Effect of Adaptive Communication Support on Human-AI Collaboration

Anonymous Author

Abstract

Effective human-AI collaboration requires agents to adopt their roles and levels of support based on human needs, task requirements, and complexity. Traditional human-AI teaming often relies on a pre-determined robot communication scheme, restricting teamwork adaptability in complex tasks. Leveraging strong communication capabilities of Large Language Models (LLMs), we propose a **H**uman-**R**obot **T**eaming Framework with **M**ulti-Modal **L**anguage feedback (HRT-ML), a framework designed to enhance human-robot interaction by adjusting the frequency and content of language-based feedback. HRT-ML framework includes two core modules: a *Coordinator* for high-level, low-frequency strategic guidance, and a *Manager* for task-specific, high-frequency instructions, enabling passive and active interactions with human teammates. To assess the impact of language feedback in collaborative scenarios, we conducted experiments in an enhanced Overcooked-AI game environment with varying levels of task complexity (easy, medium, hard) and feedback frequency (inactive, passive, active, superactive). Our results show that as task complexity increases relative to human capabilities, human teammates exhibited stronger preference towards robotic agents that can offer frequent, proactive support. However, when task complexities exceed the LLM's capacity, noisy and inaccurate feedback from superactive agents can instead hinder team performance, as it requires human teammates to increase their effort to interpret and respond to the large amount of communications, with limited performance return. Our results offers a general principle for robotic agents to dynamically adjust their levels and frequencies of communications to work seamlessly with human and achieve improved teaming performance.

Introduction

Human-robot collaboration has been extensively studied and applied across diverse scenarios, demonstrating strong potential towards enhanced efficiency and performance (Jahanmahin et al. 2022; Chuah and Yu 2021; Park et al. 2020; Gordon et al. 2020; Xiao et al. 2020; Liu et al. 2023b; 2024). As task complexity increases, agent adaptability becomes increasingly essential for seamless teamwork. Previous work has developed methods and tools for robot to adapt their actions based on inferred human objectives (Liu et al. 2024), trust level (Chen et al. 2020), and individual preferences (Bıyık et al. 2022). Recent work also begin to incorporate language-based feedback (Özdemir et al. 2022; Sharma et al. 2022) to enable more direct communications and lower user barriers. However, robot communications in these approaches primarily focused on relative simple commands such as "pick up the book" or "move left a little", which do not reflect the level of human-robot communications required in real-world applications.

Recent advancements in large language models (LLMs) have brought powerful reasoning (Zhang et al. 2023; 2024; Agashe, Fan, and Wang 2023; Guan et al. 2023), natural language understanding (Liu et al. 2023a; Wu et al.), contextual awareness (Deng et al. 2023), and generalization capabilities (Ge et al. 2024), enabling more advanced, prolonged communications (Bubeck et al. 2023; Ouyang et al. 2022; Hong et al. 2023). These methods has empowered agents to process ambiguous and complex instructions from human (Liu et al. 2023a), engage in more natural and dynamic conversations (Wu et al. ; Hou, Tamoto, and Miyashita 2024), and learn from a diverse set of inputs (Ge et al. 2024; Sun et al. 2024).

However, even in these LLM-enhanced communications, human teammates continue to play a predominant role in requesting specific tasks and providing suggestions during collaboration (Liu et al. 2023a). Schoenegger et al. (2024)'s study suggested that state-of-the-art LLMs can often match or surpass human performance in various domains. Based on these results, we hypothesize that allowing LLM agents to more proactively participate in (Tanneberg et al. 2024) or even initiate communications with human teammates can enhance teaming performance and efficiency.

To test this hypothesis and systematically evaluate the impact of different forms of language feedback provided by robots on collaboration efficiency and human satisfaction, in this study we develop **HRT-ML**, a human-robot teaming framework incorporating multi-modal language feedback to support dynamic, context-aware teaming styles. To enable the robot to provide effective language feedback, HRT-ML comprises two main modules: a *Coordinator*, which manages overall collaboration strategies and delivers low-frequency or passive instructions and feedback, and a *Manager*, which determines appropriate subtasks based on the coordinated plan at each stage, offering high-frequency in-

structions. Combining *Coordinator* and *Manager* allows the robotic agent to provide different forms of instruction at different frequencies. To investigate how the form and frequency of language feedback influence teaming performance, we performed user studies using four different agent active levels: Inactive, Passive, Active, and Superactive. We find that as the environment becomes more challenging, participants exhibit stronger preference for agents that provide active support, whereas in simpler tasks, frequent agent feedback is often perceived negatively, reducing overall performance and human satisfaction. These findings reveal the importance for the active level of robot language support to dynamically adapt based on task complexity and team capability. Inaccurate or overly-frequent feedback can decrease team efficiency, as participants must take extra time to understand and correct these suggestions.

