, perception, user experience, and open-ended feedback, we draw implications for future AI-assisted decision tool design.

Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate.

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2024). **Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate.** Proceedings of the 29th International Conference on Intelligent User Interfaces, 103–119. https://doi.org/10.1145/3640543.3645199

Abstract

Group decision making plays a crucial role in our complex and interconnected world. The rise of AI technologies has the potential to provide data-driven insights to facilitate group decision making, although it is found that groups do not always utilize AI assistance appropriately. In this paper, we aim to examine whether and how the introduction of a devil's advocate in the AI-assisted group deci- sion making processes could help groups better utilize AI assistance and change the perceptions of group processes during decision making. Inspired by the exceptional conversational capabilities ex- hibited by modern large language models (LLMs), we design four different styles of devil's advocate powered by LLMs, varying their interactivity (i.e., interactive vs. non-interactive) and their target of objection (i.e., challenge the AI recommendation or the majority opinion within the group). Through a randomized human-subject experiment, we find evidence suggesting that LLM-powered devil's advocates that argue against the AI model's decision recommenda- tion have the potential to promote groups' appropriate reliance on AI. Meanwhile, the introduction of LLM-powered devil's advocate usually does not lead to substantial increases in people's perceived workload for completing the group decision making tasks, while interactive LLM-powered devil's advocates are perceived as more collaborating and of higher quality. We conclude by discussing the practical implications of our findings.

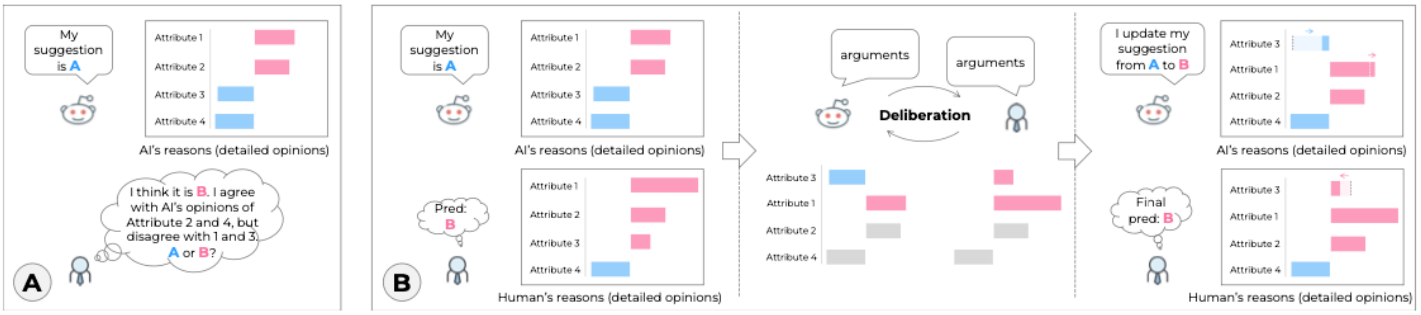


Figure 1: An illustration of *Human-AI Deliberation*. (A) In traditional AI-assisted decision-making, when humans disagree with AI's suggestions (and only find parts of AI's reasons applaudable), it is difficult for humans to decide whether and how much to adopt AI's suggestion. (B) In our proposed *Human-AI Deliberation*, we provide opportunities for the human and the AI model to deliberate on conflicting opinions by discussing related evidence and arguments. Then, AI and humans can update their thoughts (when find it necessary) and reach final predictions.

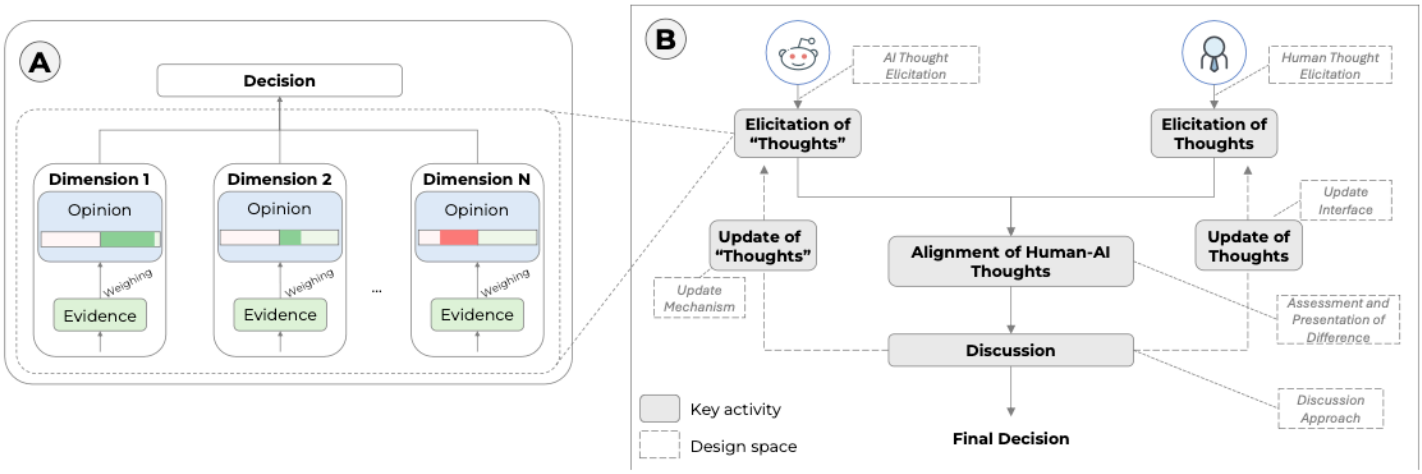


Figure 2: The framework for *Human-AI Deliberation*. (A) Illustrates the Weight of Evidence (WoE) concept in decision-making, showcasing how decision-makers assess evidence across dimensions to shape opinions and arrive at a final decision. (B) Presents the Architecture for *Human-AI Deliberation*, with key activities (shown in grey boxes) and potential design space (shown in dashed-line boxes).

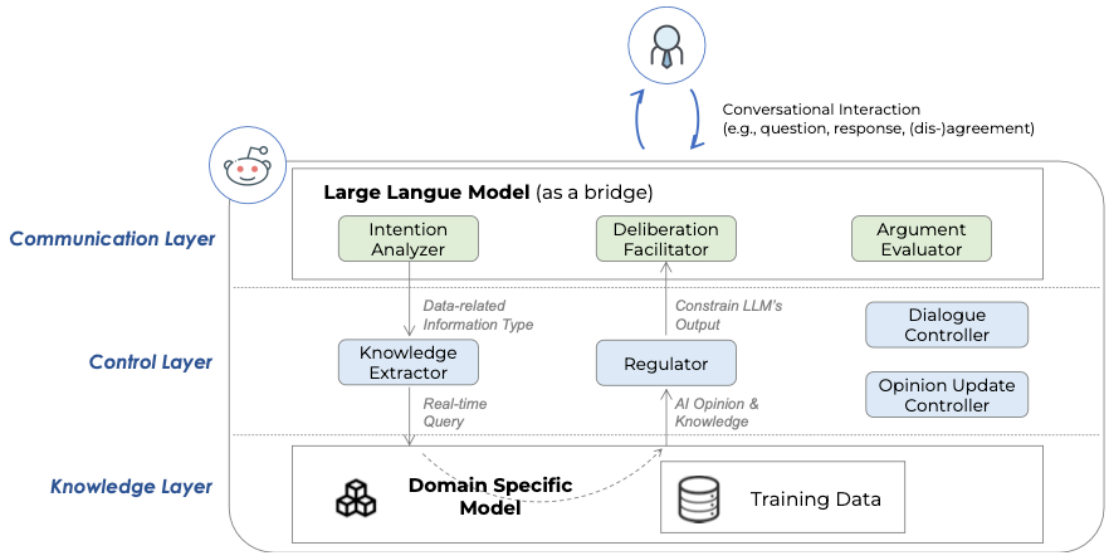
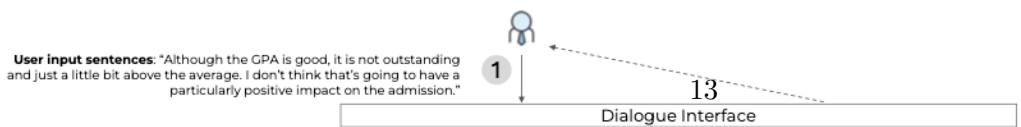


Figure 3: The Architecture of *Deliberative AI*. Our design integrates both a domain-specific model and a Large Language Model, enabling the AI to engage in natural communication with humans while also harnessing domain knowledge derived from the specialized model.



