Predicting Teamwork Performance Using a Large Language Model to Annotate Repair and Grounding Utterances in Task-Based Conversations

Akshitha Kartigueyan
Department of Computer Science
Purdue University
West Lafayette, IN, USA
0009-0007-6872-3343

Joseph P. Salisbury
ISR Innovations
Riverside Research
Melbourne, FL, USA
0000-0002-9448-9024

***Abstract*— Effective teamwork relies heavily on the quality of communication among team members. Conversational dynamics, such as the use of repair and grounding mechanisms—conversational strategies that maintain mutual understanding—play a crucial role in promoting cohesion within a team. Here, we describe an approach that leverages a large language model (LLM) to detect repair and grounding (R&G) utterances during a cooperative task and evaluate how these factors can predict team performance. To demonstrate this, we collected and analyzed video data from YouTube of the cooperative multiplayer puzzle game Keep Talking and Nobody Explodes. Player communication was transcribed using speech-to-text tools to generate speaker-labeled transcripts. Utterances were labeled for R&G mechanisms by an LLM using few-shot learning. Statistical analyses reveal distinct patterns in communication between successful and unsuccessful bomb defusal trials, which enabled the development of a model to predict task outcome.**

Keywords—Team communication, conversational dynamics, repair and grounding, task performance prediction

I. Introduction

Effective communication is essential to achieving successful outcomes in high-stakes collaborative tasks. In such scenarios, teams often operate under significant time constraints and high levels of stress, making the quality of interactions crucial for performance. Effective communication enables teams to coordinate actions, share information accurately, and make timely decisions, which are vital components of success in high-stakes environments such as emergency medical response, military operations, air traffic control, space missions, and disaster relief efforts [1], [2], [3].

Given the potentially life-threatening nature of these tasks, the ability to objectively assess and enhance teamwork communication and performance is of paramount importance. Despite the significance of effective communication, there remains substantial challenges in accurately assessing and improving team interactions in real-time. Current team training assessment is often limited, relying on high-level checklists and post-hoc, subjective evaluations of performance. Traditional approaches often rely on manual observation and coding of team interactions, which are time-consuming, labor-intensive, and subject to human bias [4]. The development of artificial intelligence (AI) that can objectively assess effective communication offers substantial benefits to enhance team training and performance evaluation [5]. This capability would be particularly useful for continuous performance monitoring and adaptive training programs, where timely insights are crucial for corrective action and skill development [6], [7].

Recent advances in large language models (LLMs) have made it substantially more feasible for an AI to analyze conversation dynamics to identify behaviors that contribute to team success [8], [9]. For example, these systems could be trained to automatically detect repair and grounding (R&G) in conversations, which are essential for maintaining mutual understanding and addressing misunderstandings in high-stakes tasks [10]. Grounding is the interactive process by which mutual understanding between individuals is constructed and maintained, ensuring that participants are perceiving and accepting each other's utterances [11], [12], [13], [14]. As part of maintaining common ground, repair strategies can be used to detect and resolve communicative problems and clarify misunderstandings [15]. These can include self-corrections [16], where a speaker recognizes and rectifies their own mistake, and other-initiated repairs [17], [18], where another team member points out and helps to correct the error. R&G helps ensure that all team members are on the same page, enhancing coordination and reducing the likelihood of errors. Thus, an AI that can detect R&G can provide feedback on communication effectiveness, helping teams recognize and correct poor communication habits before they become entrenched.

Toward this goal, our study aims to evaluate: 1) the ability of LLMs to annotate R&G utterances in team conversations; and 2) to identify conversational factors that are predictive of task performance and teamwork expertise. We leverage YouTube videos of the cooperative multiplayer game *Keep Talking and Nobody Explodes* (KTaNE) as a practical source of data to develop and demonstrate methods for analyzing conversations that are relevant to teamwork in high-stakes tasks. In KTaNE, one player ("the defuser") must disarm a virtual bomb based on instructions provided by other players (the "experts"), who have access to a bomb defusal manual but cannot see the bomb. This setup necessitates clear, precise, and effective communication under time pressure, closely mimicking the communication demands found in high-stakes real-world scenarios. The game's requirement for continuous verbal interaction and real-time problem-solving makes it a rich source of data for conversational analysis [19].