In summary, our contributions are as follows:

- developed a human-robot teaming framework incorporating multi-modal language feedback to support dynamic, context-aware teaming styles.
- performed user studies to determine how forms and frequency of agent language feedback influence teaming performance
- discovered how the active level of robot language support should adapt based on task complexity and team capability, to best enhance teaming performance

To the best of the authors' knowledge, this is the first study that systematically explores the effect of different types of LLM-based multimodal language feedback on teaming performance across a wide range of task complexities.

Testbed: Overcooked-AI

To test the influence of language feedback on human-robot collaboration, we chose Overcooked-AI (Carroll et al. 2020), a platform designed to assess multi-agent coordination skills. In this game, players are motivated to collaborate actively to maximize their score by completing orders within a time limit. A score of 60 points was awarded when the correct soup is served. Partial points will be awarded to incomplete or incorrect soups based on the number of missing/incorrect components. To cook a soup, chefs need to finish specific subtasks in sequence according to the recipes (see Fig. 1A), and to finish subtasks, chefs need to move and interact with the environment. This process can be applied to any collaborative setting: first, reasoning through subtasks to achieve the overall goal, then selecting low-level atomic actions to complete each subtask.

Subtasks The subtasks in an overcooked environment for a single agent can be broken down into three main parts:

1. **Gathering Ingredients**: Chefs must first pick up the correct ingredients, such as onions or tomatoes, and place them into the cooking pot according to the recipe requirements.
2. **Cooking**: Once the ingredients are placed in the pot, the agent has to start cooking. A timer on the pot signals when the soup is ready to be served.

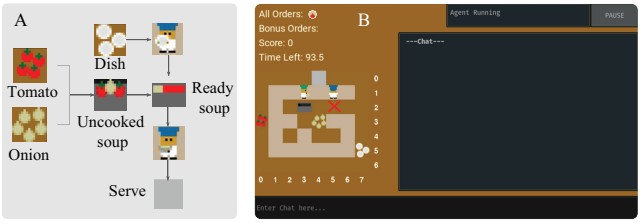

Figure 1: (A) Cooking process to complete an order. (B) The designed human-AI collaboration interface (left: game layout, right: communication panel). The red cross represent an example of the intermediate empty counter used to collaborate.

3. **Serving**: When the soup is ready, the chefs must collect and clean the dish from the dish dispenser, pour the soup into the dish, and deliver it to the serving location.

In multi-agent collaborative scenarios, each agent has a different path cost for completing each subtask, such as picking up an onion versus a tomato. Some subtasks may be unachievable for certain agents. Furthermore, by decomposing a subtask into multiple smaller subtasks that can be performed by multiple agents, the overall time cost can be reduced. For example, in Fig. 1, the green agent might place the onion at the red cross point (5,2), allowing the blue agent to pick it up from there and add it to the pot.

Atomic Action To finish a specific subtask, such as picking up an onion, the agent must execute a sequence of atomic actions, including movement commands like *up*, *down*, *left*, *right*, *stay*, and *interact* for picking up or placing objects.

Human-Robot Teaming Framework with Multi-Modal Language Feedback

To provide adaptive language feedback in human-robot collaboration, HRT-ML includes two core components: the *Coordinator* and the *Manager*. The *Coordinator* leads high-level strategy discussions with humans, generating a final coordination plan and offering low-frequency feedback. In contrast, the *Manager* handles detailed subtask allocation and provides high-frequency feedback to guide human players. Humans and other agents will receive the feedback and execute low-level actions.

Coordinator

The *Coordinator* is responsible for designing overall collaboration strategies and discussing them with the human partner. To support this, we employed a structured chain-of-thought approach to make the suggestions and discussions more effective. It begins by retrieving relevant information, including the layout, the overall collaboration goal, rules, and the agent's state formatted as a text description (Figure 2) using a prompt template, and then queries GPT-4o with it. It first analyzes the possible subtasks required to achieve the overall goal, then evaluates the difficulty of each subtask for each agent, and generates a plan that maximizes

