Task Proficiency-Aware Dialogue Analysis in a Real-Time Cooking Game Environment

Kaito Nakae **Michimasa Inaba**
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo, Japan
n2530097@edu.cc.uec.ac.jp, m-inaba@uec.ac.jp

Abstract

Real-time collaborative dialogue tasks require dynamic, instantaneous decision-making and seamless coordination between participants, yet most existing studies on collaborative dialogues primarily focus on turn-based textual environments. This study addresses the critical gap in understanding human-human interaction patterns within dynamic, real-time collaborative scenarios. We present a novel dataset collected from a real-time collaborative cooking game environment inspired by the popular game "Overcooked." Our dataset comprises detailed annotations of participants' task proficiency levels, game scores, game action logs, and transcribed voice dialogues annotated with dialogue act tags. Participants exhibited a broad range of gaming experience, from highly proficient players to those with minimal exposure to gaming controls. Through comprehensive analysis, we explore how individual differences in task proficiency influence dialogue patterns and collaborative outcomes. Our findings reveal key dialogue acts and adaptive communication strategies crucial for successful real-time collaboration. Furthermore, this study provides valuable insights into designing adaptive dialogue systems capable of dynamically adjusting interaction strategies based on user proficiency, paving the way for more effective human-AI collaborative systems. The dataset introduced in this study is publicly available at: https://github.com/UEC-InabaLab/OverCookedChat.

1 Introduction

Human-AI collaboration is emerging as a critical research area within the field of AI (Wang et al., 2020; Lai et al., 2021; Vössing et al., 2022). Moving beyond the traditional scope of AI as mere assistance or automation tools, this paradigm aims to tackle more sophisticated problems by synergistically leveraging the complementary strengths of both humans and AI (Vats et al., 2024). Reflecting this trend, there is a surge of research focusing on collaborative frameworks where AI is treated not just as a tool but as a partner, enabling humans and AI to work together effectively towards shared goals (Charakorn et al., 2020; Puig et al., 2021; Sarkar et al., 2022; Zhang et al., 2024; Wang et al., 2024).

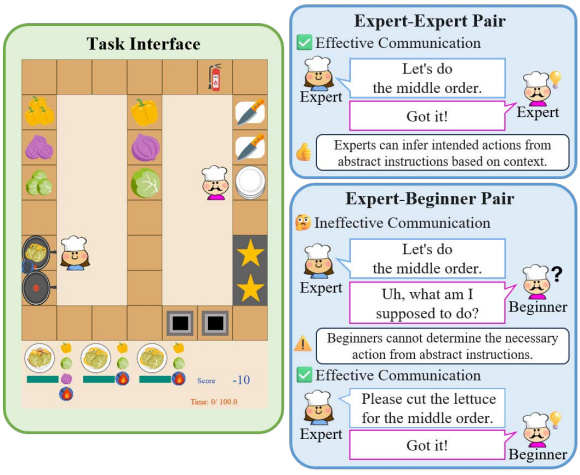

Figure 1: The collaborative cooking game environment (left) and contrasting communication examples (right). While abstract instructions suffice between experts, specific instructions are necessary when collaborating with a beginner, highlighting the need for proficiency-adaptive communication.

While research exists on human-AI collaboration for task completion through dialogue (Wu et al., 2024; Mehta et al., 2024), many existing environments primarily feature turn-based cooperation, often neglecting the crucial aspect of real-time interaction. Addressing this gap, Liu et al. extended the collaborative cooking game environment (Wu et al., 2021), which was modeled after the popular game Overcooked, specifically to assess human-AI collaboration in a real-time setting. This environment imposes strict time limits and necessitates high-frequency interactions and

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 764–779
Aug 25–27, 2025.

rapid responses from both the human player and the AI agent, simulating the pressures of time-sensitive teamwork. Within this demanding context, they developed an agent possessing the requisite response speed and reasoning abilities to engage in collaborative tasks with humans in this real-time environment (Liu et al., 2024). However, their agent primarily reacts to the human user's actions and explicit instructions, effectively casting the user into a fixed, directive role (i.e., the commander). This imposed asymmetry forces the user to constantly monitor the agent and provide guidance, potentially increasing their cognitive load and hindering the emergence of a truly synergistic and adaptive partnership. Such a rigid interaction structure may ultimately limit the effectiveness and fluidity of the collaboration.

Therefore, to foster natural and seamless collaboration, it is crucial for AI agents to adapt flexibly to their human partners. In this study, we focus specifically on adapting to the user's proficiency with the task. Considering user proficiency is vital in environments requiring mutual understanding, such as collaborative work settings. For instance, a reactive dialogue strategy where the agent primarily awaits user instructions might be effective for highly proficient users who prefer to lead. This same strategy could hinder progress and frustrate less experienced users who might benefit from more proactive suggestions or guidance.

To investigate effective dialogue strategies tailored to individual user characteristics, particularly proficiency, this study undertakes the collection of human-human collaborative task data. We employ a collaborative cooking game environment previously utilized in (Wu et al., 2021; Liu et al., 2024) as the platform for this data collection, as shown in Figure 1. The popularity of the original Overcooked game also facilitates participant recruitment, particularly for finding individuals with prior experience and potentially higher task proficiency, which is beneficial for our study focusing on proficiency-based adaptation. Our primary goal is to analyze this dataset to elucidate the interplay between user proficiency levels, their communication behaviors, and overall task performance, thereby generating actionable knowledge for designing agents capable of adapting to individual users.

The main contributions of this work are as follows: (1) We collect and present a novel dataset of human-human collaborative interactions within a real-time, high-pressure task environment. (2) We conduct a detailed analysis of this dataset, identifying dialogue strategies crucial for successful real-time collaboration, with a specific focus on how these strategies vary based on user proficiency levels. (3) We offer actionable findings and design guidelines, grounded in our empirical analysis, to inform the creation of collaborative AI agents that adapt to individual user proficiency.

2 Related Work

2.1 Collaborative Dialogue Games

Several dialogue-based game environments have been proposed to investigate collaborative behaviors (Anderson et al., 1991; Narayan-Chen et al., 2019; Jayannavar et al., 2020; Kim et al., 2019). For instance, in a Minecraft-based environment (Narayan-Chen et al., 2019), two users, designated as a builder and an architect, collaborate via text chat. The architect is given a target structure and must instruct the builder on how to construct it. In all these aforementioned environments, speakers are assigned fixed roles, such as builder and architect, where one participant typically needs to guide the other.

More recently, research has also explored environments that do not impose fixed roles on participants (Ichikawa and Higashinaka, 2023; Jeknic et al., 2024). In these settings, there are typically no strict time constraints for selecting the next action, and situations demanding immediate judgment are less common. This often simplifies the process of inferring partner intentions and adjusting dialogue strategies, potentially leading to smoother task progression.