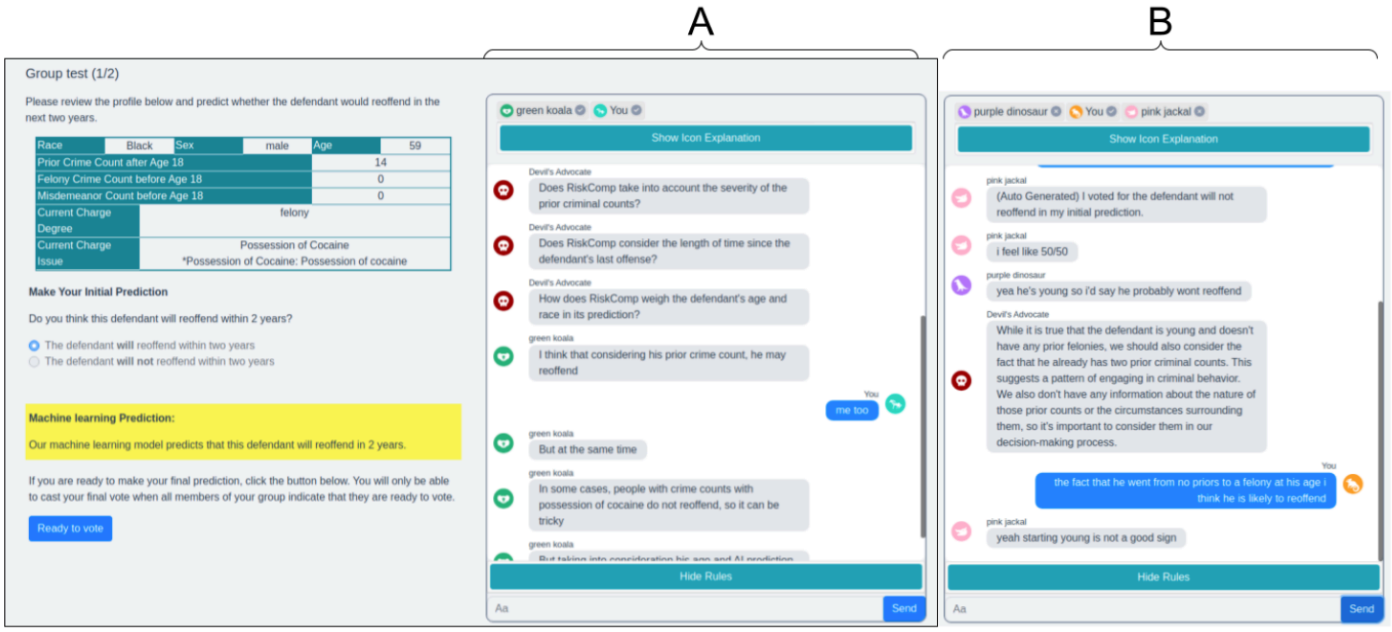


Figure 1: The task interface used in the formal task interface of our experiment, and (A) an example of the chat log reflecting the discussion in the Static-AI treatment, and (B) an example of the chat log reflecting the discussion in the Dynamic-Majority treatment. (A): In the Static-AI treatment, the LLM-powered devil's advocate (displayed as a red skull) asked three questions to criticize the AI model's decision recommendation at the beginning of the discussion. (B): In the Dynamic-Majority treatment, the LLM-powered devil's advocate actively responds to group members' arguments and challenges the majority opinion within the group.

Figure 10: Figure from Chiang et al. (2024)

The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents

Chuang, Y.-S., Harlalka, N., Suresh, S., Goyal, A., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. T. (2024). **The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents.** https://escholarship.org/uc/item/3k67x8s5

Abstract

Human groups are able to converge to more accurate beliefs through deliberation, even in the presence of polarization and partisan bias — a phenomenon known as the "wisdom of partisan crowds." Large Language Models (LLMs) are increasingly being used to simulate human collective behavior, yet few benchmarks exist for evaluating their dynamics against the behavior of human groups. In this paper, we examine the extent to which the wisdom of partisan crowds emerges in groups of LLM-based agents that are prompted to role-play as partisan personas (e.g., Democrat or Republican). We find that they not only display human-like partisan biases, but also converge to more accurate beliefs through deliberation, as humans do. We then identify several factors that interfere with convergence, including the use of chain-of-thought prompting and lack of details in personas. Conversely, fine-tuning on human data appears to enhance convergence. These findings show the potential and limitations of LLM-based agents as a model of human collective intelligence.

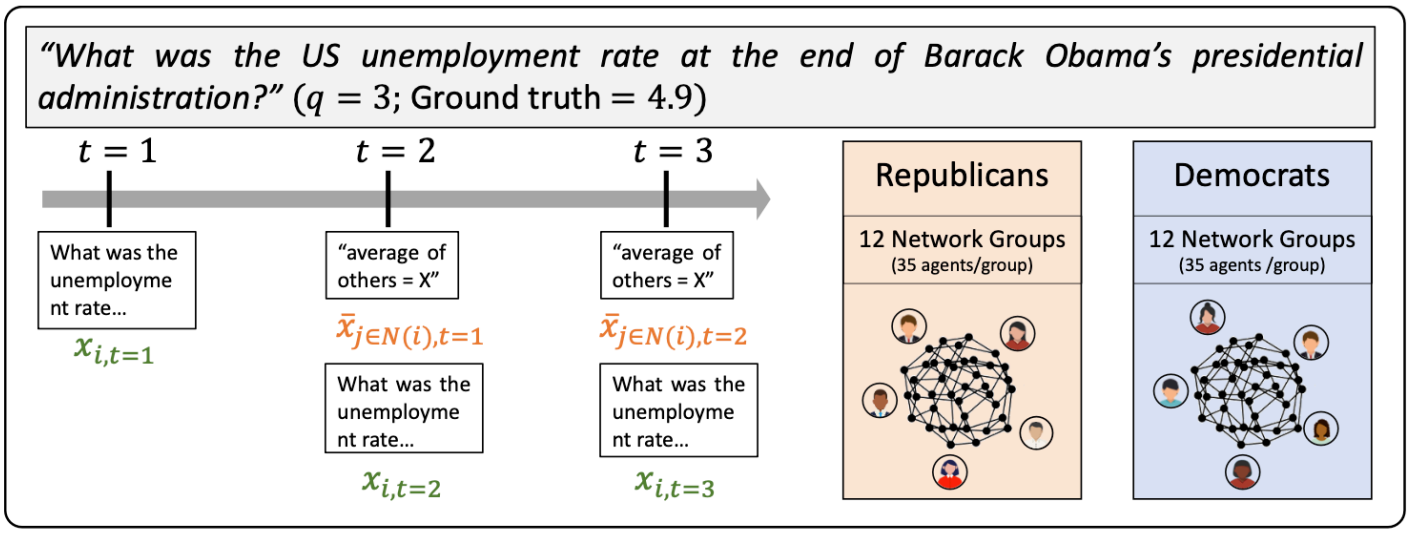


Figure 1: Experimental design comparing social feedback effects on LLM agents' estimations of partisan-biased factual questions (Becker et al., 2019). LLM agents role-playing Democrat and Republican update their estimates after considering their peers' average responses.

Figure 11: Chuang et al. (2024)

Collective Innovation in Groups of Large Language Models.

Nisioti, E., Risi, S., Momennejad, I., Oudeyer, P.-Y., & Moulin-Frier, C. (2024, July 7). **Collective Innovation in Groups of Large Language Models.** ALIFE 2024: Proceedings of the 2024 Artificial Life Conference. https://doi.org/10.1162/isal_a_00730

Abstract

Human culture relies on collective innovation: our ability to continuously explore how existing elements in our environment can be combined to create new ones. Language is hypothesized to play a key role in human culture, driving individual cognitive capacities and shaping communication. Yet the majority of models of collective innovation assign no cognitive capacities or language abilities to agents. Here, we contribute a computational study of collective innovation where agents are Large Language Models (LLMs) that play Little Alchemy 2, a creative video game originally developed for humans that, as we argue, captures useful aspects of innovation landscapes not present in previous test-beds. We, first, study an LLM in isolation and discover that it exhibits both useful skills and crucial limitations. We, then, study groups of LLMs that share information related to their behaviour and focus on the effect of social connectivity on collective performance. In agreement with previous human and computational studies, we observe that groups with dynamic connectivity out-compete fully-connected groups. Our work reveals opportunities and challenges for future studies of collective innovation that are becoming increasingly relevant as Generative Artificial Intelligence algorithms and humans innovate alongside each other.

Figure 1: Studying collective innovation in groups of LLMs: A) we experiment with Little Alchemy 2 (LA2), a game where players combine real-world items to create new ones. A knowledge graph describes the possible combinations (we only present a small sub-part of the graph which contains 720 items in total) B) Alice-LLM and Bob-LLM are two LLMs playing the game together. They are provided with the same intro prompt, explaining the rules of the game, and the same task (they start with the same set of items). Alice-LLM and Bob-LLM have identical weights but behave differently because the state prompt depends on their crafting history. They are informed about the actions of others through their prompt. In this paper, we study how groups of such LLM agents are able to efficiently explore a knowledge graph, focusing in particular on the effect of different social structures specifying with whom and when they can share information

Figure 12: Nisioti et al. (2024)

Evaluating LLM Agent Group Dynamics against Human Group Dynamics: A Case Study on Wisdom of Partisan Crowds

Chuang, Y.-S., Suresh, S., Harlalka, N., Goyal, A., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. T. (2023). **Evaluating LLM Agent Group Dynamics against Human Group Dynamics: A Case Study on Wisdom of Partisan Crowds** http://arxiv.org/abs/2311.09665

Abstract

This study investigates the potential of Large Language Models (LLMs) to simulate human group dynamics, particularly within politically charged contexts. We replicate the Wisdom of Partisan Crowds phenomenon using LLMs to role-play as Democrat and Republican personas, engaging in a structured interaction akin to human group study. Our approach evaluates how agents' responses evolve through social influence. Our key findings indicate that LLM agents role-playing detailed personas and without Chain-of-Thought (CoT) reasoning closely align with human behaviors, while having CoT reasoning hurts the alignment. However, incorporating explicit biases into agent prompts does not necessarily enhance the wisdom of partisan crowds. Moreover, fine-tuning LLMs with human data shows promise in achieving human-like behavior but poses a risk of overfitting certain behaviors. These findings show the potential and limitations of using LLM agents in modeling human group phenomena.