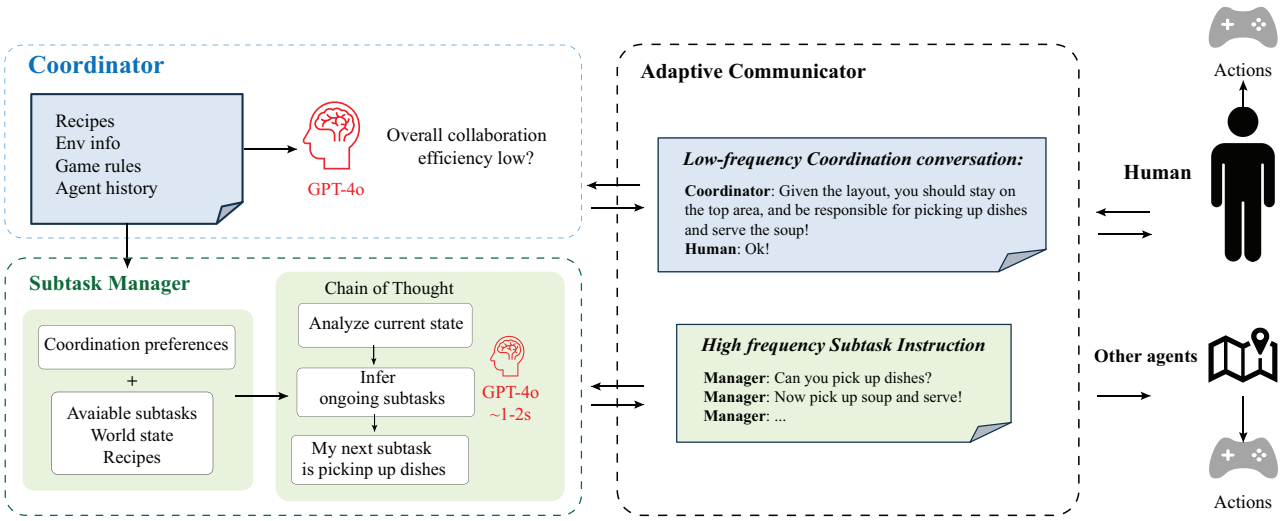

Figure 2: Flowing Human-Robot Teaming Framework (HRT-ML). It contains two modules: the *Coordinator* and the *Manager*. The *Coordinator* considers human preferences, formulates overall strategies, and provides low-frequency feedback to guide the collaboration. The *Manager* processes these strategies along with the state information to generate high-frequency subtask instructions for both human players and greedy planners, facilitating efficient task coordination and execution.

collaboration efficiency. For example, based on the environment (Fig. 1B), our *Coordinator* suggests: *Given the layout, the human (blue agent) should focus on picking up the tomato and serving the dish, while I will handle the onion and place the dish at location (5, 2) or (6, 2).* We note that the coordination querying is made conversational, allowing humans to continuously revise the plan through continuous discussion after reviewing the plan. During these coordination discussions, the game remains paused until the human chooses to end the conversation. The conversation can be initiated proactively by the *Coordinator* at a low frequency or by the human. When a human request is made, we incorporate their suggestions into the coordination plan prompt for initially querying GPT-4, explicitly asking it to evaluate the feasibility of these preferences and propose a corresponding coordination plan.

Manager The Manager assigns subtasks at each stage, providing high-frequency support to guide the human toward the final goal. First, the Manager uses a subtask filter to identify feasible subtasks based on the agent's and environment's current states, selecting from the potential subtask list from the overall plan created by the Coordinator. Next, it converts the selected subtasks, coordination plan, current agent states, and environment state into a language-based prompt, querying GPT-4 to generate the next subtask. This query is generated immediately upon completing the current subtask, with an average latency of $\sim$ 1.2 seconds. The determined subtask in every query will then be converted into language feedback and sent to humans.

Greedy Planner Once given the target subtask, the human determines the atomic actions needed to complete it. For collaborative tasks with an autonomous agent, a greedy planner using Depth First Search (DFS) finds the optimal path with the lowest action cost to finish the subtask.

Multi-Modal feedbacks

Building on the proposed HRT-ML framework, we introduce agents that provide four different language feedbacks described as follows: Inactive Feedback, Passive Feedback, Active Feedback, and Superactive Feedback.

- *Inactive Feedback agent (IFA)*: The IFA collaborates with humans without language communication and coordination. Only the Manager generates target subtasks for the greedy planner.
- *Passive Feedback agent (PFA)*: The PFA starts to provide passive feedback only when a human requests. The human player takes the role of the leader, while the agent acts as the follower, passively responding to human requests. If there are no specific human commands, PFA will behave like an IFA.
- *Active Feedback agent (AFA)*: The AFA collaborates with humans as peers. In this mode, both humans and agents can reach out to give language feedback. For user study, we prompt the GPT-4o coordinator in low frequency ($\sim$ 20s) to analyze human conversation history and suggest coordination strategies.
- *Superactive Feedback agent (SFA)*: The SFA treats humans as novices, acting as supervisors by providing frequent, continuous guidance on every subtask.

Data collection

In this section, we aim to explore the agent's ability to provide language feedback to help improve human satisfaction

and teaming efficiency. Based on the proposed HRT-ML, we consider four types of language feedback and three different layouts: easy, medium, and difficult, with increasing task difficulty and map complexity (Fig. 3). Human participants will play on each layout paired with an autonomous agent with varying levels of language feedback. Further details will be provided in the following section.