2.2 Human-AI Collaboration

Collaborative task environments based on Overcooked have become widely used benchmarks for research aiming to develop AI agents capable of effective human-AI coordination. Notable examples of Overcooked-style environments include the one employed in this study (Wu et al., 2021; Liu et al., 2024), Overcooked-AI (Carroll et al., 2019), and CuisineWorld (Gong et al., 2024).

Among these, Overcooked-AI (Carroll et al., 2019) has arguably garnered the most significant research attention. This environment primarily focuses on coordination through actions alone, as it does not support dialogue between players. Consequently, it has served as a popular benchmark

for reinforcement learning approaches to multi-agent coordination (Charakorn et al., 2020; Sarkar et al., 2022; Wang et al., 2024). More recently, Overcooked-AI has also been utilized as a testbed for studying human-AI collaboration (Le Guillou et al., 2023; Zhang* et al., 2025). However, compared to the environment used in our work, Overcooked-AI features significantly simplified task mechanics. For instance, actions like chopping and mixing ingredients are absent, and only a single cooking method is available. We observed that when humans play this simplified version using dialogue, the low task difficulty can lead to relatively simple and potentially monotonous interaction patterns. Therefore, we opted not to use Overcooked-AI for this study, as our focus is on analyzing richer collaborative dialogues that emerge in more complex, time-pressured scenarios.

Beyond cooking scenarios, other research avenues explore human-AI collaboration in different settings. For instance, the Watch-And-Help (WAH) (Puig et al., 2021) utilizes a virtual household environment where AI agents are developed to assist humans with domestic chores (Zhang et al., 2024). These studies focus on enabling AI to understand implicit human needs and proactively offer assistance within this collaborative context.

A notable limitation common to much of the prior work discussed here, including research leveraging both Overcooked-style environments like Overcooked-AI and platforms like WAH, is the general lack of consideration for the user's proficiency or expertise level with the task. These studies often implicitly assume a uniform user capability or do not explicitly model how collaboration dynamics might change based on varying levels of user skill. This gap highlights the need for research into adaptive strategies that account for user proficiency.

3 Data Collection

3.1 Collaborative Game Environment

For the data collection, we utilized our extended version of the collaborative cooking game environment presented in (Wu et al., 2021). Within this environment, we collected human-human voice dialogue data and game logs. This environment involves a time-constrained task with complex interdependencies, specifically requiring cooperation and communication between players. The objective of the game is to achieve a high score by preparing, cooking, and serving dishes according to incoming orders. Throughout the task, a constant number of orders appear, each with its own time limit. Failure to serve an order within its time limit incurs a score penalty. To complete an order, players must follow a sequence of steps: retrieve the necessary ingredients based on the order, chop them, combine the chopped ingredients, cook them in a pot, plate the dish before it overcooks, and serve it.

The game environment updates at 40 frames per second (FPS). In each frame, each player can execute one action: move (up, down, left, or right), wait, or interact with an object (e.g., pick up an ingredient, use a tool). In this study, we enhanced the original game environment (Wu et al., 2021) by improving the interface and operability and by increasing task complexity through the addition of one new ingredient and one new cooking method. Within our environment, up to four different types of ingredients and two cooking methods are utilized. Players control either pink or blue characters and collaboratively complete incoming orders.

We designed several maps for the experiment: two single-player maps used for the initial practice and proficiency assessment phases, and four distinct collaborative maps for the main pair-based sessions. All maps are detailed in Appendix A.

3.2 Experimental Procedure

Participants were recruited from the general public. The data collection sessions were conducted in person. Participants interacted with the collaborative game environment using Sony DualSense controllers. Participants first received an overview of the collaborative environment. To familiarize them with the controls and game rules, they initially engaged in a solo practice session on a dedicated training map for 200 seconds. Following the practice, their task proficiency was measured using a different map designed for assessment. Participants performed this task individually for 100 seconds, and the score obtained during this session was recorded as their proficiency score.

After the proficiency measurement, participants were paired up for the main data collection phase. Each pair collaborated on four distinct game maps sequentially, with a time limit of 100 seconds per map. (Detailed descriptions of each map are provided in Appendix A.) To gather data across different partners, participants were re-paired with a

Cooperation	I was able to cooperate with my partner.
Communication	I was able to communicate effectively with my partner.
Role Division	We were able to divide roles appropriately.
Self Guidance	I was able to give accurate opinions or instructions.
Partner Guidance	My partner was able to give accurate opinions or instructions.
Self Adaptation	I was able to adjust my actions to my partner.
Partner Adaptation	My partner was able to adjust their actions to me.
Relative Proficiency	I believe I was more proficient in the task than my partner.

Table 1: Post-Task Survey Items

different person for each of the four maps. This procedure allowed us to collect four sets of interaction data (dialogue and in-game actions) per participant, each with a unique partner.

Immediately following the completion of each collaborative map session, participants individually completed a post-task questionnaire. The questionnaire consisted of eight items, listed in Table 1. Participants responded to each item using a 5-point Likert scale, ranging from "1: Strongly disagree" to "5: Strongly agree".

The data collection procedure was approved by the institutional review board (IRB) or ethics committee at the authors' institution.

3.3 Data Collection Results

54 participants took part in the collaborative sessions. After excluding data instances with technical issues, primarily due to microphone malfunctions, the final dataset used for analysis comprises collaborative task data and corresponding post-task survey responses from 111 unique participant pairs. The audio recordings captured during the collaborative sessions were transcribed into text using Google's Gemini 2.5 Pro Preview 03-25. Any errors resulting from the automatic speech recognition process were manually reviewed and corrected to ensure transcription accuracy.

Table 2 presents the overall statistics of the collected dataset. Note that "Actions in Games," as listed in the table, refers to the count of in-game actions excluding idle time or wait actions. This total comprises "Move Actions" (player movement) and "Interact Actions" (e.g., picking up items, using tools, delivering orders). Figure 2 illustrates the distribution of the proficiency scores measured prior to the collaborative tasks. As the figure shows, the participants exhibited a wide range of proficiency levels. While some participants achieved high scores (e.g., exceeding 100 points), a notable portion (n=9) recorded negative scores, confirming the diversity in task expertise within our participant pool.