II. Methods

A. Summary of Approach

To demonstrate predicting task performance and expertise based on R&G, we developed the workflow summarized in Figure 1. Briefly, videos were identified on YouTube of KTaNE and downloaded so transcripts could be processed with speaker partitioning using Amazon Transcribe. The assignment of utterances to speakers was further refined using an LLM with few-shot prompting [20]. We then assessed the ability for LLMs to label R&G utterances in the diarized transcript using the approach from [21], comparing performance of two LLMs at different temperatures against manually annotated transcript segments. The best performing LLM was then used to label the complete dataset. Frequency of various R&G utterances were normalized and used as features, along with KTaNE puzzle difficulty, to train a model to predict task outcome and expertise.

B. Video Selection and Annotation

To obtain data for analyzing team communication, we conducted a targeted search for "Let's Play"-style [22] videos on YouTube. The search term *Keep Talking and Nobody Explodes Let's Play* was used. The following selection criteria were established to ensure the relevance and quality of the data:

1. Only videos that included the entire bomb defusal process without edits were selected.
2. Videos where players introduced additional challenges or engaged in behaviors that distort the typical gameplay experience were excluded.
3. Only videos featuring exactly two players were selected: one acting as the bomb defuser and the other as the expert providing instructions.
4. For channels with multiple KTaNE videos, preference was given to those showcasing new pairings of players to increase data diversity. When possible, preference was given to pairings of opposite genders to minimize speaker misassignment.

These criteria were selected to help ensure data we selected would provide a robust foundation for analyzing the impact of conversational dynamics on task performance.

During the selection process, a subset of videos featuring professional bomb defusers (retired US Army Airborne Infantry) playing the game were identified. These videos were notable as participants clearly had training experience for how to communicate effectively under stress. Consequently, these videos were selected and analyzed as a separate "professional" group for comparative analysis against the "amateur" player population.

To collect performance metrics for analysis, videos were inspected by the authors for start and stop times of each KTaNE puzzle, the puzzle name, the number of sub-modules in each puzzle, and whether each puzzle was successfully defused. In total, 10 videos were selected for analysis (Table 1). The chosen videos represent a mix of typical gameplay and expert-level communication, which we hypothesized would enable identifying factors that contribute to effective teamwork in high-stakes environments.

TABLE I. Summary of Videos Selected For Analysis

YouTube ID	Expertise Level	# Success	# Fail	Length (mm:ss)
FJ6M03H2-RU	Amateur	4	2	33:10
vF5jFFkL1p0	Amateur	2	7	41:01
BdvdZlh1Xdo	Amateur	4	4	43:23
BjjzSbSEwXo	Amateur	1	3	18:21
AlqDjkER5Ws	Amateur	0	6	29:07
GVyfxHMYDH4	Amateur	9	4	57:25
7a96RyJVfD8	Amateur	3	2	21:10
dl78TfbnahI	Amateur	2	4	20:48
BYunaBkn9Ng	Professional	3	0	8:19
ESuAQHt5Dus	Professional	2	1	8:43
Totals	8 Amateur; 2 Pro	30	33	281:27

C. Speech-to-Text with Diarization

To generate transcripts with speaker assignments, we first used Amazon Transcribe to produce a transcript of the video with timestamps associated with each speaker. Prior to upload, we edited out advertisements that appeared at the beginning or end of some videos featuring the voice of another person. For Amazon Transcribe job settings, we specified the language as English and requested speaker partitioning with a maximum of two speakers. Amazon Transcribe provides two outputs: a subtitle file in the SRT format, which contains segments of utterances and the associated timestamps, and a JSON file which includes timestamps of speaker diarization. We developed a custom script to cross-reference timestamps between the SRT and JSON files to compile a transcript with alternating speaker IDs followed by their utterances. In some cases, discrepancies in timestamps resulted in blank entries. To address this, the script assigned the last complete or incomplete sentence from the previous utterance to the blank entry. If this adjustment created a new blank segment, the script removed the original blank entry and merged the previous and subsequent utterances under the same speaker.