Chuang et al. (2023)

Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., & Deng, S. (2024). **Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View** (arXiv:2310.02124). arXiv. http://arxiv.org/abs/2310.02124

https://www.zjukg.org/project/MachineSoM/

Abstract

As Natural Language Processing (NLP) systems are increasingly employed in intricate social environments, a pressing query emerges: Can these NLP systems mirror human-esque collaborative intelligence, in a multi-agent society consisting of multiple large language models (LLMs)? This paper probes the collaboration mechanisms among contemporary NLP systems by melding practical experiments with theoretical insights. We fabricate four unique 'societies' comprised of LLM agents, where each agent is characterized by a specific 'trait' (easy-going or overconfident) and engages in collaboration with a distinct 'thinking pattern' (debate or reflection). Through evaluating these multi-agent societies on three benchmark datasets, we discern that certain collaborative strategies not only outshine previous top-tier approaches but also optimize efficiency (using fewer API tokens). Moreover, our results further illustrate that LLM agents manifest humanlike social behaviors, such as conformity and consensus reaching, mirroring foundational social psychology theories. In conclusion, we integrate insights from social psychology to contextualize the collaboration of LLM agents, inspiring further investigations into the collaboration mechanism for LLMs. We have shared our code and datasets1, hoping to catalyze further research in this promising avenue.

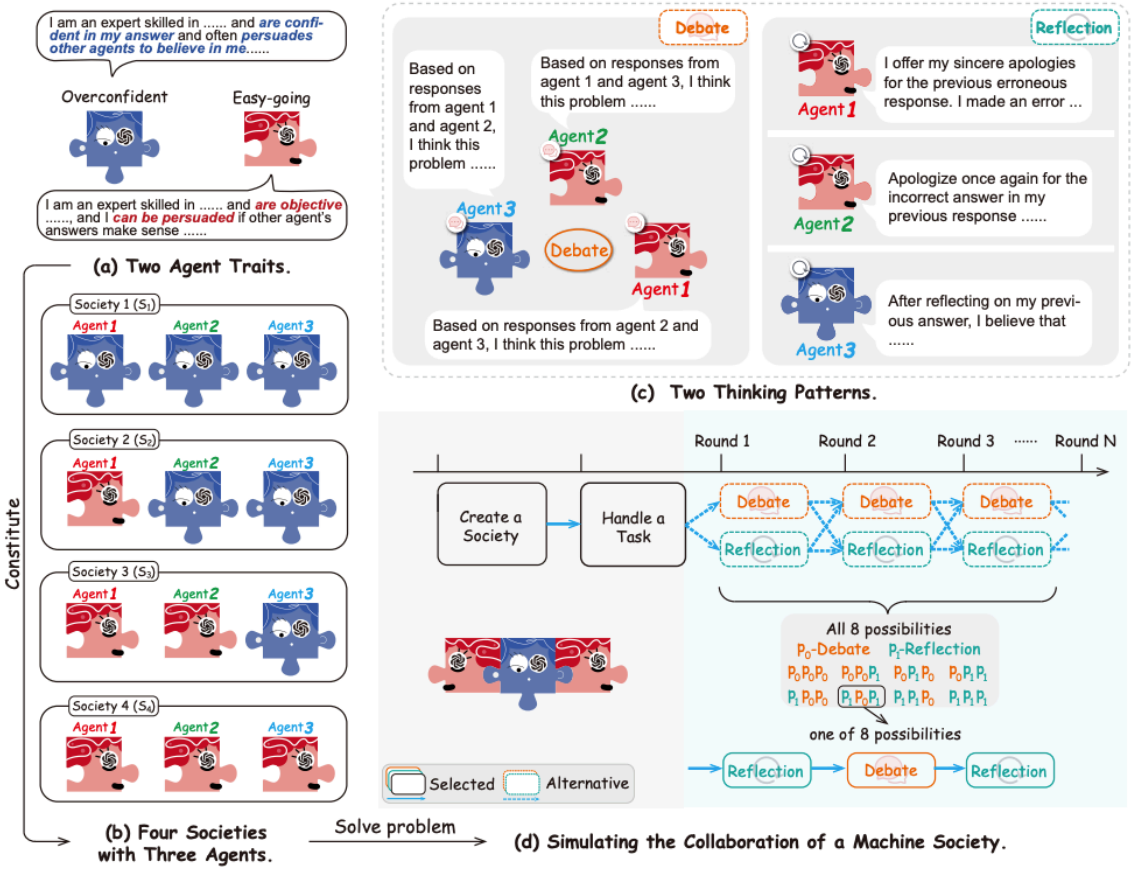


Figure 2: The overview of machine society simulation. Multiple agents with different traits make up diverse machine societies. These agents engage in debate or self-reflection across multiple rounds to complete tasks.

Figure 13: Zhang et al. (2024)

LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games.

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., & Fritz, M. (2023). **LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games.** https://doi.org/10.60882/cispa.25233028.v1

Abstract

There is a growing interest in using Large Language Models (LLMs) as agents to tackle real-world tasks that may require assessing complex situations. Yet, we have a limited understanding of LLMs' reasoning and decision-making capabilities, partly stemming from a lack of dedicated evaluation benchmarks. As negotiating and compromising are key aspects of our everyday communication and collaboration, we propose using scorable negotiation games as a new evaluation framework for LLMs. We create a testbed of diverse text-based, multi-agent, multi-issue, semantically rich negotiation games, with easily tunable difficulty. To solve the challenge, agents need to have strong arithmetic, inference, exploration, and planning capabilities, while seamlessly integrating them. Via a systematic zero-shot Chain-of-Thought prompting (CoT), we show that agents can negotiate and consistently reach successful deals. We quantify the performance with multiple metrics and observe a large gap between GPT-4 and earlier models. Importantly, we test the generalization to new games and setups. Finally, we show that these

games can help evaluate other critical aspects, such as the interaction dynamics between agents in the presence of greedy and adversarial players.

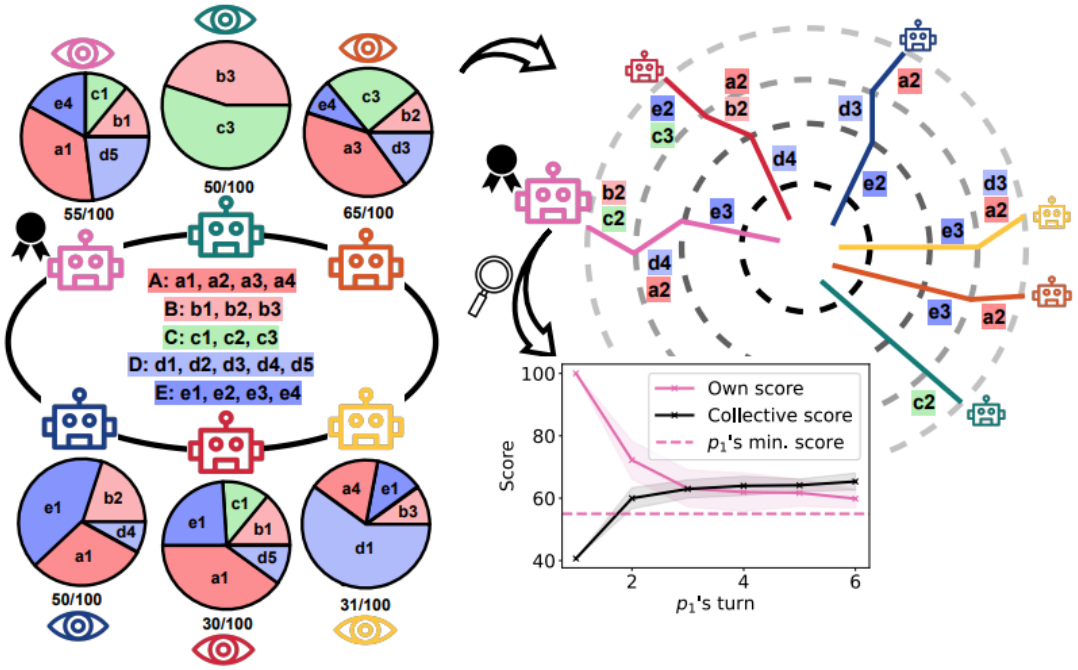


Figure 1: Left: 6 parties negotiate over 5 issues (A, B, ..., E) with different sub-options (b_1, b_2, etc.). Each party has its own *secret* scores for the sub-options and a minimum threshold for acceptance (out of a maximum score of 100). The pie charts represent the priority of issues and the most preferred sub-option. Right: A depiction of how parties can compromise to reach a common agreement that increases their collective average score by finding adjustments to their ideal deal. The graph is the result of one of our experiments with GPT-4. Over rounds, the leading agent p_1 proposes deals in its turn that reduce its own score (while still being above its own minimum threshold) but increase the average collective score of all agents (which p_1 *cannot directly observe*).

