Large Language Models for Automated Grading and Synthetic Data Generation in Communication-Based Training Assessment

Joseph P. Salisbury
Riverside Research, Melbourne, FL
jsalisbury@riversideresearch.org

David M. Huberdeau
Riverside Research, Lexington, MA
dhuberdeau@riversideresearch.org

Abstract

Effective communication is critical in high-stakes tasks, particularly in scenarios requiring precision and coordination under time pressure. Here, we explore the potential of large language models (LLMs) to evaluate communication performance and generate synthetic conversation data for training and assessment purposes. We present a proof-of-concept study focused on a highly structured task: the interaction between a forward observer and a fire direction center during a call for fire mission. Using a rubric-based approach, the LLM graded transcripts of forward observer communications, distinguishing between varying levels of trainee performance with high reliability and alignment to expected outcomes. Additionally, we demonstrate the utility of LLMs in generating synthetic transcripts that simulate varying performance levels. While this study is centered on the call for fire, the approach has broader implications for training assessment in complex, communication intensive tasks. Our results suggest that LLMs can serve as effective tools for both grading and data generation, enabling scalable solutions for improving performance in high-stakes domains.

Introduction

Effective communication is critical in high-stakes domains where coordination, precision, and timeliness can directly impact outcomes. Fields such as military operations, healthcare, aviation, and emergency response require individuals and teams to exchange information under conditions of uncertainty, time pressure, and complex task demands. In these contexts, communication failures are consistently cited as contributing factors to errors and adverse events (Salas, Sims, and Burke 2005; Patterson et al. 2004). Consequently, assessing and improving communication performance has become a central focus in training programs for these domains.

In the military, for example, the successful execution of tasks such as fire support operations relies on precise and structured communication protocols (Department of the Army 2017). Forward Observers (FOs), as part of a Fire Support Team (FiST), must convey critical information about targets, locations, and firing adjustments to the Fire Direction Center (FDC). These exchanges must adhere to established procedures, such as the "call for fire" (CFF) protocol, to minimize errors that could result in missed targets or unintended collateral damage. Similarly, in other high-stakes settings, communication protocols like surgical safety checklists or air traffic control phraseology are used to standardize information exchange and reduce variability in performance (Lingard et al. 2004; Cushing 1994).

Despite the importance of communication, assessing its quality during training remains a significant challenge. Traditional assessment methods often rely on human evaluators using standardized rubrics to observe and score performance. While these methods provide valuable insights, they are inherently limited by subjectivity, variability between raters, and scalability issues in large training programs (Downing 2004). These challenges are particularly acute in high-stakes domains where training must simulate complex, high-pressure scenarios to prepare individuals for real-world tasks. Automating the evaluation process offers a potential solution, reducing the burden on human evaluators while providing consistent, objective feedback.

Recent advancements in artificial intelligence, particularly large language models (LLMs) such as GPT-4, have demonstrated capabilities in natural language understanding, text evaluation, and content generation, making them suitable candidates for assessing structured communication tasks. These developments suggest a potential paradigm shift in how communication training and assessment are approached, with implications for scalability, reliability, and accessibility. Thus, to address the challenges associated with assessing communication performance, this study investigates the use of LLMs as tools for automating communication evaluation.

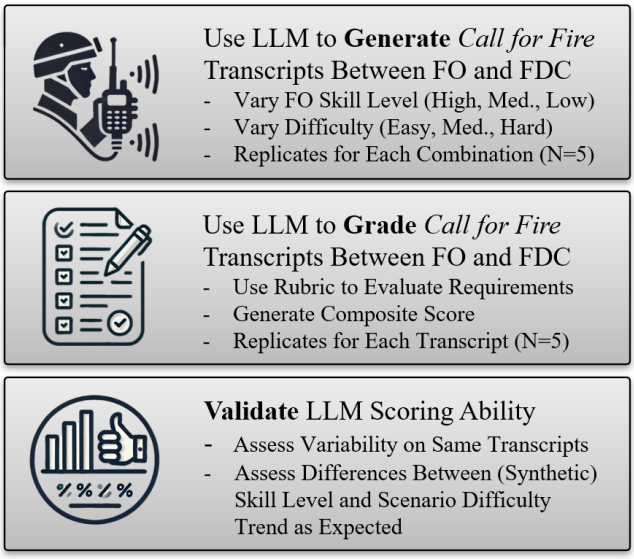

Figure 1: Overview of steps to demonstrate automated scoring.