Dialogues / Games	111
Utterances	3,412
Words in Utterances	22,844
Actions in Games	48,781
- Move Actions	34,649
- Interact Actions	14,132

Table 2: Statistics of the Dataset

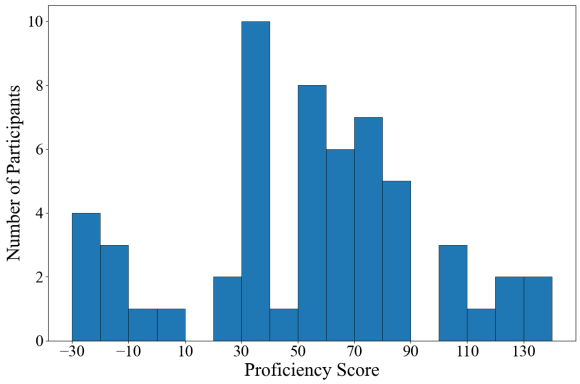

Figure 2: Proficiency Score Distribution

3.4 Dialogue Act Annotation

To facilitate the analysis of communication patterns, we annotated all transcribed utterances in our dataset with dialogue act (DA) tags. The DA tagset was designed primarily based on established schemes: the Switchboard Dialog Act Corpus (SwDA) (Jurafsky et al., 1997; Shriberg et al., 1998; Stolcke et al., 2000) and the ISO 24617-2 standard for dialogue act annotation (ISO 24617-2 DA) (Bunt et al., 2017, 2020). While most tags in our scheme directly correspond in name and function to tags within these established frameworks, we also introduced two additional tags: Encouragement and Advice. We deemed these necessary to capture specific communicative functions considered important for effective collaboration in our task setting. A complete list of the tags used, along with their descriptions, is provided in Appendix B.

The dialogue act tagging was performed automatically using the Gemini 2.5 Pro Preview 03-25. To ensure annotation consistency and mitigate potential randomness, we set the model's temperature parameter to 1.0 and performed the anno-

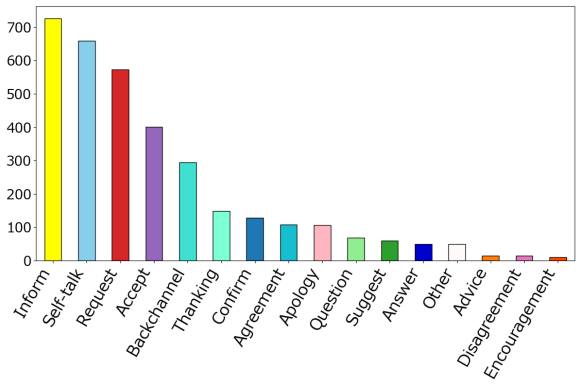

Figure 3: DA Annotation Results

tation process five times independently for each utterance. The final DA tag assigned was determined by majority vote. We first evaluated the internal consistency of this automated process using Krippendorff's α across the five annotation runs, achieving a score of 0.81, which indicates high reliability of the automated procedure itself. The specific prompts and procedures employed for the automatic annotation are detailed in Appendix C.

Furthermore, to assess the quality against human judgment, we compared the results of our automated annotation (majority vote) with manual annotations performed by a crowdworker. For this comparison, a subset of 201 utterances from 5 dialogues was annotated by a crowdworker. The Cohen's κ coefficient between the automated (majority vote) and manual annotations was calculated to be 0.604. While this value represents moderate agreement, it is lower than the inter-annotator agreement reported for SwDA ($\kappa = 0.80$) (Jurafsky et al., 1997). We attribute this difference partly to the nature of our dataset, which includes many short, context-dependent utterances (e.g., single words like "Lettuce") that are inherently ambiguous, making the annotation task challenging. Considering this inherent ambiguity in our real-time collaborative dialogue data, we believe the achieved agreement level indicates that the automatic annotation results are of reasonably high quality for the purpose of our analyses.

4 Analysis

4.1 Post-Task Survey Analysis

To investigate the relationship between participants' subjective experiences captured in the post-task survey and their objective task performance (game score), we conducted an analysis using Spearman's rank correlation coefficient (ρ). Given that the difficulty varied across game maps, we normalized the game scores for each map using robust Z-scores before calculating the correlations. This normalization accounts for potential variations in score distributions due to differing map challenges. The correlation matrix resulting from this analysis is presented in Figure 4.

First, strong positive correlations were observed among several subjective items: Cooperation, Communication, Role Division, Self Adaptation, and Partner Adaptation. This suggests that participants perceived these aspects of collaboration as highly interrelated; effective communication and clear role division likely contributed to feelings of mutual adaptation and successful cooperation.

Second, a positive correlation was also found between Partner Guidance and Partner Adaptation ($\rho = 0.61$). This indicates that participants who felt their partner provided appropriate guidance were also more likely to perceive their partner as adaptive. This could imply that effective guidance, which likely requires observing the partner's actions and needs, is interpreted as a key component of adaptive behavior from the partner's side.

Regarding the relationship between subjective ratings and objective performance, moderate positive correlations were found between the robust Z-score and both Role Division ($\rho = 0.45$) and Partner Adaptation ($\rho = 0.46$). This aligns with the expectation that effective role allocation and mutual adaptation are crucial for achieving higher scores in the collaborative task environment.

Interestingly, the correlation between the score and Self Adaptation was weaker ($\rho = 0.32$) compared to that with Partner Adaptation. Comparing the average ratings, participants rated their own adaptation (Self Adaptation, mean = 3.35) slightly lower than their partner's adaptation (Partner Adaptation, mean = 3.58). This suggests a tendency for participants to slightly underestimate their own adaptability, or potentially reflects the inherent difficulty in objectively assessing one's own adaptive behaviors during a fast-paced task.

Further analysis using Spearman's rank correlation was conducted to explore the relationships between the game score (robust Z-score) and objective behavioral metrics derived from the game logs and dialogue data. These metrics included utterance counts (Utts), interaction counts (Interactions, excluding 'Move' actions), and the pre-measured proficiency scores (ProScore), cal-

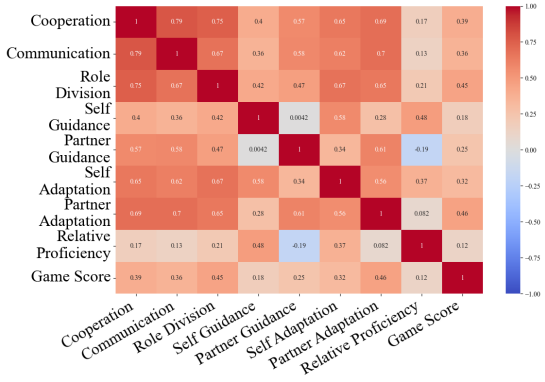

Figure 4: Spearman correlation between post-task survey results and cooperative game scores

Figure 5: Spearman correlation between post-survey results and dialogue/game features

culated separately for the individual participant (Self), their partner (Partner), and the pair combined (Total). The results of this analysis are summarized in Figure 5.

We found a moderate positive correlation ($\rho = 0.42$) between the game score and the total number of interactions performed by the pair (defined as the sum of their in-game actions, excluding 'Move' actions). This suggests that higher interaction frequency is associated with more efficient task execution and, consequently, better performance in the cooking task.