Occasionally, Amazon Transcribe struggled to diarize conversations involving an exchange of brief phrases between

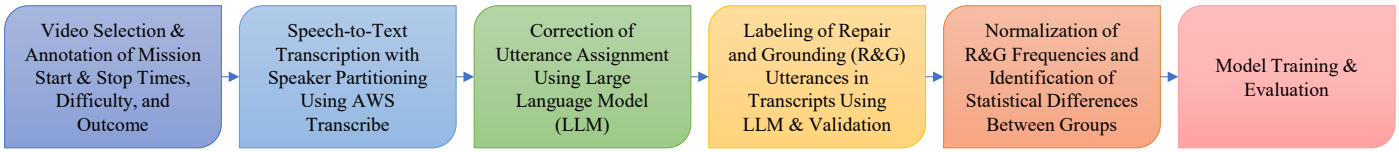

Fig. 1. Overview of steps to develop predictive models based on LLM-annotated repair and grounding (R&G) utterances from online videos.

speakers. We observed this particularly in videos with players of the same gender. To rectify this and further refine transcripts where this was notable, we utilized Claude 3.5 Sonnet for LLM-based utterance assignment correction using a prompt with examples of a transcript before and after manual speaker reassignment based on inspection of the source video. While we did not rigorously evaluate the accuracy of this approach, inspection of transcripts where this approach was applied showed notable improvement in speaker assignment when compared with source videos.

D. R&G Labeling in Transcripts

To label R&G utterances, we used the prompt from [21]. To compare performance of various pre-trained LLMs, we sent prompts with speaker-labeled transcripts to OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet. To create a standard for comparing performance of LLMs, we manually labeled segments from two transcripts, including 128 utterances (~10 minutes of gameplay dialogue). To assess variability in LLM responses, we evaluated LLM's across a range of temperatures (T = 0, 0.1, 0.3, 0.5, 1) in replicates of five. During preliminary data analyses, we noted that successful runs featured heavy use of the adjacency pair [23] *proactive-grounding* followed by *other-repetition*. To evaluate its importance, we determined the frequency of this sequential pair in each puzzle attempt and included it in subsequent analyses as its own feature.

E. Data Normalization

To ensure comparability across different gameplay sessions and puzzles, we applied a data normalization process to the R&G labeled utterances. This process aimed to account for variation in the length and complexity of puzzles by standardizing the occurrence of R&G mechanisms relative to the total number of utterances labeled within each puzzle. We segmented labeled transcripts puzzle by puzzle, with each segment corresponding to a discrete defusal attempt. For each puzzle, we tallied the number of utterances labeled with specific R&G mechanisms. To normalize data, we divided the raw count of each R&G mechanism for a given puzzle by the total count of all R&G-labeled utterances in that puzzle.

F. Statistical Analysis

To evaluate the relationship between conversational dynamics and task performance, we conducted statistical analyses using the normalized values of R&G mechanisms for each puzzle. We performed two sets of group comparisons: 1) samples from puzzles that were successfully defused ("Success" samples) were compared with samples where the bomb detonated ("Failure" samples); 2) samples from professional military bomb defusers ("Professional" samples) were compared with samples from the amateur players ("Amateur" samples). To compare the normalized R&G values between groups, we performed independent T-tests using SciPy [24]. To filter out R&G mechanisms that were too infrequent, we only included those with a mean normalized value of greater than 0.02 in at least one sample group under test. To account for multiple comparisons, we used the two-stage false discovery rate Benjamini/Hochberg correction [25], [26]. We choose an alpha level of 0.1 to allow for the detection of differences that could guide future, more rigorous studies.

G. Model Training and Evaluation

To predict both outcome ("Success" vs. "Failure") and "expertise" ("Professional" vs. "Amateur"), we trained classification models using the normalized R&G frequencies. To limit the number of features and focus on the most informative variables, we selected the three most significant features from each group comparison. Given that task outcome may also be influenced by the complexity of the puzzle, we included "mission difficulty" as a fourth feature in the outcome model. Difficulty was calculated based on the mission structure of KTaNE using the formula:

$$\text{difficulty} = (\text{section \#} - 1) + \frac{(\text{mission \# in section} - 1)}{\text{\# of missions in section}} \quad (1)$$

This formula normalizes mission difficulty across sections of increasing difficulty with varying numbers of missions.