Layouts with different complexities

To test the performance of the four agents and the influence of language feedback, we implemented four overcooked maps (Fig. 3). One of the maps is an introductory map, designed for participants to get familiar with the game and operation. The other three maps have varying levels of difficulty, which we refer to as easy, medium, and hard. The more challenging maps have more "dead-ends", requiring humans and agents to have a better coordination strategy to maintain the team efficiency. Furthermore, the hard map introduces complexity in task orders by requiring multiple ingredients, which demands that humans and the agent reason about orders containing both tomatoes and onions.

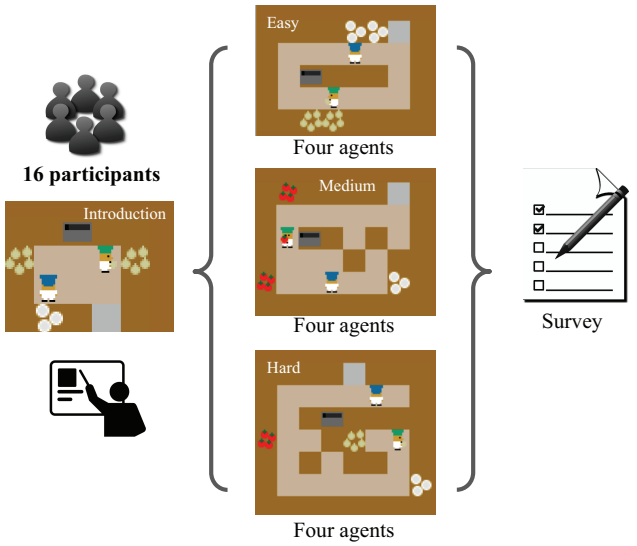

Figure 3: Overview of the human study procedure involving 16 participants, beginning with an introduction map to learn game mechanics and agent behavior followed by trials on Easy, Medium, and Hard Layout with four different agents. Post-session surveys were conducted to collect data on participant satisfaction, engagement, trust, and feedback.

Participants

In this study, we recruited 16 participants (9 male and 7 female) aged between 23 and 33 to evaluate the performance of collaborative agents. We selected participants with varying levels of familiarity with digital agents in game environments: 10 participants with self-reported video game time between 1-10 hours per week, 5 participants reported 10-20 hours per week, and 1 reported 20-30 hours per week. Additionally, 13 of the 16 participants reported prior familiar with Large Language Models (LLMs) and embodied language agents, while 3 of the 16 participants reported no prior knowledge or experiences with language agents.

Procedures

Participants were first asked to complete a consent form per Institutional Review Board (IRB) protocol. Subsequently, prior to the formal trials they were given an opportunity to play with the four agents in an introduction map (Fig. 3) to explore various language feedback styles. During this introduction phase participants were guided through system operations, game rules, and agent functionalities. After the introduction phase, participants proceeded to independently collaborate with each of the four agent types, on easy, medium, and hard layouts (Fig. 3), to complete the maximum score within the given time limit (60 s for each layout). For each participant, the scenario (*i.e.,* agent type $\times$ layout difficulty) was set to show up in a randomized order. We collected a total of 192 experiment trials, 12 trials per participant.

We collected game scores and step-by-step action logs for each experiment shown in (Fig. 3). After completing the teaming scenario, participants were also asked to fill out a survey rating their satisfaction, engagement, and trust level for each experiment, on a seven-point Likert scale. Participants were also asked to specify their preferred language feedback level for each layout.

Additionally, participants were asked to provide improvement suggestions on the language feedback provided by agents (*e.g., "If you were to play with this agent in an Overcooked game competition, what changes or improvements would you suggest for the agent's feedback?"*), and state the reason for their satisfaction ratings. For more details, the full questionnaire is available at [1].

Results and Discussion

The purpose of our data collection and analysis was to test the following hypotheses: *An increased level of language support will result in an increase in the perceived level of trustworthiness and intelligence of the agent, and improve overall team performance.*

Surprisingly, our data suggested while language feedback could indeed facilitate human trust, perceived agent intelligence, and team efficiency, the desired level of language support exhibited a different relationship than hypothesized. We report our findings in the subsequent sections.

Language feedback builds human trust and perceived intelligence

Trust is a key factor in human-AI collaboration, shaping human experiences and significantly influencing long-term collaboration efficiency (Chen et al. 2020). In this section, we report the perceived levels of trust and intelligence after participants teamed with four types of agents across three

[1]Link to the questionnaire:
https://docs.google.com/forms/d/1xSRPH1mkrZxuKXQ0xvjl9Cv4yi0l-aBblpwyG5MbX34/viewform?edit_requested=true

layouts. Both intelligence and trust levels were measured using a 7-point Likert scale, ranging from "Very Untrustworthy" (or "Very Unintelligent") to "Very Trustworthy" (or "Very Intelligent"). We found that as the agent's support level increased from Inactive to Superactive, intelligence rating increased monotonically (Fig. 4). A similar trend was observed in the trust ratings. This suggested that as hypothesized, active communication can facilitate building trust and preceived intelligence. Another interesting observation was that, as the agents become more active, the standard deviation for trust ratings and intelligent ratings also increased. This suggested that both trust and intelligence levels also become more influenced by individual human preferences when the agent takes on a more active role in assisting.