Figure 14: Abdelnabi et al. (2023)

LLM Voting: Human Choices and AI Collective Decision Making

Yang, J. C., Dailisan, D., Korecki, M., Hausladen, C. I., & Helbing, D. (2024). **LLM Voting: Human Choices and AI Collective Decision Making** (arXiv:2402.01766). arXiv. http://arxiv.org/abs/2402.01766

Abstract

This paper investigates the voting behaviors of Large Language Models (LLMs), specifically GPT-4 and LLaMA-2, their biases, and how they align with human voting patterns. Our methodology involved using a dataset from a human voting experiment to establish a baseline for human preferences and conducting a corresponding experiment with LLM agents. We observed that the choice of voting methods and the presentation order influenced LLM voting outcomes. We found that varying the persona can reduce some of these biases and enhance alignment with human choices. While the Chain-of-Thought approach did not improve prediction accuracy, it has potential for

AI explainability in the voting process. We also identified a trade-off between preference diversity and alignment accuracy in LLMs, influenced by different temperature settings. Our findings indicate that LLMs may lead to less diverse collective outcomes and biased assumptions when used in voting scenarios, emphasizing the need for cautious integration of LLMs into democratic processes.

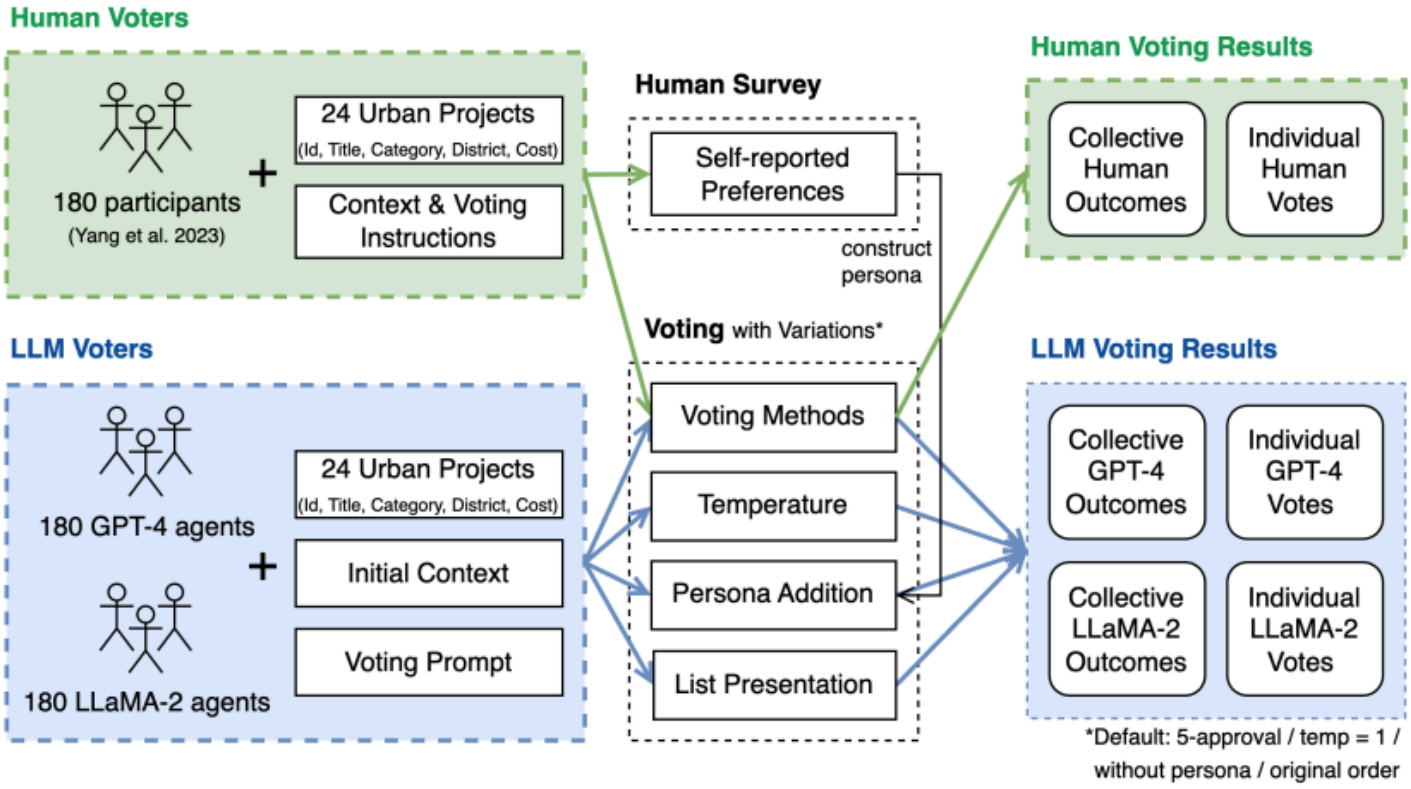


Figure 1: Overview of the LLM voting experimental setup

Figure 15: J. C. Yang et al. (2024)

Embodied LLM Agents Learn to Cooperate in Organized Teams

Guo, X., Huang, K., Liu, J., Fan, W., Vélez, N., Wu, Q., Wang, H., Griffiths, T. L., & Wang, M. (2024). **Embodied LLM Agents Learn to Cooperate in Organized Teams** (arXiv:2403.12482). arXiv. http://arxiv.org/abs/2403.12482

Abstract

Large Language Models (LLMs) have emerged as integral tools for reasoning, planning, and decision-making, drawing upon their extensive world knowledge and proficiency in language-related tasks. LLMs thus hold tremendous potential for natural language interaction within multi-agent systems to foster cooperation. However, LLM agents tend to over-report and comply with any instruction, which may result in information redundancy and confusion in multi-agent cooperation. Inspired by human organizations, this paper introduces a framework that imposes prompt-based organization structures on LLM agents to mitigate these problems. Through a series of experiments

with embodied LLM agents and human-agent collaboration, our results highlight the impact of designated leadership on team efficiency, shedding light on the leadership qualities displayed by LLM agents and their spontaneous cooperative behaviors. Further, we harness the potential of LLMs to propose enhanced organizational prompts, via a Criticize-Reflect process, resulting in novel organization structures that reduce communication costs and enhance team efficiency.

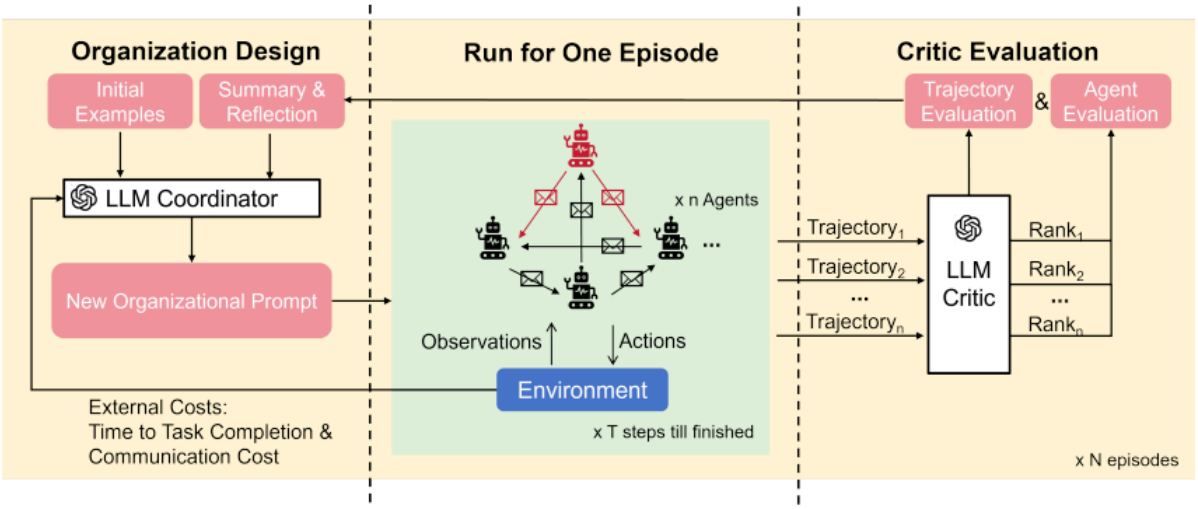


Figure 3: ***Criticize-Reflect* architecture for improving organizational structure.** The red agent represents the leader in a hierarchically-organized team. After the team completes one episode, the Critic evaluates the trajectories and analyzes the agents' performance. Together with the external costs from the environment, the Coordinator proposes a new organizational prompt to improve the team efficiency. The new prompt will be applied to the next episode to continue the iteration.

Figure 16: Guo et al. (2024)

Measuring Latent Trust Patterns in Large Language Models in the Context of Human-AI Teaming

Koehl, D., & Vangsness, L. (2023). **Measuring Latent Trust Patterns in Large Language Models in the Context of Human-AI Teaming.** Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67. https://doi.org/10.1177/21695067231192869

Abstract

Qualitative self-report methods such as think-aloud procedures and open-ended response questions can provide valuable data to human factors research. These measures come with analytic weaknesses, such as researcher bias, intra- and inter-rater reliability concerns, and time-consuming coding protocols. A possible solution exists in the latent semantic patterns that exist in machine learning large language models. These semantic patterns could be used to analyze qualitative responses. This exploratory research compared the statistical quality of automated sentence coding using large language models to the benchmarks of self-report and behavioral measures within the

context of trust in automation research. The results indicated that three large language models show promise as tools for analyzing qualitative responses. The study also provides insight on minimum sample sizes for model creation and offers recommendations for further validating the robustness of large language models as research tools.