To classify both outcome and expertise, we employed a bagging classifier with a support vector machine (SVM) as the base estimator using scikit-learn [27], [28]. The SVM model used a linear kernel with C = 10 to control regularization, while the bagging ensemble leveraged 10 estimators to improve performance through bootstrap aggregation. To evaluate model performance, we used a stratified train-test split. To predict outcome, we reserved 25% of the data for testing. To predict expertise, we used a 50% split given the limited number of professional trials (6 total). Given there were significantly fewer professional samples than amateur samples, we used the Synthetic Minority Over-sampling Technique (SMOTE) [29] to create a balanced dataset for training. SMOTE was configured with 2 nearest neighbors due to the small number of professional samples. We first assessed model performance through 5-fold stratified cross-validation. To further validate the robustness of the model's accuracy, we conducted a bootstrap analysis [30]. We assessed performance using standard classification metrics, including overall accuracy, precision, recall, and F1-score.

III. Results

A. LLM Labeling Performance

Overall, Claude 3.5 Sonnet produced the most accurate results, achieving 80.5±0.1% (mean ± S.D.) accuracy in labeling utterances for R&G mechanisms at a temperature T = 0 across five replicates. In contrast, GPT-4o exhibited a significantly lower accuracy of 48.5±0.1% at T = 0 (Figure 2), which decreased to 39.3±4.9% at T = 1. Claude 3.5 Sonnet's accuracy remained consistently high (~80%) across a range of T, although variance increased slightly with T, up to 78.9±1.9% at T = 1. We selected Claude 3.5 Sonnet at T = 0 for further analyses due to its optimal balance of accuracy and consistency. Note that not all R&G mechanisms were in the annotated segments, and some had relatively few instances.

B. Statistical Analysis of R&G Frequency

To explore the relationship between R&G mechanisms and task performance, we conducted statistical comparisons of normalized R&G features across two sets of groupings: 1) task outcome (successfully defusing the bomb vs. failure), and 2) player expertise (professional bomb defusers vs. amateurs).

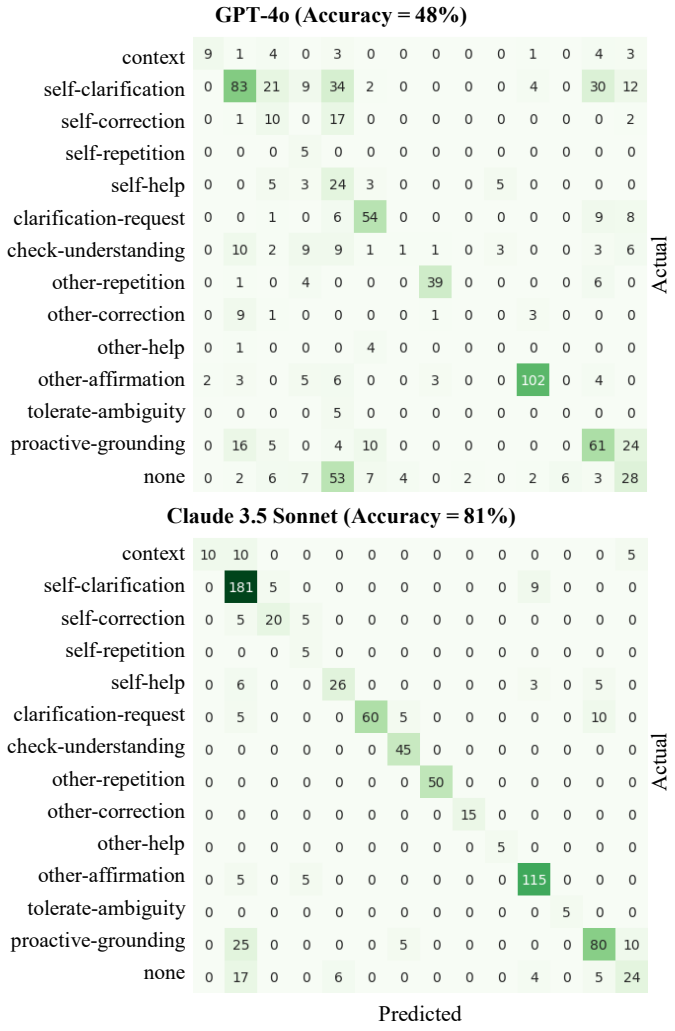

Fig. 2. Confusion matrices showing Claude 3.5 Sonnet provided more accurate labeling compared to GPT-4o ($T = 0$).