4.2 Communication Analysis

To identify dialogue strategies associated with higher task performance, we conducted multiple regression analyses using dialogue act (DA) tag frequencies as predictors, with the robust Z-score of game score as the dependent variable. We selected the predictor variables through a two-step process. First, we identified candidate predictors by selecting dialogue act tags that occurred at least 50 times across the entire dataset. Second, this initial pool was refined using a combination of stepwise regression (based on AIC) and adjustments based on domain knowledge: We added the 'Request' tag, as it is potentially crucial for guidance especially between experts and beginners, and we removed the 'Backchannel' tag, considering its relatively passive or reactive nature in active collaboration. This process yielded the 7 final predictor tags: Inform, Suggest, Agreement, Thanking, Apology, Confirm, and Request.[1]

In addition to analyzing the entire dataset (ALL, n=111), we performed separate analyses based on the difference in proficiency between partners within each pair. We calculated the absolute difference (ΔP) between the proficiency scores of the two participants in each pair. Based on this difference, pairs were categorized into three groups: Low-Difference (Low-Diff; n=36) for pairs with $\Delta P \leq 20$, Medium-Difference (Med-Diff; n=35) for pairs with $20 < \Delta P \leq 85$, and High-Difference (High-Diff; n=40) for pairs with $\Delta P > 85$. The slight unevenness in group sizes is due to multiple pairs having identical proficiency score differences falling at the category boundaries.

4.2.1 Overall Analysis

The results of the multiple regression analyses are presented in Table 3. The analysis across all pairs reveals that higher frequencies of Inform, Thanking, and Agreement are significantly associated with higher game scores. This suggests that collaborative behaviors emphasizing information sharing about the current state and expressing mutual agreement and appreciation are conducive to better performance. Conversely, higher frequencies of Suggest and Apology are significantly associated with lower scores. This indicate that frequent suggestions could reflect ongoing difficulties in reaching consensus or coordinating actions, while frequent apologies likely point to the occurrence of errors or inefficiencies during the task.

4.2.2 Impact of Proficiency Difference

We then examined the results for the three groups based on the proficiency difference between partners. The results are also shown in Table 3.

In the Low-Diff group, pairs with similar proficiency levels, a higher frequency of Apology was

[1]We checked for multicollinearity using the Variance Inflation Factor (VIF); all values were below 5, indicating it was not a significant concern.

DA	ALL	Low-Diff	Med-Diff	High-Diff	Low-Diff-H	Low-Diff-L
Inform	0.059*	0.043	0.094*	0.056*	-0.024	0.169
Request	0.017	-0.030	-0.074	0.119*	-0.278*	-0.039
Thanking	0.160*	0.214	0.204	0.082	0.446*	0.341
Confirm	-0.067	-0.050	0.136	-0.109	-0.046	-0.189
Agreement	0.178*	0.085	0.196*	0.388*	0.745*	0.021
Apology	-0.151*	-0.385*	-0.074	0.152	-0.385*	-0.648
Suggest	-0.252*	-0.220	-0.310*	-0.276	-0.256	-0.529

Table 3: Results of multiple regression analyses predicting game score (robust Z-score) from the selected seven dialogue act frequencies. Standardized coefficients (β) are shown for the entire dataset (ALL), subgroups based on proficiency difference (Low-Diff, Med-Diff, High-Diff) and further subdivision of Low-Diff (Low-Diff-L, Low-Diff-H. * $p < 0.05$ (Benjamini-Hochberg corrected).

significantly associated with lower scores. This pattern suggest that frequent apologies, likely indicating the occurrence of errors or inefficiencies during the task, are particularly detrimental to performance when partners have similar skill levels, perhaps because neither partner can easily compensate for the mistakes. The results for the Med-Diff group roughly mirrored those of the overall dataset (ALL), with Inform, Agreement, and Suggest showing significant associations. Notably, in the High-Diff group, pairs with a large proficiency difference, Request emerged as significant positive predictor of performance, a pattern not observed in the other groups. This suggests that in high-difference pairings, explicit requests facilitate better coordination and outcomes. Agreement also showed a positive association in this group.

4.2.3 Analysis within Low-Difference Pairs

Recognizing that the Low-Diff group could comprise pairs where both partners have high proficiency or both have low proficiency, we further divided this group based on the median of their combined proficiency scores. This resulted in a high-proficiency pair subgroup (Low-Diff-H; n=23) and a low-proficiency pair subgroup (Low-Diff-L; n=17). The results are also shown in Table 3. In the Low-Diff-H (High-Proficiency Pairs) subgroup, where both partners were highly proficient, the overall trend was similar to the ALL group, but Inform was no longer a significant predictor. This might imply that highly skilled players can effectively infer necessary information from the visual game state, reducing the relative importance of explicit information sharing through dialogue. Furthermore, and perhaps counter-intuitively, a higher frequency of Request acts was significantly associated with lower scores in this high-proficiency pair group. It might indicate that frequent explicit instructions between experts can lead to inefficiencies, potentially disrupting autonomous workflows or causing micromanagement overhead. This suggests that for expert-expert collaboration in this environment, excessive reliance on explicit requests may be less effective than coordination based on mutual anticipation and minimal, targeted communication. However, Thanking and Agreement remained significant positive predictors, suggesting that fostering a cooperative atmosphere is still beneficial even among experts. Interestingly, in the Low-Diff-L (Low-Proficiency Pairs) subgroup, no dialogue act tags showed a significant association with the score. This could suggest that for pairs where both partners lack proficiency, variations in communication strategies have less impact on the outcome compared to the overriding factor of their fundamental skill limitations.

DA	Med-Diff	High-Diff
$\text{Inform}_{\text{high}}$	0.125^{+}	0.017
Inform_{low}	0.106^{+}	0.058^{*}
Request_{high}	-0.029	0.132^{+}
Request_{low}	0.069	0.193^{+}
Agreement_{high}	0.206^{+}	0.356^{+}
Agreement_{low}	0.009	0.413^{+}
Suggest_{high}	-0.301^{+}	-0.161
Suggest_{low}	-0.104	-0.396

Table 4: Results of multiple regression analysis predicting game score (robust Z-score) using dialogue act frequencies separated by high-proficiency (Tag_{high}) and low-proficiency (Tag_{low}) speakers in Med-Diff and High-Diff pairs. * $p < 0.05$, $^{+}$ $p < 0.10$ (Benjamini-Hochberg corrected).

4.2.4 Speaker Proficiency within Pairs

To further analyze speaker-specific dynamics in the Med-Diff and High-Diff groups, we conducted additional set of multiple regression analyses. Analyzing all original predictor tags separately for each speaker (high vs. low proficiency) would result in a large number of variables (16 predic-

tors) relative to the sample size in these subgroups (n=35 and n=40, respectively), potentially leading to model instability and unreliable estimates. Therefore, to maintain model robustness while focusing on key communicative functions, we selected four core dialogue acts based on our preceding analyses and their fundamental roles in coordination: Inform (for establishing shared situational awareness), Suggest and Request (for planning and directing actions), and Agreement (for confirming mutual understanding and facilitating smooth interaction). The results of this speaker-specific analysis are presented in Table 4.