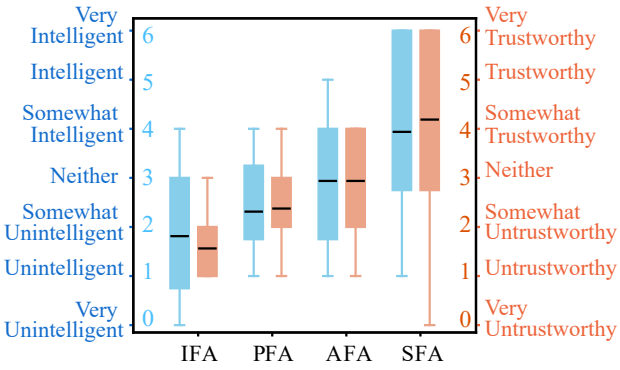

Figure 4: Human perceived agent intelligence level (blue bar) and trust level (red bar) represented on a seven-point Likert scale, ranging from "Very Unintelligent/Untrustworthy" to "Very Intelligent/Trustworthy".

Appropriate Language feedback improves collaboration efficiency

We used the game score to evaluate the team perfromance and collaboration efficiency between humans and agents. Overall, the team scored more points in easy layout with all agent types (53.6 in average) as compared to the medium (30.6 in average) and hard (27.3 in average) layouts. This is not surprising, as the easy layout has simpler maps and fewer subtasks, requiring minimal coordination to complete. In contrast, medium and hard layouts introduced more complex subtasks and dependencies, requiring greater coordination and team effort, which could increase the chance of errors during the task and result in lower scores.

We found that the team performance with "active" agents (PFA, AFA, and SFA), which engaged in language feedback and coordination with human, achieved higher scores on almost all difficulty levels than the "inactive" agent (IFA), which did not engage in any communication (Fig. 5). In addition, as the difficulty level increases from easy to hard, the team performance with the passive agent (PFA) decreased significantly, from close to 40 points to around 10 points (approximately 75% of performance drop). This result suggested that as hypothesized, language feedback can play a significant role in human-robot teaming, especially in complex tasks.

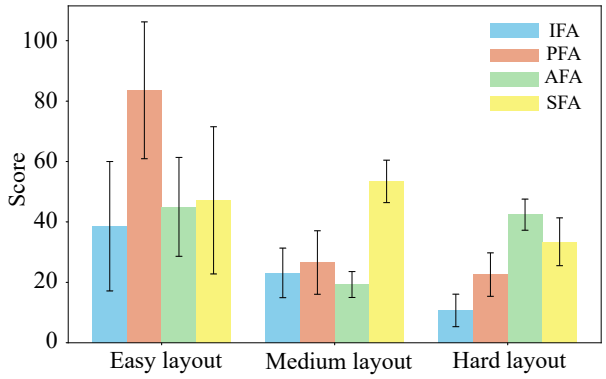

Figure 5: Game scores of all participants paired with different agents across various layouts. The score distribution of each agent type, IFA, PFA, AFA, SFA, is represented by the blue, red, green, and yellow boxes, respectively.

However, the team performance was not always better with more active agents. In the easy layout, team with PFA performed best, scoring 83.6 in average, significantly higher than IFA, AFA, and SFA (Fig. 5). It was expected that PFA performed better than IFA in the easy task, as it provided effective support with minimal interference, enabling participants to benefit from its assistance as needed and allowing humans to take the lead. This was also supported by user feedback from the survey – *"Robot listening to the user and following the commands would help the game better. In terms of intelligence, passive robots work better but still lack user assistance"*. Unexpectedly, while AFA and SFA offered more active feedback, they achieved lower scores than the PFA (Fig. 5), suggesting that constant communication was less effective for simple tasks, and may even distract humans. As the complexity increases to the medium level, the team with SFA exhibited a huge increase in score (Fig. 5), exceeding the performance of PFA, implying that as task complexity increased, the frequent support provided by SFA became more valuable. Similarly, in the hard layout, the SFA and AFA demonstrated superior performance as compared to IFA and PFA (Fig. 5), underscoring the value of active guidance and high-frequency support in facilitating collaboration and helping improve task execution in complex tasks.

Interestingly, despite SFA's higher frequency of active support compared to AFA, overall performance still decreased due to the high complexity of the hard layout compared to AFA. Participant feedback highlighted this challenge, *e.g.,* "*the robot/agent is not as smart as me,*" and, "*I have to give it a long instruction set, and I am more intelligent than it in the hard layout.*" One interpretation of these responses was that participants found the agent's support insufficient for complex tasks. As a result, instead of reducing cognitive load, the frequent suggestions from the agent

required human to carefully think about responses, which ultimately increased cognitive demands and decreased team efficiency. Another interpretation is that psychologically, humans may have the preference to demonstrate and maintain intellectual superiority in challenging tasks when teaming with AI agents. As a result, how AI agents communicate suggestions may greatly influence humans' acceptance rate and team efficiency.