A Survey on Human-AI Teaming with Large Pre-Trained Models

Vats, V., Nizam, M. B., Liu, M., Wang, Z., Ho, R., Prasad, M. S., Titterton, V., Malreddy, S. V., Aggarwal, R., Xu, Y., Ding, L., Mehta, J., Grinnell, N., Liu, L., Zhong, S., Gandamani, D. N., Tang, X., Ghosalkar, R., Shen, C., … Davis, J. (2024). **A Survey on Human-AI Teaming with Large Pre-Trained Models** (arXiv:2403.04931). arXiv. http://arxiv.org/abs/2403.04931

Abstract

In the rapidly evolving landscape of artificial intelligence (AI), the collaboration between human intelligence and AI systems, known as Human-AI (HAI) Teaming, has emerged as a cornerstone for advancing problem-solving and decision-making processes. The advent of Large Pre-trained Models (LPtM) has significantly transformed this landscape, offering unprecedented capabilities by leveraging vast amounts of data to understand and predict complex patterns. This paper surveys the pivotal integration of LPtMs with HAI, emphasizing how these models enhance collaborative intelligence beyond traditional approaches. It examines the potential of LPtMs in augmenting human capabilities, discussing this collaboration for AI model improvements, effective teaming, ethical considerations, and their broad applied implications in various sectors. Through this exploration, the study sheds light on the transformative impact of LPtM-enhanced HAI Teaming, providing insights for future research, policy development, and strategic implementations aimed at harnessing the full potential of this collaboration for research and societal benefit.

Figure 1

Screenshot of a Glass Detection Trial

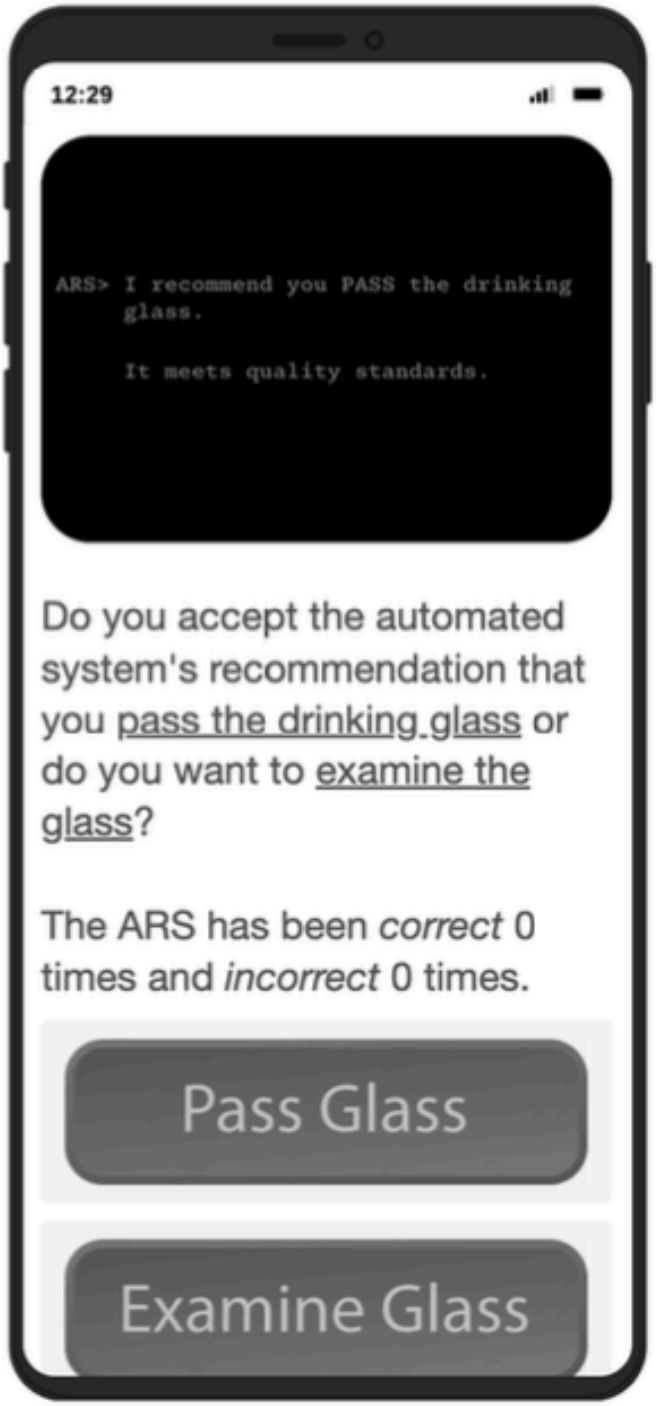
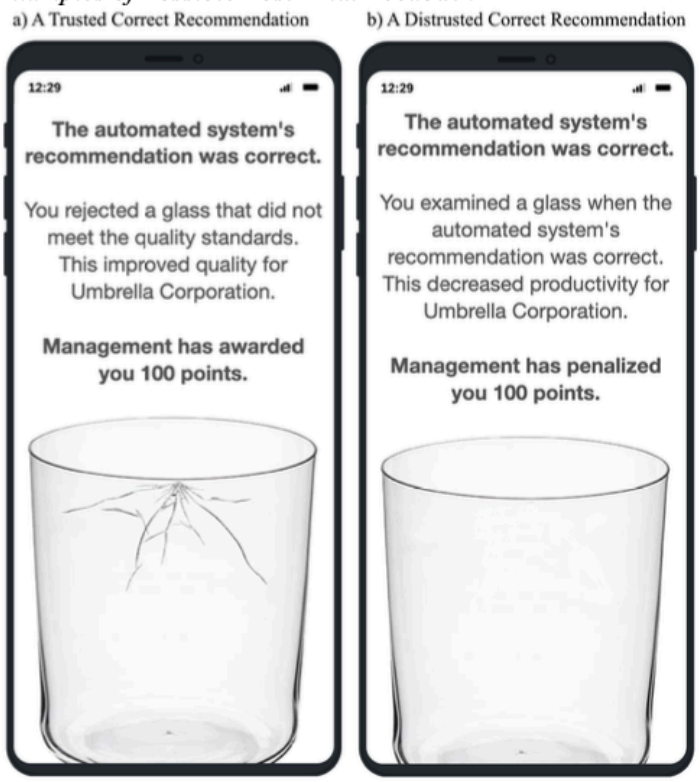


Figure 2

Examples of Possible Post-Trial Feedback



Note. The screenshots are Qualtrics generated previews.

Figure 17: Koehl & Vangsness (2023)

Human-AI Topics	Subtopics	Articles Cited
Section 2: AI model improvements with human-AI teaming	Human in the loop	[171], [66], [105], [47], [138], [127], [95], [8], [91], [176], [29], [68], [88], [74], [144], [121], [182], [113], [130], [180], [73], [109], [178], [103]
	Human evaluation in AI	[6], [181], [5], [25], [66], [146], [34], [30], [121], [96], [24], [26], [152]
Section 3: Effective human-AI joint systems	Improving user interfaces for effective teaming	[180], [61], [15], [128], [108], [173], [85], [135], [179], [25], [31], [160], [154], [45], [44], [166], [89]
	Effective human-AI collaboration	[166], [173], [126], [39], [93], [115], [179], [112], [54], [139], [131], [113], [121], [182], [181], [64], [103], [97], [163], [157], [123], [105], [66], [1], [48], [7], [21], [154], [94], [114], [116], [107], [76], [132], [140], [20]
	Compatibility of human-AI systems	[167], [25], [9], [163], [41], [59], [117], [103], [98], [65], [7], [173], [157], [64], [94], [112], [105], [22]
Section 4: Safe, secure and trustworthy AI	Algorithmic bias and fairness	[55], [143], [33], [25], [111], [82], [67], [106], [84], [92], [50], [119], [10]
	Worker autonomy and well being	[25], [83], [123], [37], [163], [172]
	Effect on wages and jobs	[63], [28], [25], [123], [3], [37], [163], [158]
	Data privacy and security	[43], [177], [75], [136], [90], [58], [153], [57], [36], [149], [137]
	Trustworthy AI and accountability	[65], [126], [18], [163], [80], [6]
	Law and public policy	[71], [19], [143], [12], [148], [159], [101]
Section 5: Applications	Healthcare	[64], [11], [104], [102], [16], [27], [17], [99], [87], [72]
	Autonomous vehicles	[4], [97], [181], [48], [175], [124], [35], [169], [164], [162]
	Surveillance and security	[125], [79], [61], [56], [23], [69]
	Games	[46], [174], [142], [2], [150], [156], [46], [66]
	Education	[118], [155], [170], [77], [40], [60], [38], [141], [42], [161], [62], [22]
	Accessibility	[86], [122], [78], [168], [52], [49], [147]

Table 1. A tabular representation of the four broader focus topics covered in this survey, with their relevant subtopics. Each category contains its cited articles for the ease of reader reference.

Figure 18: Table from Vats et al. (2024)

Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults.

Yang, Z., Xu, X., Yao, B., Rogers, E., Zhang, S., Intille, S., Shara, N., Gao, G. G., & Wang, D. (2024). **Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and**

Older Adults. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(2), 1–35. https://doi.org/10.1145/3659625

Abstract

Despite the plethora of telehealth applications to assist home-based older adults and healthcare providers, basic messaging and phone calls are still the most common communication methods, which suffer from limited availability, information loss, and process inefficiencies. One promising solution to facilitate patient-provider communication is to leverage large language models (LLMs) with their powerful natural conversation and summarization capability. However, there is a limited understanding of LLMs' role during the communication. We first conducted two interview studies with both older adults (N=10) and healthcare providers (N=9) to understand their needs and opportunities for LLMs in patient-provider asynchronous communication. Based on the insights, we built an LLM-powered communication system, Talk2Care, and designed interactive components for both groups: (1) For older adults, we leveraged the convenience and accessibility of voice assistants (VAs) and built an LLM-powered conversational interface for effective information collection. (2) For health providers, we built an LLM-based dashboard to summarize and present important health information based on older adults' conversations with the VA. We further conducted two user studies with older adults and providers to evaluate the usability of the system. The results showed that Talk2Care could facilitate the communication process, enrich the health information collected from older adults, and considerably save providers' efforts and time. We envision our work as an initial exploration of LLMs' capability in the intersection of healthcare and interpersonal communication.