When comparing the frequency of R&G mechanisms between "Success" (successful bomb defusal) and "Failure" (bomb detonation) sessions, we observed several significant differences (Figure 3A). The frequency of *self-correction* was significantly lower in successful runs compared to failed runs (raw p-value = 0.020; adjusted p-value = 0.060), suggesting that more frequent self-correction may be indicative of communication difficulties that negatively affect team performance. In contrast, *other-repetition*, where one player repeats a statement made by the other, was significantly higher in successful runs (raw p-value = 0.010; adjusted p-value = 0.060). This pattern indicates that effective team coordination in successful trials may involve a higher reliance on repetition to confirm mutual understanding. Additionally, a significant interaction between *proactive-grounding* and *other-repetition* was observed. Specifically, the pairing of a *proactive-grounding* statement immediately followed by an *other-repetition* was found to be significantly more frequent in success runs (raw p-value = 0.020; adjusted p-value = 0.060). This suggests that in successful teams, *proactive grounding* is often reinforced by repetition, which may serve to solidify shared understanding during complex tasks.

In the comparison between "Professional" and "Amateur" groups, several patterns emerged (Figure 3B). Professionals exhibited significantly higher frequencies of *other-repetition* (raw p-value = $2.1 \cdot 10^{-6}$; adjusted p-value = $9.4 \cdot 10^{-6}$) and *proactive-grounding* (raw p-value = $7.4 \cdot 10^{-3}$; adjusted p-value = 0.022), indicating a more deliberate approach to ensuring shared understanding and task coordination. These behaviors may reflect the professionals' training in efficient communication under stress, where confirming and grounding information is critical to mission success. Conversely, the amateurs showed significantly higher frequencies of *self-repetition* (raw p-value = 0.040; adjusted p-value = 0.090) and *self-correction* (raw p-value = 0.051; adjusted p-value = 0.092). Finally, as observed in the outcome comparisons, the pairing of a *proactive-grounding* statement followed immediately by an *other-repetition* was significantly higher in the professional trials compared to the amateur trials (raw p-value = $2.1 \cdot 10^{-12}$; adjusted p-value = $1.9 \cdot 10^{-11}$). This reinforces the idea that trained professionals are more likely to use this specific interaction pattern, highlighting a communication strategy that may contribute to their superior task performance and reduced need for self-directed repair mechanisms.

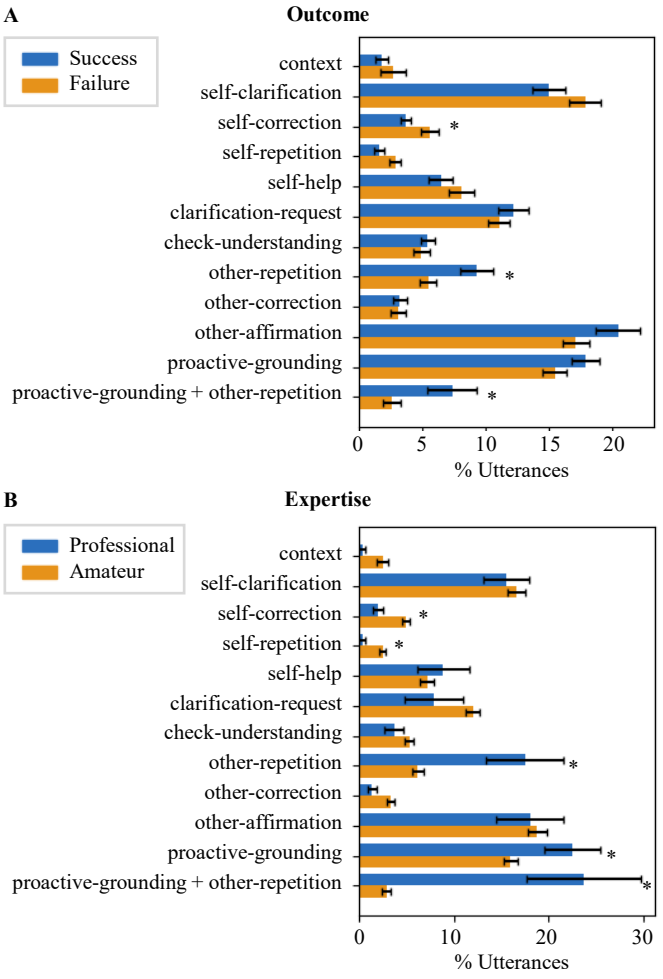

Fig. 3. Comparison of R&G mechanisms across outcome (A) and expertise groups (B). Error bars are standard error of the mean. Asterisks denote p-value < 0.1 after two-stage FDR Benjamini/Hochberg correction.