Overall, our results revealed that, the language feedback provided by LLMs can boost human-robot collaboration efficiency and increase human satisfaction. However, the proactiveness and frequency of the language feedback should be provided based on the task complexity and the capabilities of LLMs.

Humans don't always prefer the best-performing agents

Interestingly, our data suggested that human does not always prefer the agent type that helped achieve the highest game scores. For example, even though the IFA achieved the lowest scores on easy layout (Fig. 5), over 50% of participants reported IFA as their preferred agent among the four types (Fig. 6A). Similarly, even though the SFA outperformed PFA by almost two folds in terms of team score (Fig. 5), 56% of participants selected the PFA as their preferred agent, and 0% of them preferred SFA (Fig. 6A).

We believe this preference shift can be explained by the flow theory (Csikszentmihalyi 2000; Chen et al. 2024), which states that people feel most engaged when task complexity aligns with their skill level, and robots can help achieve this balance by adjusting their level of support. In the easy layout, the task was not significantly beyond human's skill level, and the need for additional help to maintain engagement was low. Therefore, while the active agents can provide help, this help did not significantly influence human's engage level and satisfaction rate (Fig. 6B). As a result, the non-active agent, which requires the lowest level of cognitive load ("cost") was preferred. As the difficulty level increases, however, the gap between task complexity and human skill level increases, resulting a disruption to the task-capability balance and reduced engagement. Meanwhile, the additional feedback and support from the more active agents, such as assigning subtasks (*e.g.,* "pick up the onion from (x, x)."), can reduce the coordination and planning effort on the human, and restore the task-capability balance and enhance human's feel of engagement and joy. At this point, the cost of communication became negligible as compared to the need for sense of achievement, making more active agents more desirable in these challenging scenarios (Fig. 6B).

To test this, we recorded participants' engagement during each game trial. As shown in Fig. 7, satisfaction levels increased with higher engagement, with a correlation test yielding a p-value of 0.00275 and a correlation coefficient of 0.93, indicating a strong relationship. This strong correlation between satisfaction and engagement indicates that engagement is the cause of different satisfaction levels reported by participants. This aligned with our theory, emphasizing the need to align agent support and feedback frequency based on task complexity relative to human capabilities.

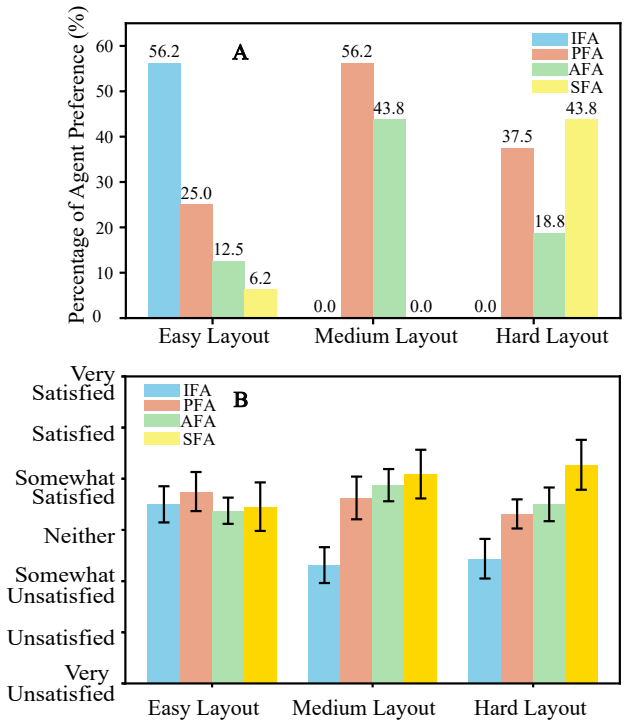

Figure 6: Participant preferences and satisfaction levels for different agent types across layouts. (A) Bar chart showing the percentage of agent preference by participants for different layouts. (B) Satisfaction ratings for each agent type in different layouts, represented on a seven-point Likert scale, ranging from "Very Unsatisfied" to "Very Satisfied," with error bars indicating the standard error of the mean (SEM) in responses.