For the Med-Diff group, the results suggest potential trends, although no predictors reached the conventional $p < 0.05$ significance level after correction (Table 4). Specifically, Inform acts tended to be positively associated with the score when initiated by either the high-proficiency partner (Inform_{high}) or the low-proficiency partner (Inform_{low}), suggesting that information sharing remains beneficial regardless of who provides it. Similarly, Agreement expressed by the higher-proficiency partner also showed a positive trend. Conversely, Suggest acts initiated by the higher-proficiency partner tended to be negatively associated with the score, suggesting that excessive top-down suggestions might disrupt collaboration in these moderately heterogeneous pairs.

Turning to the High-Diff group, the analysis reveals that Inform acts initiated by the low-proficiency partner were significantly positively associated with performance. This suggests that when less experienced partners actively share information about their status or actions, it significantly aids coordination. Furthermore, there were positive trends for Request acts initiated by both the high-proficiency partner and the low-proficiency partner. This suggests that explicit requests or instructions, regardless of who issues them, facilitate coordination when the skill gap is large. Similarly, Agreement initiated by both partners also tended to be beneficial.

5 Design Guidelines for Collaborative Dialogue Systems

Based on the analyses presented in this paper, we propose the following design guidelines for dialogue systems intended to engage in collaborative tasks with human users.

First, a fundamental strategy for the system should involve proactive communication regarding its own status and intentions (Inform), coupled with frequent expressions of Agreement and Thanking in response to the user's utterances and actions. Our analysis indicated that these dialogue acts are generally associated with higher task performance (as discussed in Section 4.2.1). Furthermore, given that effective Role Division was found to be strongly correlated with task success (Section 4.1), proactive information sharing is crucial for establishing and maintaining appropriate roles between the user and the system.

Second, it is critical for the system to dynamically assess or infer the proficiency difference between itself and the human user during the collaborative process and adapt its interaction strategy accordingly. If the system possesses high task proficiency and the perceived proficiency difference with the user is relatively small (Low-Diff or Med-Diff scenarios), continuing with the aforementioned strategy of proactive information sharing, agreement, and thanking appears effective (Section 4.2.1, 4.2.2 and 4.2.3). However, when the system is highly proficient but perceives a large proficiency gap with the user (High-Diff scenario), the system should adopt a more directive role, actively issuing Requests or instructions to guide the user and take leadership in the task (Section 4.2.4). Conversely, if the system has low proficiency while the user is highly proficient (another instance of a High-Diff scenario, but with roles reversed), the system should be designed to encourage or solicit guidance and instructions from the user, effectively positioning the user as the leader (Section 4.2.4). Finally, in situations where both the system and the user have low proficiency (Low-Diff-L scenario), our findings suggest that communication strategies alone may be insufficient to ensure efficient collaboration (Section 4.2.3). In such cases, the primary focus should be on improving the system's fundamental task execution capabilities before sophisticated adaptive communication strategies can become truly effective.

6 Conclusion

This paper presented an analysis of a novel human-human interaction dataset from a real-time collaborative game, investigating how communication patterns and subjective experiences correlate with task success, particularly considering

partner proficiency differences. Our findings revealed that while some communication acts consistently aid collaboration, the effectiveness of others depends heavily on the partners' relative skill levels. Based on these insights, we proposed design guidelines for adaptive collaborative dialogue systems.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 25H01382 and HAYAO NAKAYAMA Foundation for Science & Technology and Culture Research Grants (25-A2-33).

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, and 1 others. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The iso standard for dialogue act annotation, second edition. In *12th Edition of its Language Resources and Evaluation Conference (LREC 2020)*, pages 549–558. European Language Resources Association (ELRA).

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso 24617-2 standard. *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 109–135.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32.

Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. 2020. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. In *Neural Information Processing. ICONIP*, pages 395–402. Springer.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Li Fei-Fei, Jianfeng Gao, and Hoi Vo. 2024. MindAgent: Emergent gaming interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3154–3183, Mexico City, Mexico. Association for Computational Linguistics.

Takuma Ichikawa and Ryuichiro Higashinaka. 2023. Modeling collaborative dialogue in minecraft with action-utterance model. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 75–81.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a minecraft dialogue. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2589–2602.

Isidora Jeknic, David Schlangen, and Alexander Koller. 2024. A dialogue game for eliciting balanced collaboration. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 477–489, Kyoto, Japan. Association for Computational Linguistics.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Yi Lai, Atreyi Kankanhalli, and Desmond Ong. 2021. Human-ai collaboration in healthcare: A review and research agenda. In *Hawaii International Conference on System Sciences*.

Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2023. Trusting artificial agents: Communication trumps performance. AAMAS '23, page 299–306, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. 2024. Llm-powered hierarchical language agent for real-time human-ai coordination. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1219–1228.

Nikhil Mehta, Milagro Teruel, Xin Deng, Sergio Figueroa Sanz, Ahmed Awadallah, and Julia Kiseleva. 2024. Improving grounded language understanding in a collaborative environment by interacting with agents through help feedback. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1306–1321, St. Julian's, Malta. Association for Computational Linguistics.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B. Tenenbaum, Sanja Fidler, and Antonio Torralba. 2021. Watch-and-help: A challenge for social perception and human-{ai} collaboration. In *International Conference on Learning Representations*.

Bidipta Sarkar, Aditi Talati, Andy Shih, and Dorsa Sadigh. 2022. Pantheonrl: A marl library for dynamic training interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 13221–13223.

Elizabeth Shriberg, Rebecca Bates, Paul Taylor, Andreas Stolcke, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3–4):439–487.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371.

Vanshika Vats, Marzia Binta Nizam, Minghao Liu, Ziyuan Wang, Richard Ho, Mohnish Sai Prasad, Vincent Titterton, Sai Venkat Malreddy, Riya Aggarwal, Yanwen Xu, and 1 others. 2024. A survey on human-ai teaming with large pre-trained models. *arXiv preprint arXiv:2403.04931*.

Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing transparency for effective human-ai collaboration. *Information Systems Frontiers*, 24(3):877–895.

Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6.

Xihuai Wang, Shao Zhang, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. 2024. Zsc-eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. *Advances in Neural Information Processing Systems*, 37:47344–47377.

Guande Wu, Chen Zhao, Claudio Silva, and He He. 2024. Your co-workers matter: Evaluating collaborative capabilities of language models in blocks world. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4941–4957, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sarah A. Wu, Rose E. Wang, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. 2021. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2024. Building cooperative embodied agents modularly with large language models. In *The Twelfth International Conference on Learning Representations*.

Shao Zhang*, Xihuai Wang*, Wenhao Zhang, Chaoran Li, Junru Song, Tingyu Li, Lin Qiu, Xuezhi Cao, Xunliang Cai, Wen Yao, Weinan Zhang, Xinbing Wang, and Ying Wen. 2025. Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration. *ACL 2025*.