C. Model Evaluation Results

The model predicting outcome (success vs. failure) using the three most significant features (*other-repetition*, *self-correction*, and *proactive-grounding*), as well as puzzle difficulty, performed well, achieving a test set accuracy of 88%. The model demonstrated good generalization capabilities, with a mean cross-validation score of 74±12% across 5 stratified folds. This variance suggests some fluctuation in performance across different subsets of the data but remains within an acceptable range given the exploratory nature of this analysis. The precision and recall for predicting success were 80% and 100%, respectively, resulting in an F1-score of 89%. For failure predictions, precision was 100% and recall was 75%, with an F1-score of 86%. This indicates that the model was more conservative in predicting failure (i.e., fewer false positives) while being highly sensitive in identifying successful outcomes. The overall accuracy was consistent across both labels, with a weighted average F1-score of 87%. Bootstrap mean accuracy of 73±11% indicates moderate variability across resampled training sets. This suggests that while the model performed well on the test set, some sensitivity to sample variation exists, particularly given the relatively small sample size.

The model predicting expertise (professional vs. amateur) achieved higher performance, with an accuracy of 97% on the test set. Cross-validation also showed strong generalization, with a mean cross-validation score of 98±4%, demonstrating very little variation across folds. The model exhibited excellent precision for both professional and amateur players, with values of 100% and 97%, respectively. Recall for amateurs was 100%, but recall for professionals was lower at 67%, resulting in an F1-score of 80% for professionals and 98% for amateurs. The lower recall for professionals is attributable to the small sample size and the model's difficulty in perfectly identifying all professional cases. However, given the large imbalance in the dataset, the overall performance is strong, with a macro-averaged F1-score of 89%. The bootstrap analysis reinforced the robustness of the model, with an accuracy of 93±3% (mean ± S.D.), indicating high consistency across resampled datasets. A summary of model evaluation metrics is provided in Table 2.

TABLE II. SUMMARY METRICS FOR MODELS

Model Target	Test Set Accuracy	Cross-Validation	Bootstrap Accuracy
Outcome	88 %	72 ± 12 %	73 ± 11 %
Expertise	97 %	98 ± 4 %	93 ± 3 %

IV. DISCUSSION

A. Summary of Findings

This study set out to evaluate the use of conversational dynamics, specifically R&G mechanisms, as predictors of task performance and player expertise in a high-stakes teamwork task. By leveraging the cooperative game KTaNE, we explored how effective communication can influence success and distinguish between professional and amateur players. The major findings from both the labeling of R&G utterances and statistical analyses of team performance provide valuable insights into how communication mechanisms shape team outcomes.

Leveraging the R&G annotation technique used in [21], we found Claude 3.5 Sonnet demonstrated superior performance in labeling R&G, with consistently high accuracy across multiple temperatures. Its stability and precision at temperature = 0 make it a more reliable model for conversational analysis, particularly in the context of this study's goal to assess conversational factors predictive of task performance.

The statistical analyses reveal distinct patterns in communication between successful and unsuccessful bomb defusal trials, as well as between professional and amateur players. Teams that successfully defused bombs relied more heavily on *other-repetition* and *proactive-grounding*, mechanisms that ensure clear communication and reinforce shared understanding among team members. In contrast, unsuccessful teams exhibited higher frequencies of *self-correction*, suggesting that these teams may have experienced more internal communication difficulties or uncertainty, which could have contributed to their failures. Similarly, professional bomb defusers exhibited more frequent use of *other-repetition* and *proactive-grounding*, reinforcing the notion that team-oriented mechanisms are integral to maintaining cohesion and ensuring task success. This differentiation highlights the role of training and experience in shaping communication strategies, particularly in environments that demand precise and efficient interaction under pressure.