Adaptively assigning subtasks and providing language feedback

Our findings revealed the importance for LLM agents to adapt their language feedback by considering the relationship between task complexity, T_h, human capability, C_h, and the LLM's capability, C_l. Below we propose a simple adaptation strategy for LLM agents to select their support level and language feedback frequency:

- $C_h > T$ and $C_l < T$: Here human capability surpasses the task's complexity, and the LLM capability is not sufficient to address the challenging tasks without human guidance. Based on our results, a passive (PFA) to relative infrequent (AFA) feedback style would allow the agent to provide sufficient support to improve team performance and request human help when needed, while keeping communication frequency to a minimal to avoid overhead on the human side. This way, human teammates with higher capability level could guide the agent on high-level coordination strategies and specific subtask execu-

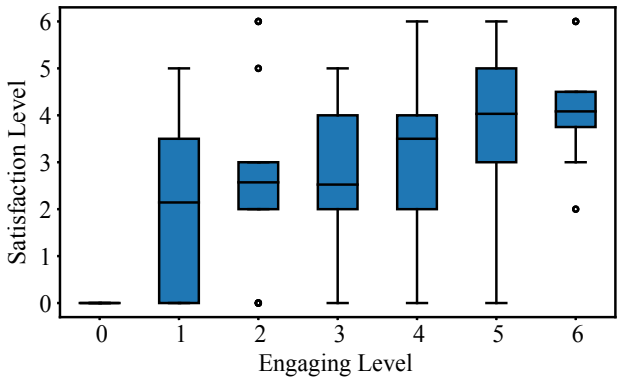

Figure 7: Illustration of the relationship and variability between satisfaction level and the engagement of participants for all experimental trials. The box plot shows how satisfaction levels correspond to participants with an engagement level from 0 (Very distracted) to 6 (Very engaged).

tion, keeping them engaged.

- $C_h < T$ and $C_l > T$: Here task complexity exceeds human capability, while LLM is fully capable of executing the task. In this case, extra active agent feedback and support (SFA) are crucial for maintaining team performance, and help reduce the gap between human capability and task complexity.
- $C_h < T$ and $C_l < T$: Here task complexity exceeds human capability. However, LLM capability is also not sufficient to address the challenging task neither. In this case, language feedback from LLM is often not useful in resolving the challenges that human is struggling with, and can even be misleading. High-frequency communications from LLM in this scenario would require additional effort from human to respond, potentially further heightening the anxiety that human is already experiencing (Lenzner et al. 2010), decreasing team performance and human engagement. Our results suggested that a more passive (PFA) or relative infrequent (AFA) feedback style would result in better teaming performance in this case.
- $C_h > T$ and $C_l > T$: Here both human and LLM capabilities surpass the task's complexity. The active feedback style (AFA) could allow the human and agents to communicate their needs at a comfortable pace, and improve collaboration efficiency.

Conclusion

In this work, we introduced HRT-ML, a flexible human-robot teaming framework designed to provide adaptive communication feedback to humans at varying levels and frequencies. The HRT-ML framework comprises two core modules: a *Coordinator* for high-level, low-frequency strategic guidance and a *Manager* for task-specific, high-frequency instructions, allowing collaborating with humans across four distinct feedback styles: Inactive, Passive, Active, and Superactive.

Our user study results demonstrated that language-based feedback from LLMs can significantly enhance collaboration performance and foster human trust, and that as task complexity increases, more frequent, proactive support is desired. However, our study also revealed that it is critical for the agent to select their language feedback frequency based on task complexity, human capability, and agent capability. Overly frequent feedback in simple tasks or from less capable agents does not effectively increase team performance and satisfaction, and could even increase effort and reduce human engagement. Based on these findings, we proposed a simple principle that allows agents to adapt their language feedback style according to perceived task challenges, human capabilities, and LLM capabilities.

Limitations and future work

In this work, we focused on investigating the effect of agent feedback frequency on team performance. As such, each agent was set to a constant active level, and cannot dynamically adjust their level of support and language feedback throughout the task. In real-world scenarios, task complexity and human skill levels for different sub task can vary dynamically, requiring agents to adjust their communications and behaviors accordingly. Future work should explore methods for adaptive agent to estimate human cognitive load, capabilities, and engagement, and adjust LLM feedback in real time to enable better team performance and adaptability. The results from our study provide the basis for designing and implementing such real-time feedback adjustments. Going forward, these adaptive response and feedback capabilities can empower future LLM agents to flexibly support human needs in a wide variety of task complexities, fostering true partnerships and enhancing teamwork outcomes.

Acknowledgment

References

Agashe, S.; Fan, Y.; and Wang, X. E. 2023. Evaluating multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*.

Bıyık, E.; Losey, D. P.; Palan, M.; Landolfi, N. C.; Shevchuk, G.; and Sadigh, D. 2022. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research* 41(1):45–67.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Carroll, M.; Shah, R.; Ho, M. K.; Griffiths, T. L.; Seshia, S. A.; Abbeel, P.; and Dragan, A. 2020. On the utility of learning about humans for human-ai coordination.

Chen, M.; Nikolaidis, S.; Soh, H.; Hsu, D.; and Srinivasa, S. 2020. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)* 9(2):1–23.