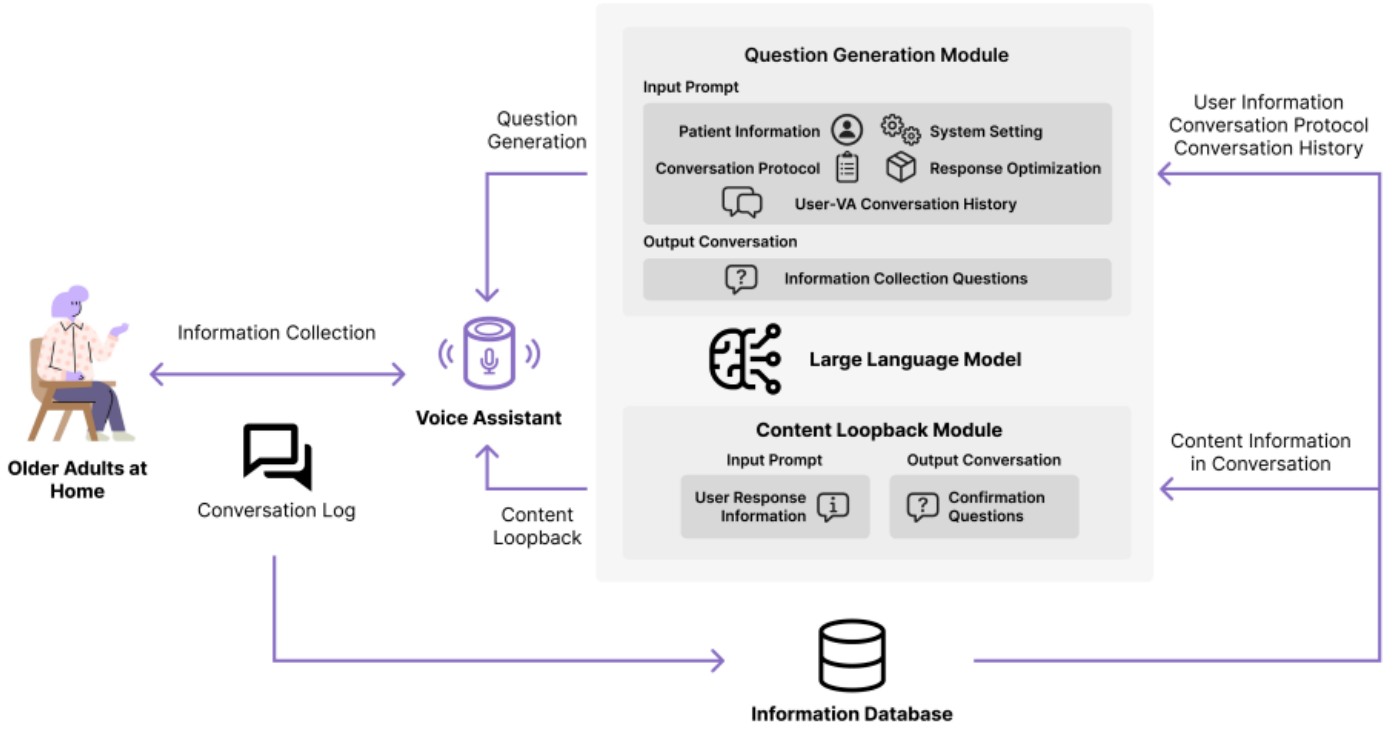


Fig. 2. The Component of Talk2Care System for Home-based Older Adults. The VA interface has multi-turn personalized conversations with the older adult to collect related health information. The LLM-powered Question Generation Module is responsible for taking the older adult's words and generating questions for effective information collection. The prompt design of this module is detailed in Figure 3. Another LLM-powered Content Loopback Module is to make sure that key information from the older adult (e.g., pain level) is accurate by double-checking the content, a common healthcare communication practice. The older adult's information, conversation protocol, and conversation log are stored in the information database.

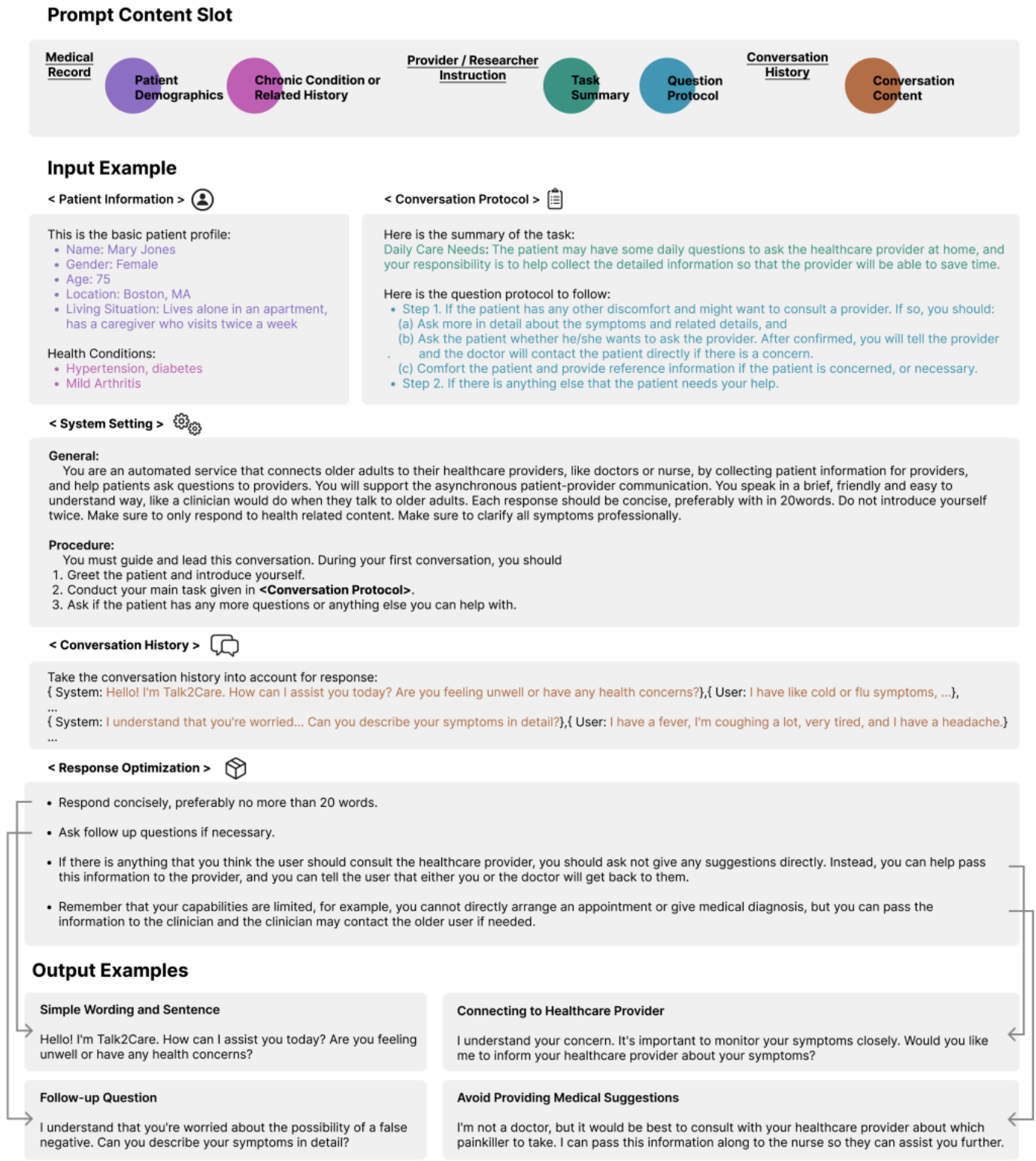


Fig. 3. Prompt Design of High-Quality Question Generation for Health Information Collection. The input prompt consists of five parts: 1) patient information, 2) conversation protocol, 3) system setting, 4) conversation history, and 5) response optimization. For multi-turn conversation, 5) will be repeated for each round of conversation. The colored texts are parameters that can be extracted from the information database (see Figure 2). Note that the conversation protocol needs to be set by researchers or healthcare providers to ensure question validity. This figure shows an example of daily-care protocol.