A Cooking Game Environment

(a) Practice (b) Proficiency measurement (c) Standard

(d) Concave (e) Partition (f) Expanded Concave

Figure 6: The Designed maps in the Overcooked environment.

Figure 6a shows the map used for the initial practice session, while Figure 6b depicts the map used for proficiency measurement. Both are designed with a simple structure suitable for single-player task execution and feature longer time limits for order completion compared to the collaborative maps, facilitating familiarization and assessment, respectively.

Figures 6c, 6d, 6e, and 6f show the four collaborative maps used for data collection, which participants played sequentially in this order. Figure 6c represents a standard layout for collaborative play, featuring a spatially balanced arrangement of cooking stations. This design allows for relatively easy movement to different stations and facilitates flexible mutual assistance between players based on the evolving situation. Figure 6d is designed to encourage clearer role division through its structure and station placement. Movement between the left and right areas requires passing through a central passage, promoting a style where each player tends to focus on specific roles while coordinating their efforts. Figure 6e features the most rigid role division structure among the maps used, completely separating the two players via a central partition. Completing orders necessitates strong interdependence and relies heavily on effective communication for coordinating actions. Figure 6f shares a similar structure to Figure 6d but introduces greater complexity by incorporating two cooking methods and the maximum of four ingredient types, requiring players to handle more intricate orders. Furthermore, while the player on the right side can efficiently perform tasks from retrieving vegetables to chopping, the diversity of orders creates an asymmetric workload distribution. Consequently, the player on the left, who typically has a lighter task load,

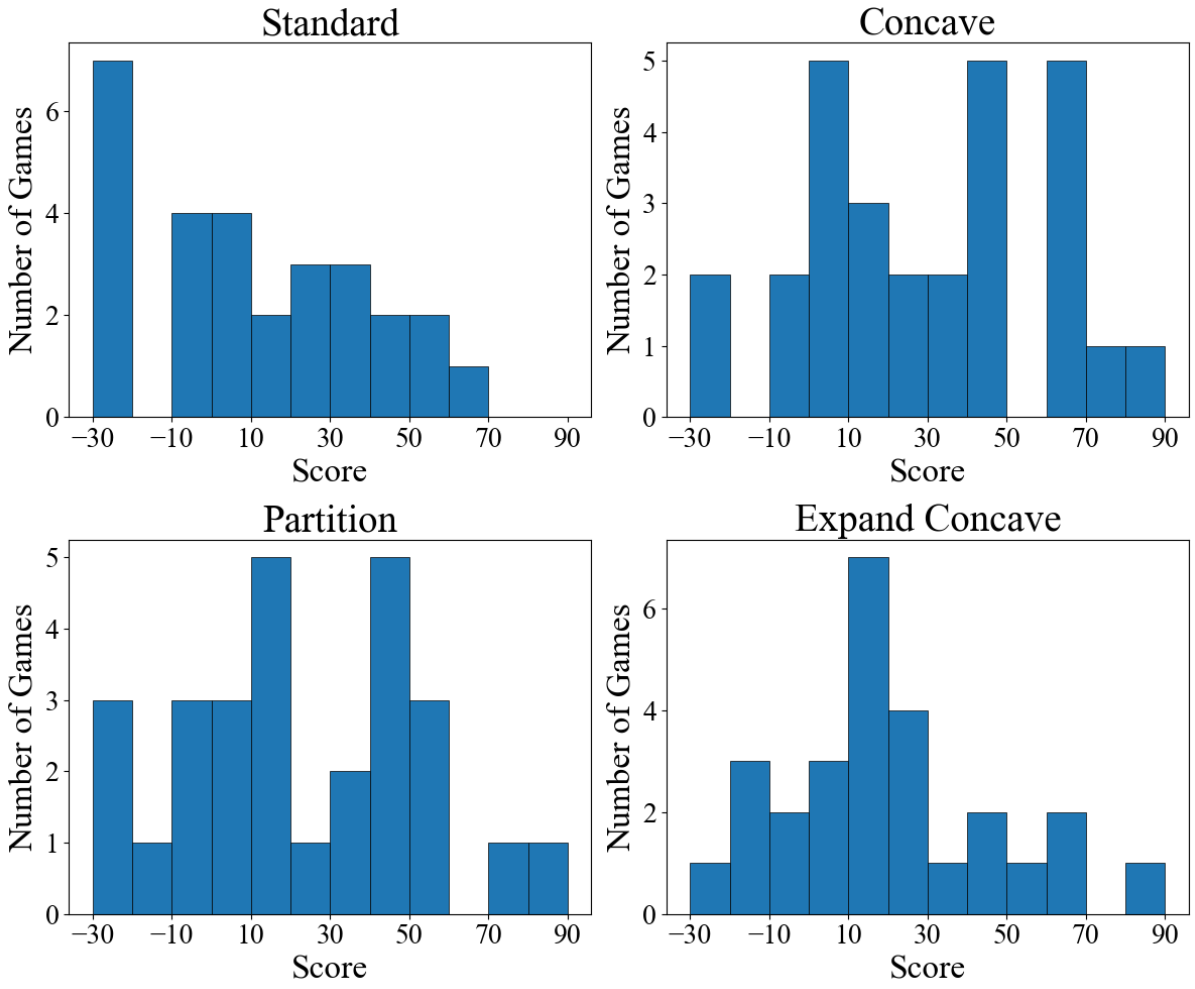

Figure 7: Score Distribution for Four Maps.

often assumes a crucial role in planning and leading the overall cooking process.

Figure 7 displays the distribution of game scores for each collaborative map. On the Standard map, scores tend to be skewed towards the lower end. This is likely because it was the first collaborative map participants played, making it challenging for them to establish efficient coordination patterns immediately. In contrast, the Concave and Partition maps show scores distributed across a wider range, suggesting that performance on these maps is more sensitive to the success or failure of the pair's coordination and chosen strategy. The Expand Concave map also shows a tendency towards lower scores, although less pronounced than the Standard map. This is presumably because the map's structure and the complexity of orders hinder efficient collaboration, making it difficult to achieve high scores.

B Dialogue Act

A list and description of the dialogue act (DA) tags used in this study are presented in Table 5. Among the total of 16 tag types shown in the table, Advice and Encouragement are tags specifically defined for this research to capture nuances relevant to the collaborative task. The tags Inform, Request, Confirm, Agreement, Apology, Question, Suggest, Answer, Disagreement, and Other originate from the ISO 24617-2 DA standard (Bunt et al., 2017, 2020). The remaining four tags, namely Self-talk, Accept, Thanking, and Backchannel, are derived from the SwDA scheme (Jurafsky et al., 1997; Shriberg et al., 1998; Stolcke et al., 2000). The descriptions and definitions for each tag, also provided in Table 5, were developed by referencing existing literature and adapting them to the specific context and requirements of our collaborative cooking game environment.