Chen, H.; Alghowinem, S.; Breazeal, C.; and Park, H. W. 2024. Integrating flow theory and adaptive robot roles: A conceptual model of dynamic robot role adaptation for the enhanced flow experience in long-term multi-person human-robot interactions. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 116–126.

Chuah, S. H.-W., and Yu, J. 2021. The future of service: The power of emotion in human-robot interaction. *Journal of Retailing and Consumer Services* 61:102551.

Csikszentmihalyi, M. 2000. *Beyond boredom and anxiety*. Jossey-bass.

Deng, Y.; Lei, W.; Huang, M.; and Chua, T.-S. 2023. Rethinking conversational agents in the era of llms: Proactivity, non-collaborativity, and beyond. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 298–301.

Ge, Y.; Hua, W.; Mei, K.; Tan, J.; Xu, S.; Li, Z.; Zhang, Y.; et al. 2024. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems* 36.

Gordon, E. K.; Meng, X.; Bhattacharjee, T.; Barnes, M.; and Srinivasa, S. S. 2020. Adaptive robot-assisted feeding: An online learning framework for acquiring previously unseen food items. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9659–9666. IEEE.

Guan, C.; Zhang, L.; Fan, C.; Li, Y.; Chen, F.; Li, L.; Tian, Y.; Yuan, L.; and Yu, Y. 2023. Efficient human-ai coordination via preparatory language-based convention. *arXiv preprint arXiv:2311.00416*.

Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Hou, Y.; Tamoto, H.; and Miyashita, H. 2024. " my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.

Jahanmahin, R.; Masoud, S.; Rickli, J.; and Djuric, A. 2022. Human-robot interactions in manufacturing: A survey of human behavior modeling. *Robotics and Computer-Integrated Manufacturing* 78:102404.

Lenzner, T.; Kaczmirek, L.; Lenzner, A.; and xxxx, x. 2010. Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied cognitive psychology* 24(7):1003–1020.

Liu, J.; Yu, C.; Gao, J.; Xie, Y.; Liao, Q.; Wu, Y.; and Wang, Y. 2023a. Llm-powered hierarchical language agent for real-time human-ai coordination. *arXiv preprint arXiv:2312.15224*.

Liu, S.; Wilson, C. G.; Krishnamachari, B.; and Qian, F. 2023b. Understanding human dynamic sampling objectives to enable robot-assisted scientific decision making. *ACM Transactions on Human-Robot Interaction*.

Liu, S.; Wilson, C. G.; Lee, Z. I.; and Qian, F. 2024. Modelling experts' sampling strategy to balance multiple objectives during scientific explorations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, 452–461. New York, NY, USA: Association for Computing Machinery.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744.

Özdemir, O.; Kerzel, M.; Weber, C.; Lee, J. H.; and Wermter, S. 2022. Language-model-based paired variational autoencoders for robotic language learning. *IEEE Transactions on Cognitive and Developmental Systems* 15(4):1812–1824.

Park, D.; Hoshi, Y.; Mahajan, H. P.; Kim, H. K.; Erickson, Z.; Rogers, W. A.; and Kemp, C. C. 2020. Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned. *Robotics and Autonomous Systems* 124:103344.

Schoenegger, P.; Park, P. S.; Karger, E.; Trott, S.; and Tetlock, P. E. 2024. Ai-augmented predictions: Llm assistants improve human forecasting accuracy. *arXiv preprint arXiv:2402.07862*.

Sharma, P.; Sundaralingam, B.; Blukis, V.; Paxton, C.; Hermans, T.; Torralba, A.; Andreas, J.; and Fox, D. 2022. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*.

Sun, Y.; Salami Pargoo, N.; Jin, P.; and Ortiz, J. 2024. Optimizing autonomous driving for safety: A human-centric approach with llm-enhanced rlhf. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 76–80.

Tanneberg, D.; Ocker, F.; Hasler, S.; Deigmoeller, J.; Belardinelli, A.; Wang, C.; Wersing, H.; Sendhoff, B.; and Gienger, M. 2024. To help or not to help: Llm-based attentive

support for human-robot group interactions. *arXiv preprint arXiv:2403.12533*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Xiao, J.; Wang, P.; Lu, H.; and Zhang, H. 2020. A three-dimensional mapping and virtual reality-based human–robot interaction for collaborative space exploration. *International Journal of Advanced Robotic Systems* 17(3):1729881420925293.

Zhang, C.; Yang, K.; Hu, S.; Wang, Z.; Li, G.; Sun, Y.; Zhang, C.; Zhang, Z.; Liu, A.; Zhu, S.-C.; et al. 2023. Proagent: Building proactive cooperative ai with large language models. *arXiv preprint arXiv:2308.11339*.

Zhang, H.; Wang, Z.; Lyu, Q.; Zhang, Z.; Chen, S.; Shu, T.; Du, Y.; and Gan, C. 2024. Combo: Compositional world models for embodied multi-agent cooperation. *arXiv preprint arXiv:2404.10775*.