To demonstrate this, we selected the CFF as it is both highly structured, adhering to established communication protocols, and mission-critical, with significant consequences for errors. Our study examines two complementary capabilities of LLMs: automated grading and synthetic data generation (Figure 1). In automated grading, the LLM is tasked with evaluating FO communications against a predefined rubric. This rubric assesses key aspects of communication performance, including adherence to protocol, clarity, and accuracy. By analyzing the consistency and sensitivity of LLM-generated scores across different performance levels and scenarios, our objective is to provide support that LLMs can reliably grade structured communication tasks. For this proof-of-concept, an LLM is also used to generate synthetic transcripts simulating FOs of various skill competency levels. These transcripts enable the development and testing of automated assessment methods, providing a scalable alternative to real-world data, which can be difficult and time-consuming to collect.

Related Works

LLMs are increasingly influencing educational practices, particularly through their potential to automate assessments and provide personalized feedback. Numerous studies have explored the integration of LLMs into essay evaluation processes, highlighting their potential to enhance grading efficiency and consistency (Ishida et al. 2024; Golchin et al. 2024; Katuka, Gain, and Yu 2024; Xiao et al. 2024). For instance, Ishida et al. examined LLMs' performance in assessing student essays across various scenarios, finding a strong correlation between LLM and faculty assessments. Similarly, Golchin et al. investigated the feasibility of utilizing LLMs to replace peer grading in massive open online courses. Their findings indicated that LLMs, particularly when guided by instructor-provided answers and rubrics, produced grades more aligned with those assigned by instructors compared to traditional peer grading methods. While these studies underscore the potential of LLMs in educational assessment, they primarily focus on general academic tasks such as essay grading and programming assignments. There remains a gap in research concerning the application of LLMs to highly structured, domain-specific communication critical to high-stakes tasks.

LLMs have also significantly enhanced the field of conversational analysis, particularly in evaluating and simulating dialogues. Studies have demonstrated the efficacy of LLMs in assessing dialogue quality across multiple dimensions, offering a unified approach to evaluating open-domain conversations (Lin and Chen 2023). LLMs have been employed to simulate user interactions within conversational recommender systems, serving as cost-effective proxies for real users in system evaluation (Yoon et al. 2024). LLMs have also been utilized to generate synthetic dialogues for training purposes, enabling the creation of high-quality conversational data without the ethical and privacy concerns associated with human data collection (Chen et al. 2023). While these studies underscore the potential of LLMs in conversational analysis, they primarily focus on open-domain dialogues and general conversational systems. In contrast, we explore here the application of LLMs to highly structured, domain-specific communication tasks that are critical in high-stakes environments.

The importance of CFF training has led to various dialogue systems and natural language processing (NLP) techniques to be developed to enhance military training prior to the emergence of LLMs. Early efforts include the development of Radiobot-CFF, a spoken dialogue system designed to engage in CFF radio dialogues, assisting soldiers in mastering artillery fire request procedures (Roque et al. 2006; Roque and Traum 2006). This system utilized an information-state dialogue manager to manage interactions, aiming to provide realistic training scenarios. Similarly, the IF-Soar agent was introduced to simulate the role of the FDC in processing and coordinating CFF information from an FO (Stensrud, Taylor, and Crossman 2006). By emulating FDC responses, IF-Soar provided trainees with a comprehensive understanding of the CFF process, enhancing their readiness for real-world operations. These initiatives highlight the application of dialogue systems and NLP in military training, particularly concerning the CFF. However, they primarily rely on rule-based approaches and lack the adaptability inherent in modern LLMs. Our study addresses this gap by leveraging LLMs to assess and generate synthetic data for CFF training, offering a more flexible and scalable solution to enhance the effectiveness of military communication training programs.

Methods

Task Overview: The Call for Fire

In modern military operations, indirect fire support plays a critical role in achieving tactical objectives. Indirect fire involves using artillery, mortars, or other ranged weaponry to engage targets that are beyond the line of sight of the firing unit. To ensure accuracy and minimize risk, this process requires precise coordination between multiple roles within a FiST. Two key roles are the FO and FDC. The FO is a trained specialist responsible for identifying and locating enemy targets from a forward position, typically closer to the battlefield than the firing unit. The FDC is the central hub that processes and coordinates all fire mission requests from FOs. The FO operates under potentially hostile conditions, requiring both technical expertise and the ability to communicate effectively under pressure. Their role is essential for ensuring that indirect fire is delivered accurately and without unintended collateral damage.

The CFF is the structured communication process by which an FO requests indirect fire support and includes six key elements (Figure 2), which are communicated to the FDC across three transmissions. When ready, the FDC responds with a Message to Observer (MTO), detailing the ammunition and guns to be used. Adjusting rounds are fired to verify accuracy, and the FO provides incremental adjustments to the FDC until the fire is accurate on target and ready to fire for effect. This protocol ensures all necessary information is communicated effectively. Errors, omissions, or ambiguities in this process can lead to mission failure or unintended consequences.