Tag	Description
Inform	Utterances sharing task-related facts, states, or actions with the partner.
Self-talk	Utterances in which the speaker verbalises their own thoughts or feelings; no call for action or information directed to the partner.
Request	Utterances asking the partner to perform some action, expressed imperatively or politely.
Accept	Utterances that accept or agree to a proposal, request, or confirmation from the partner.
Thanking	Utterances expressing gratitude to the partner.
Backchannel	Short reactive tokens showing attention or understanding without adding new content (e.g., "uh-huh", "okay").
Confirm	Utterances checking the correctness of information or mutual understanding (e.g., "Is this okay?").
Agreement	Utterances explicitly expressing agreement or alignment with the partner's opinion.
Apology	Utterances in which the speaker apologises for their own behaviour or the situation.
Question	Utterances seeking information or opinion from the partner, excluding confirmation or suggestion questions.
Suggest	Utterances presenting a new idea or course of action for consideration.
Answer	Utterances that respond appropriately to the immediately preceding question, including replies to confirmation or suggestion questions.
Advice	Utterances pointing out improvements or giving guidance to the partner (e.g., "You should ⋯").
Disagreement	Utterances expressing a negative stance toward the partner's statement, opinion, or question (e.g., "No, that's wrong").
Encouragement	Utterances that cheer up or motivate the partner (e.g., "You can do it!").
Other	Utterances that do not fit any category above, have unclear intention, or comprise greetings or fillers.

Table 5: Dialogue act tags used in this study.

C Prompt for Dialogue Act Annotation

The following describes the prompt provided to the language model (Gemini 2.5 Pro Preview 03-25) for automatic dialogue act annotation. The prompt consisted of several sections: a task description, a list of dialogue act tags with definitions and priority rules, specific annotation guidelines, a summary of the game rules, descriptions of the input log format and the desired output format, and finally the log data itself.

Prompt for Dialogue Act Annotation

 # Task Description
You are given mixed logs of in-game actions and voice chat produced by two players collaborating in an Overcooked-inspired cooking game. For every utterance, assign exactly one dialogue-act tag from the list below. If multiple tags seem plausible, choose the single most appropriate one.

 # Dialogue Act Tag Set
- Inform: Utterances sharing task-related facts, states, or actions with the partner.
... (same as Table 5. omitted)
- Other: Utterances that do not fit any category above, have unclear intention, or comprise greetings or fillers.

 # Game Rules
- The goal is to prepare specified dishes by chopping, mixing, cooking ingredients, plating them, and placing them at the serving station within the time limit.
- Three orders are always visible on screen, each with its own time limit.
- A player can move up, down, left, or right and interact with an adjacent object.
- Ingredients are lettuce, bell pepper, tomato, and onion.
- Interacting with an ingredient crate produces that ingredient.
- Interacting with a cutting board while holding an ingredient cuts it.
- Two different cut ingredients placed at the same location can be mixed together (three or more can also be mixed).
- Cut ingredients can be cooked in a frying pan or boiled in a pot.
- Over-cooking or over-boiling causes a fire.
- Fires must be extinguished with a fire extinguisher.
- After a fire is out, the burnt dish must be discarded in the trash.
- Interacting with the trash while holding an ingredient, cut ingredient, or dish discards it.
- Interacting with the plate rack produces a plate.
- A cooked dish can be plated by interacting with the pan or pot while holding a plate.

- A plated dish placed at the serving station yields score.
- While holding a plate, a player can still pick up ingredients or cut ingredients.
- Any portable item can be placed on an empty counter by interacting with it.

 # Input format
The log is a JSON file in which game-action entries and speech-action entries are interleaved in chronological order.
Each entry has the following common keys:
- id: unique event identifier (integer)
- time: timestamp in seconds
- player_side: initial spawn side of the player
- player_id: ID of the focal player (logs include only this player)
- action_type: Type of event ('game' or 'utterance')

If action_type is 'game', the following keys are included:
- action: Specific type of game action (e.g., 'Move', 'Chop_FreshTomato ')
- position: Coordinates where the action occurred ([x, y]). For 'Move', it's the position after moving; otherwise, it's the target location of the action

If action_type is 'utterance', the following keys are included:
- utterance: The content of the utterance.
- end_time: Time the utterance ended (in seconds).

 # Output format
- Estimate the tag for all entries where "action_type" is "utterance".
- The final output should be in JSON format as follows:

```
[
  {
    " id": "int", // Same id as the target entry
    "tag": "str"  // The estimated tag
  },
...
]
```

 # Now please annotate the following log:
{GAME_AND_DIALOGUE_LOG}

D Dialogue Example

Figure 8 presents a dialogue excerpt from a High-Difference (High-Diff) pair, consisting of one high-proficiency (expert) player and one low-proficiency (beginner) player. Figure 9 shows a dialogue example from a Low-Difference High-Proficiency (Low-Diff-H) pair, where both participants had high proficiency scores. In both figures, each utterance is shown alongside its assigned dialogue act tag based on our annotation process. To aid understanding of the interaction flow, accompanying descriptions of relevant game actions and illustrations are also included.

The dialogue example in Figure 8 is characterized by a dynamic where the high-proficiency player primarily issues Request acts, guiding the interaction, while the low-proficiency player takes a more responsive role. In contrast, the example in Figure 9 demonstrates how high-proficiency partners effectively coordinate through more abstract Inform acts, accurately inferring each other's intentions from these cues and smoothly transitioning to subsequent actions.