In terms of predictive modeling, both the outcome and expertise models demonstrated strong performance. The outcome model, which incorporated key R&G features along with puzzle difficulty, exhibited solid predictive accuracy (88%), though some variability was noted in cross-validation and bootstrap results. This suggests that while the model is reliable, the inclusion of additional contextual factors could further stabilize performance across different data subsets. The expertise model performed exceptionally well, achieving an accuracy of 97%. Despite the limited number of professional samples, the use of SMOTE to balance the training data proved effective, enabling the model to learn distinguishing patterns between professionals and amateurs. This result underscores the power of R&G frequencies in identifying expertise, even in datasets where class imbalance poses challenges.

B. Implications for Research and Practice

These findings offer several important implications for both research and practice. First, the practical use of AI-annotated conversational dynamics as a metric for evaluating teamwork is demonstrated through this study. Furthermore, the identification of specific R&G mechanisms—such as *proactive-grounding, other-repetition* and the sequential pairing of these—as predictors of success supports the notion that team-oriented communication is crucial in high-stakes environments. This suggests that training programs designed to improve team performance should focus on fostering these behaviors, emphasizing the importance of maintaining mutual understanding through repetition and grounding. Moreover, the promising predictive performance of this approach suggests that AI-driven tools could be developed to monitor team performance in real-team. These tools could provide feedback to trainees and supervisors, identifying potential communication breakdowns or deviations from optimal team behavior.

C. Limitations

While the study presents strong evidence for the utility of R&G mechanisms in predicting task performance and expertise, there are several limitations to consider. The truth data set used to evaluate LLMs' ability to annotate transcripts for R&G lacked representation of several class labels and was of limited size. While the results we obtained with Claude 3.5 Sonnet surpassed those obtained with GPT-4 in [21], we did not evaluate the dataset used in that paper. Thus, while we observed similar performance as [21] with GPT-4o, we cannot claim Claude 3.5 Sonnet would show similar performance gains on other datasets. Further validation is recommended across multiple datasets annotated from different sources to avoid annotator bias. LLMs were trained to annotate through simple prompting with example annotations. Fine-tuning models on larger annotated datasets could improve the robustness of this approach. In terms of predictive modeling, the small number of professional bomb defusers limited the generalizability of the expertise model, although the use of SMOTE mitigated some of these concerns. Future research should aim to expand the dataset to include a larger sample of professionals to validate and refine the model.

V. Conclusion

This study demonstrates how R&G mechanisms can be key indicators of successful teamwork and expertise in high-stakes tasks. Claude 3.5 Sonnet proved to be an effective tool for labeling utterances, providing a solid foundation for analysis. The statistical and modeling results underscore the importance of team-oriented communication strategies, such as *other-repetition* and *proactive-grounding*, in achieving success. These findings have practical implications for designing training programs and AI-driven tools that enhance teamwork performance in critical settings.

Acknowledgment

Supported by Independent Research and Development funds provided by Riverside Research's Open Innovation Center. The authors thank Riverside Research's Commercial Innovation Center for providing access to, and support for, Amazon Web Services.

References

[1] E. Salas, N. J. Cooke, and M. A. Rosen, "On Teams, Teamwork, and Team Performance: Discoveries and Developments," *Hum Factors*, vol. 50, no. 3, pp. 540–547, Jun. 2008, doi: 10.1518/001872008X288457.

[2] N. Power, "Extreme teams: Toward a greater understanding of multiagency teamwork during major emergencies and disasters," *American Psychologist*, vol. 73, no. 4, pp. 478–490, 2018, doi: 10.1037/amp0000248.

[3] P. Haddington and E. Stokoe, "Social interaction in high stakes crisis communication," *Journal of Pragmatics*, vol. 208, pp. 91–98, Apr. 2023, doi: 10.1016/j.pragma.2023.02.014.

[4] A. K. Kalia, N. Buchler, A. DeCostanza, and M. P. Singh, "Computing Team Process Measures From the Structure and Content of Broadcast Collaborative Communications," *IEEE Transactions on Computational Social Systems*, vol. 4, no. 2, pp. 26–39, Jun. 2017, doi: 10.1109/TCSS.2017.2672980.

[5] P. W. Foltz and M. J. Martin, "Automated communication analysis of teams," in *Team effectiveness in complex organizations*, Routledge, 2008, pp. 445–466.