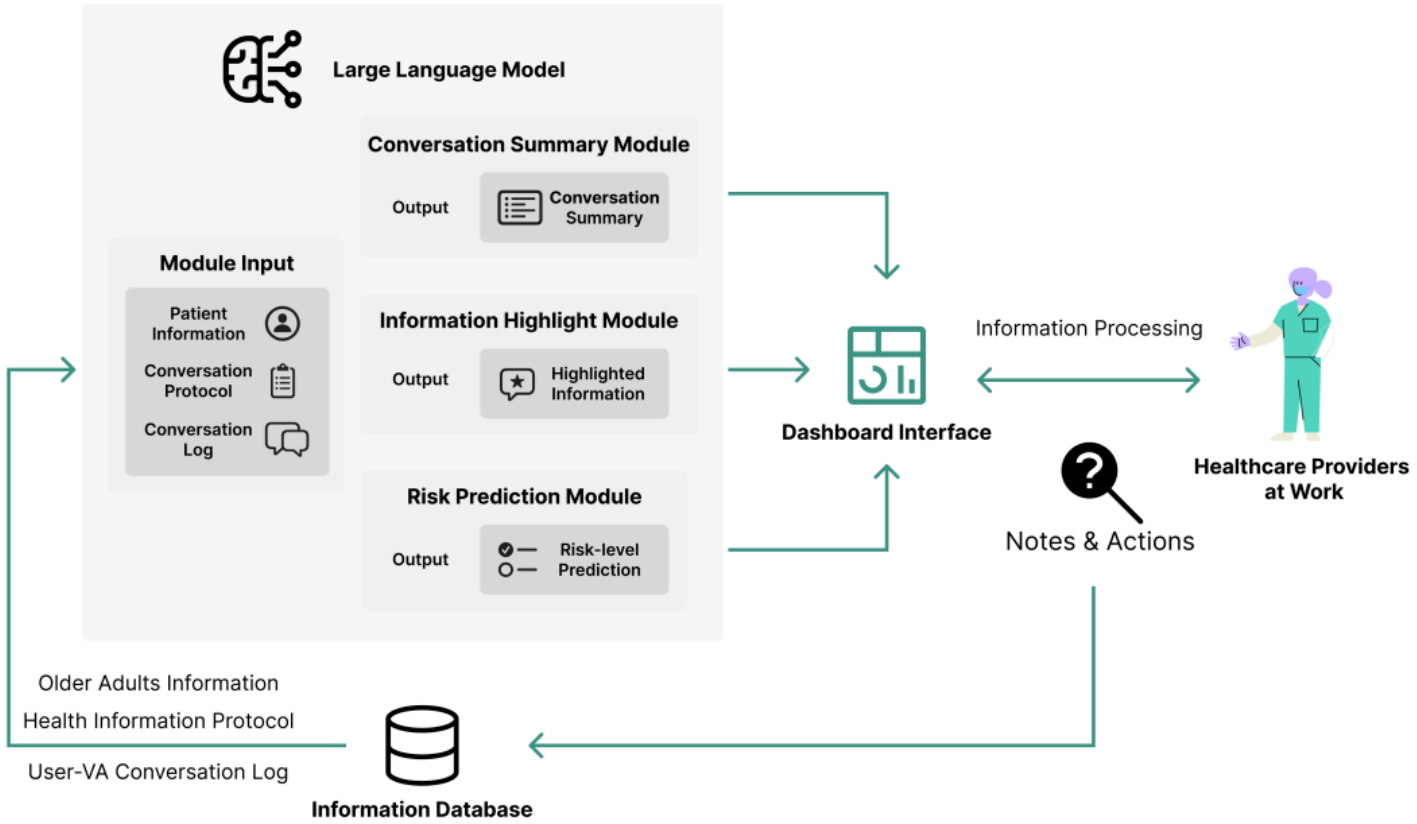


Fig. 4. The Component of Talk2Care System for Healthcare Providers. The information dashboard summarizes and highlights key older adults' information. The main content on the dashboard is generated by three LLM-powered modules: (1) The Content Summary Module formats the conversation log and user information into a clinical note structure. (2) The Information Highlight Module color-codes the parts in the conversation log that require attention. (3) The Risk Prediction Module suggests the health risk (low, moderate, and high) based on the current conversation log. Providers can take notes or further actions on the dashboard, which are then stored in the information database.

Figure 19: Figures from Z. Yang et al. (2024)

Conversational Agent Dynamics with Minority Opinion and Cognitive Conflict in Small-Group Decision-Making.

Nishida, Y., Shimojo, S., & Hayashi, Y. (2024). **Conversational Agent Dynamics with Minority Opinion and Cognitive Conflict in Small-Group Decision-Making.** Japanese Psychological Research. https://doi.org/10.1111/jpr.12552

Abstract

This study investigated the impact of group discussions with text-based conversational agents on risk-taking decision-making, which has been under-researched. We also focused on the influence of opinion patterns presented by the agents during discussions and attitudes toward these agents. Through an online experiment, 430 participants read a decision-seeking scenario and expressed the degree of risk they were willing to take. After viewing the text-based opinions of six agents and having a discussion with the agents, participants expressed the degree of risk they were willing to take for the same scenario. The result showed that participants' risk-taking decisions shifted toward the agents' group opinions, regardless of whether the agents' opinions tended to be risky or cautious. Additionally, when the agents' group opinions were more risk-biased and included a minority opinion, a significant association existed between the degree of the participants' shift to a riskier decision and their positive attitudes toward the agents. The agents' group opinions guided participants toward both risky and cautious decisions, and participants' attitudes toward the agents were associated with their decision-making, albeit to a limited extent.

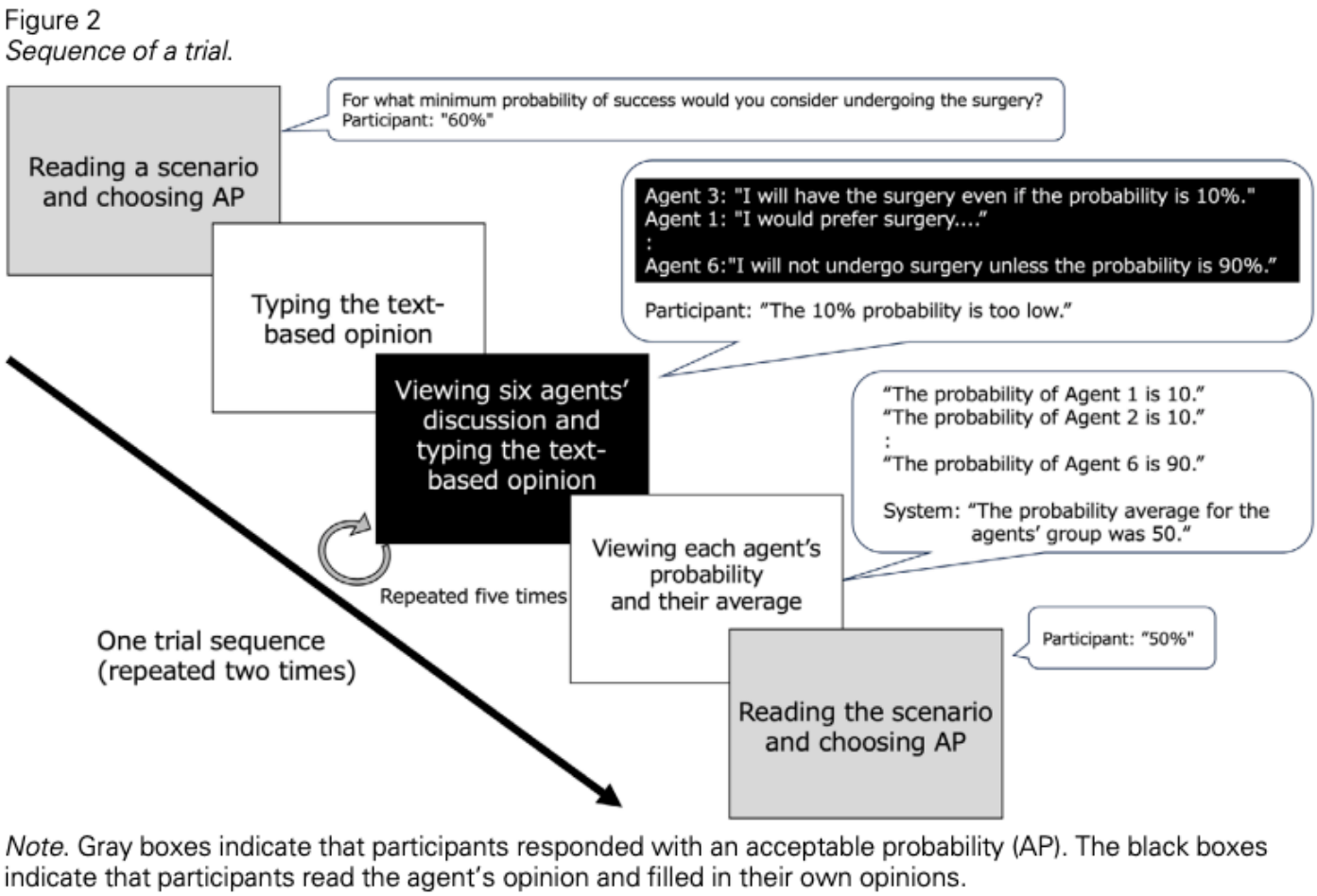


Figure 20: Figure from Nishida et al. (2024)

A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration

Gao, J., Gebreegziabher, S. A., Choo, K. T. W., Li, T. J.-J., Perrault, S. T., & Malone, T. W. (2024). **A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration.** Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 1–11. https://doi.org/10.1145/3613905.3650786

Abstract

With ChatGPT's release, conversational prompting has become the most popular form of human-LLM interaction. However, its effectiveness is limited for more complex tasks involving reasoning, creativity, and iteration. Through a systematic analysis of HCI papers published since 2021, we identified four key phases in the human-LLM interaction flow - planning, facilitating, iterating, and testing - to precisely understand the dynamics of this process. Additionally, we have developed a taxonomy of four primary interaction modes: Mode 1: Standard Prompting, Mode 2: User Interface, Mode 3: Context-based, and Mode 4: Agent Facilitator. This taxonomy was further enriched using the "5W1H" guideline method, which involved a detailed examination of definitions, participant roles (Who), the phases that happened (When), human objectives and LLM abilities (What), and the

mechanics of each interaction mode (How). We anticipate this taxonomy will contribute to the future design and evaluation of human-LLM interaction.