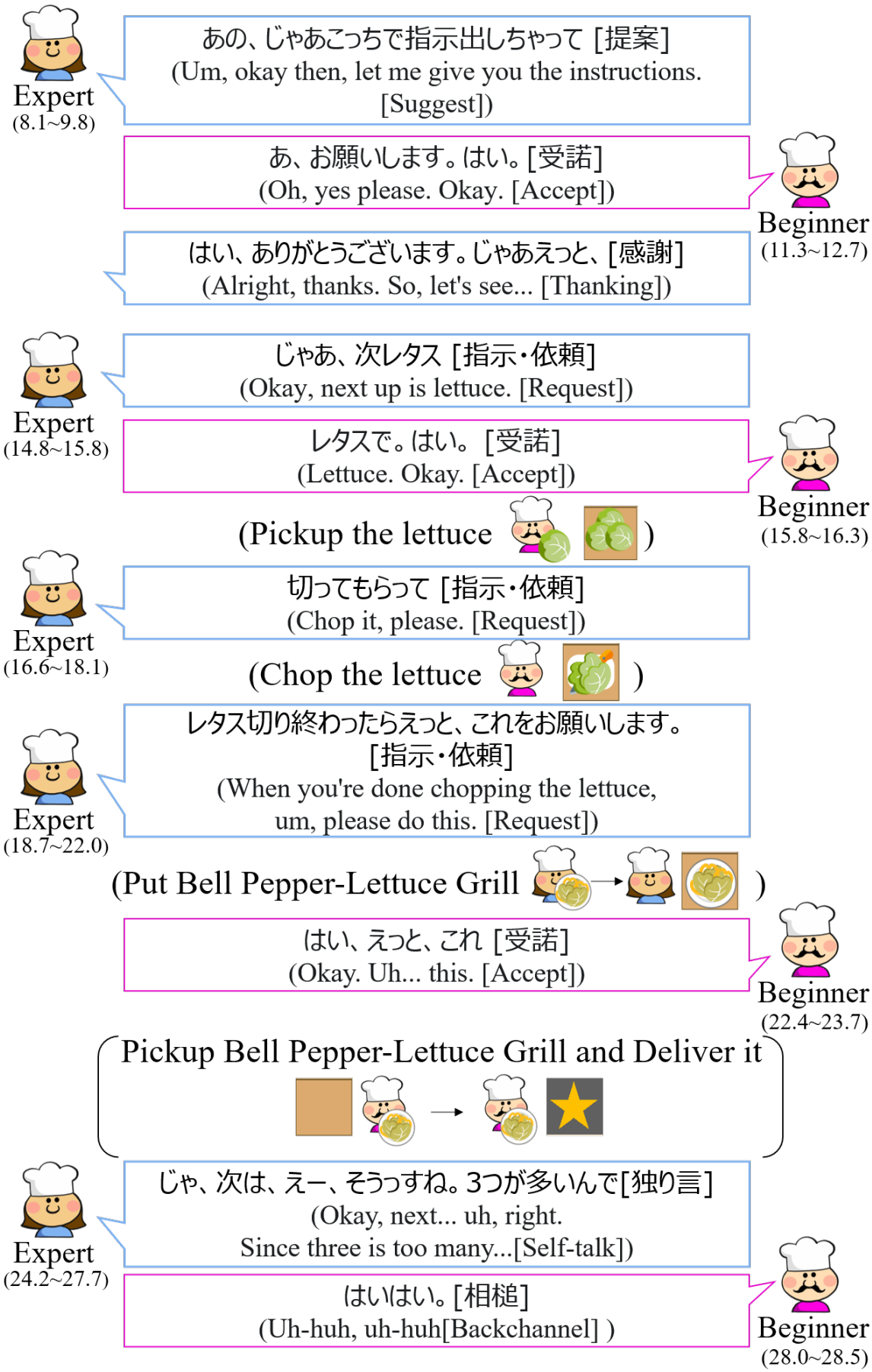

Figure 8: Dialogue examples between experts and beginners (collected in Japanese and translated to English by authors) during collaborative work.

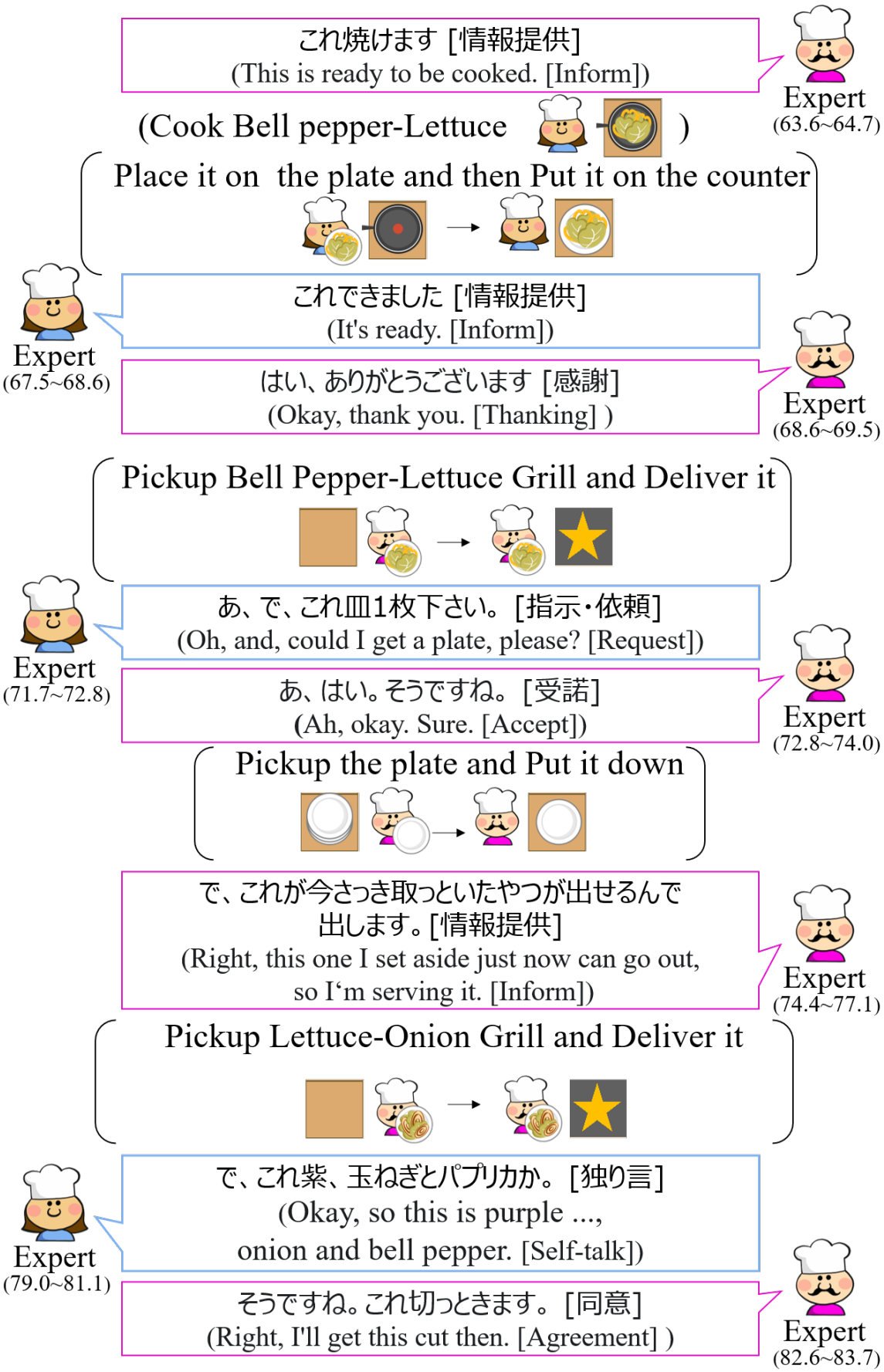

Figure 9: Dialogue examples between experts (collected in Japanese and translated to English by authors) during collaborative work.