[6] E. Salas, D. L. Reyes, and S. H. McDaniel, "The science of teamwork: Progress, reflections, and the road ahead.," *American Psychologist*, vol. 73, no. 4, pp. 593–600, May 2018, doi: 10.1037/amp0000334.

[7] K. Ernst *et al.*, "Training for High-Stakes Domains," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 67, no. 1, pp. 1356–1360, Sep. 2023, doi: 10.1177/21695067231192413.

[8] X. Zhang, H. Yu, Y. Li, M. Wang, L. Chen, and F. Huang, "The Imperative of Conversation Analysis in the Era of LLMs: A Survey of Tasks, Techniques, and Trends," Sep. 21, 2024, *arXiv*: arXiv:2409.14195. doi: 10.48550/arXiv.2409.14195.

[9] P. Soares, S. McCurdy, A. J. Gerber, and P. Fonagy, "Chatting Up Attachment: Using LLMs to Predict Adult Bonds," Aug. 31, 2024, *arXiv*: arXiv:2409.00347. doi: 10.48550/arXiv.2409.00347.

[10] F. Gervits, K. Eberhard, and M. Scheutz, "Team Communication as a Collaborative Process," *Front. Robot. AI*, vol. 3, Oct. 2016, doi: 10.3389/frobt.2016.00062.

[11] M. Baker, T. Hansen, and R. Joiner, "The Role of Grounding in Collaborative Learning Tasks," in *Collaborative Learning: Cognitive and Computational Approaches*, 1999.

[12] S. Larsson, "Grounding as a Side-Effect of Grounding," *Topics in Cognitive Science*, vol. 10, no. 2, pp. 389–408, 2018, doi: 10.1111/tops.12317.

[13] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[14] H. H. Clark and E. F. Schaefer, "Contributing to discourse," *Cognitive Science*, vol. 13, no. 2, pp. 259–294, Apr. 1989, doi: 10.1016/0364-0213(89)90008-6.

[15] S. Albert and J. P. de Ruiter, "Repair: The Interface Between Interaction and Cognition," *Topics in Cognitive Science*, vol. 10, no. 2, pp. 279–313, 2018, doi: 10.1111/tops.12339.

[16] E. Schegloff, G. Jefferson, and H. Sacks, "The Preference for Self-Correction in the Organization of Repair in Conversation," *Language*, vol. 53, pp. 361–382, Jun. 1977, doi: 10.2307/413107.

[17] E. Schegloff, "When 'others' initiate repair," *Applied Linguistics*, vol. 21, no. 2, pp. 205–243, Jun. 2000, doi: 10.1093/applin/21.2.205.

[18] P. Drew, "'Open' class repair initiators in response to sequential sources of troubles in conversation," *Journal of Pragmatics*, vol. 28, no. 1, pp. 69–101, Jul. 1997, doi: 10.1016/S0378-2166(97)89759-7.

[19] L. Rudge, "'I Cut It and I… Well Now What?'(Un) Collaborative Language in Timed Puzzle Games," *Approaches to Videogame Discourse: Lexis, Interaction, Textuality*, pp. 178–200, 2019.

[20] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 1877–1901.

[21] X. Zhang, R. Divekar, R. Ubale, and Z. Yu, "GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 300–314. doi: 10.18653/v1/2023.bea-1.26.

[22] R. Glas, "Vicarious play: Engaging the viewer in Let's Play videos," *Empedocles: European Journal for the Philosophy of Communication*, vol. 5, no. 1–2, pp. 81–86, 2015.

[23] E. A. Schegloff and H. Sacks, "Opening up Closings," vol. 8, no. 4, pp. 289–327, Jan. 1973, doi: 10.1515/semi.1973.8.4.289.

[24] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[25] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, vol. 93, no. 3, pp. 491–507, Sep. 2006, doi: 10.1093/biomet/93.3.491.

[26] Josef Perktold *et al.*, *statsmodels/statsmodels: Release 0.14.2*. (Apr. 17, 2024). Zenodo. doi: 10.5281/ZENODO.593847.

[27] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S.-Y. Bang, "Support Vector Machine Ensemble with Bagging," in *Pattern Recognition with Support Vector Machines*, S.-W. Lee and A. Verri, Eds., Berlin, Heidelberg: Springer, 2002, pp. 397–408. doi: 10.1007/3-540-45665-1_31.

[28] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[30] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.