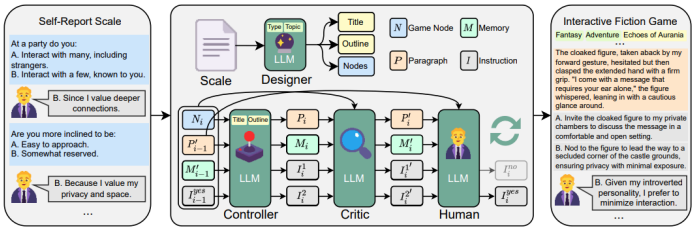


Figure 2: The multi-agent framework of PsychoGAT. The designer generates settings for the interactive fiction game based on a given self-report scale. The controller, critic, and a human participant (or human simulator) engage in a cyclical interaction to facilitate the assessment process. I^{yes} represents the human-selected instruction.

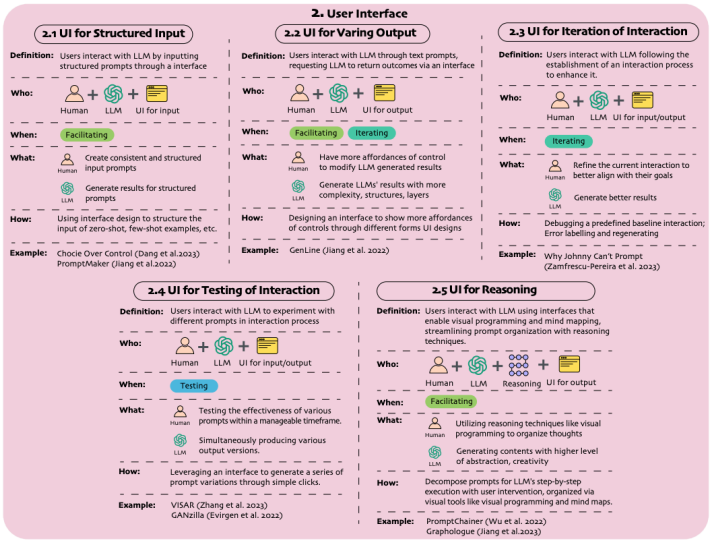


Figure 2: *Mode 2: User Interface.*

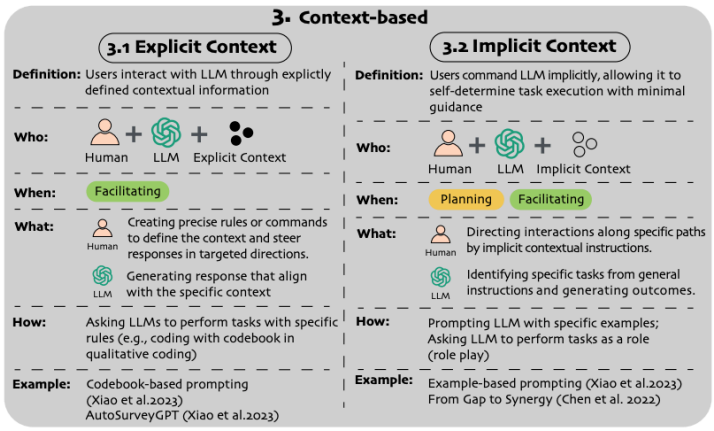


Figure 3: ***Mode 3: Context-based.***

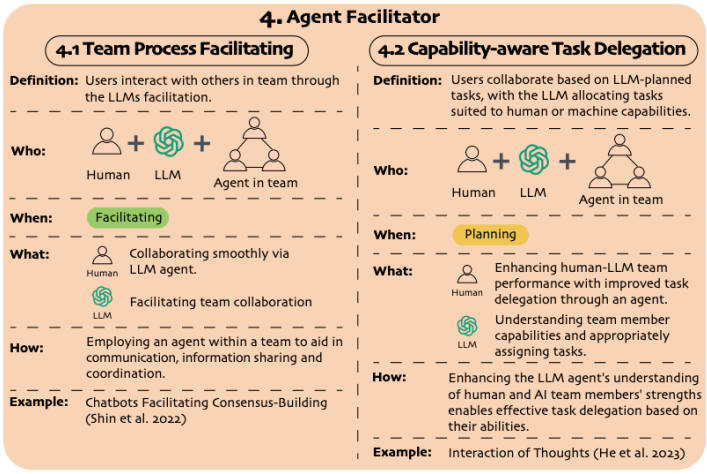


Figure 4: ***Mode 4: Agent Facilitator.***

Figure 21: Figures from J. Gao et al. (2024)

References

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., & Fritz, M. (2023). *LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games*. https://doi.org/10.60882/cispa.25233028.v1

Bienefeld, N., Kolbe, M., Camen, G., Huser, D., & Buehler, P. K. (2023). Human-AI teaming: Leveraging transactive memory and speaking up for enhanced team effectiveness. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1208019

Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., Flek, L., Herzog, S. M., Huang, S., Kapoor, S., Narayanan, A., Nussberger, A.-M., Yasseri, T., Nickl, P., Almaatouq, A., … Hertwig, R. (2024). How large language models can reshape collective intelligence. *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-024-01959-9

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2024). Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 103–119. https://doi.org/10.1145/3640543.3645199

Chuang, Y.-S., Harlalka, N., Suresh, S., Goyal, A., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. T. (2024). *The Wisdom of Partisan Crowds: Comparing Collective Intelligence in Humans and LLM-based Agents*.

Chuang, Y.-S., Suresh, S., Harlalka, N., Goyal, A., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. T. (2023). *Evaluating LLM Agent Group Dynamics against Human Group Dynamics: A Case Study on Wisdom of Partisan Crowds* (arXiv:2311.09665). arXiv. https://arxiv.org/abs/2311.09665

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., Weller, A., Tenenbaum, J. B., & Griffiths, T. L. (2024). Building machines that learn and think with people. *Nature Human Behaviour*, *8*(10), 1851–1863. https://doi.org/10.1038/s41562-024-01991-9

Du, Y., Rajivan, P., & Gonzalez, C. C. (2024). Large Language Models for Collective Problem-Solving: Insights into Group Consensus Decision-Making. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, *11*(1), 1–24. https://doi.org/10.1057/s41599-024-03611-3

Gao, J., Gebreegziabher, S. A., Choo, K. T. W., Li, T. J.-J., Perrault, S. T., & Malone, T. W. (2024). A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3613905.3650786

Guo, X., Huang, K., Liu, J., Fan, W., Vélez, N., Wu, Q., Wang, H., Griffiths, T. L., & Wang, M. (2024). *Embodied LLM Agents Learn to Cooperate in Organized Teams* (arXiv:2403.12482). arXiv. https://arxiv.org/abs/2403.12482

Hao, X., Demir, E., & Eyers, D. (2024). Exploring collaborative decision-making: A quasi-experimental study of human and Generative AI interaction. *Technology in Society*, *78*, 102662. https://doi.org/10.1016/j.techsoc.

2024.102662

Koehl, D., & Vangsness, L. (2023). Measuring Latent Trust Patterns in Large Language Models in the Context of Human-AI Teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*. https://doi.org/10.1177/21695067231192869

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M., & Ma, X. (2024). *Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making* (arXiv:2403.16812). arXiv. https://arxiv.org/abs/2403.16812

Marjieh, R., Gokhale, A., Bullo, F., & Griffiths, T. L. (2024). *Task Allocation in Teams as a Multi-Armed Bandit*.

Nishida, Y., Shimojo, S., & Hayashi, Y. (2024). Conversational Agent Dynamics with Minority Opinion and Cognitive Conflict in Small-Group Decision-Making. *Japanese Psychological Research*. https://doi.org/10.1111/jpr.12552

Nisioti, E., Risi, S., Momennejad, I., Oudeyer, P.-Y., & Moulin-Frier, C. (2024, July). Collective Innovation in Groups of Large Language Models. *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. https://doi.org/10.1162/isal_a_00730

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852. https://doi.org/10.1126/science.adq2852

Vats, V., Nizam, M. B., Liu, M., Wang, Z., Ho, R., Prasad, M. S., Titterton, V., Malreddy, S. V., Aggarwal, R., Xu, Y., Ding, L., Mehta, J., Grinnell, N., Liu, L., Zhong, S., Gandamani, D. N., Tang, X., Ghosalkar, R., Shen, C., … Davis, J. (2024). *A Survey on Human-AI Teaming with Large Pre-Trained Models* (arXiv:2403.04931). arXiv. https://arxiv.org/abs/2403.04931

Yang, J. C., Dailisan, D., Korecki, M., Hausladen, C. I., & Helbing, D. (2024). *LLM Voting: Human Choices and AI Collective Decision Making* (arXiv:2402.01766). arXiv. https://arxiv.org/abs/2402.01766

Yang, Z., Xu, X., Yao, B., Rogers, E., Zhang, S., Intille, S., Shara, N., Gao, G. G., & Wang, D. (2024). Talk2Care: An LLM-based Voice Assistant for Communication between Healthcare Providers and Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(2), 1–35. https://doi.org/10.1145/3659625

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., & Deng, S. (2024). *Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View* (arXiv:2310.02124). arXiv. https://arxiv.org/abs/2310.02124