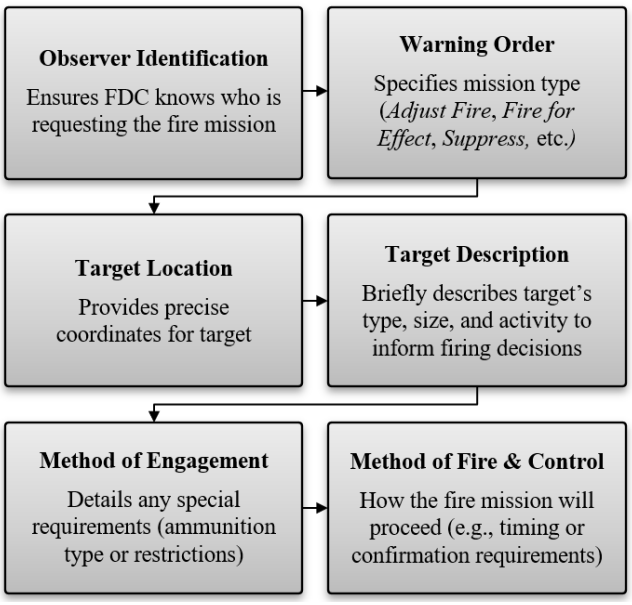

Figure 2: Essential components of a Call for Fire.

Synthetic Call for Fire Generation

To evaluate the potential of LLMs to generate realistic and varied data for evaluation, synthetic CFF transcripts were created using OpenAI's Assistants API. An FO assistant (henceforth, "agent") was defined using the GPT-4o model and the instructions "You are a U.S. Marine Corps training assistant. Use your knowledge base to perform tasks related to procedures a Forward Observer (FO) should be well trained in." To ground the agent, a relevant training manual (U.S. Marine Corps 2015) was made available to the agent via the file_search tool. To generate transcripts, the agent was provided with a detailed task prompt that included:

1. **Task Description** – A summary of the task to generate a realistic transcript of a CFF dialogue between an FO and FDC based on a scenario and FO description.
2. **Scenario Description** – Each prompt included a brief descriptive scenario to provide context for the fire mission. Scenarios varied in difficulty (Easy, Medium, Hard) and included details such as:
    - **Target type and location:** E.g., "A mechanized infantry platoon is positioned in an open field at grid coordinates AB 1234 5678."
    - **Environmental Conditions:** E.g., "The weather is clear with good visibility" or "Heavy fog limits visibility; target identification is challenging."
    - **Mission Objectives:** E.g., "Adjust fire to register the artillery" or "Fire for effect to neutralize the target."
3. **Performance Level** – Prompts specified whether the FO was a highly trained professional or varying level of trainee:
    - **High:** The FO is an experienced operator with significant battlefield experience.
    - **Medium:** The FO is a trainee who has successfully completed some basic training exercises but still makes some mistakes.
    - **Low:** The FO is a new trainee attempting the procedure for the first time.
4. **CFF Protocol Guidance** – Brief reminders of the structured communication process, including essential components of the CFF, the sequence they should be transmitted in, how the FDC should respond to the CFF, and adjustment procedures.
5. **Example Transcript** – A high-quality example was provided to illustrate desired response format.

For each combination of the three scenario difficulty levels and performance levels, five transcripts were generated, resulting in 45 synthetic transcripts total. The transcripts were manually reviewed to ensure alignment with procedure standards and scenario expectations. Subject matter experts (U.S. veterans with FO training) were consulted to provide feedback to refine the prompt until generated transcripts reached an acceptable level of procedural accuracy.

LLM Evaluation of Call for Fire Dialogues

To evaluate the synthetic CFF transcripts, the same FO agent configuration used for transcript generation was employed to grade the transcripts. For each generated transcript, the agent was tasked in a prompt with applying a structured rubric to assess FO performance. The grading rubric consisted of six items, each scored on a scale from 0 to 5. These items were chosen to evaluate key aspects of CFF communication and procedural adherence. While somewhat arbitrary, the rubric was designed to illustrate the LLM's capability to evaluate communication tasks systematically. These items included:

1. **Adherence to the CFF Three-Transmission Format** – Assessed whether the FO followed the standard structure of the CFF, including observer identification and mission type, target location, and target description/method of engagement/method of fire control.
2. **Clarity and Accuracy of Communication** – Evaluated the FO's ability to communicate clearly and concisely, avoiding unnecessary filler words, and ensuring accuracy in provided information.
3. **Performance During the MTO Phase** – Scored based on the inclusion of all required elements in the MTO (e.g., units to fire, number of rounds, and target number), and whether the FO repeated the MTO back verbatim.
4. **Execution of Fire Adjustments** – Measure the quality and precision of adjustments provided by the FO, including timely and logical corrections during the adjustment phase.
5. **Adherence to Protocol and Professionalism** – Assessed whether the FO adhered to military protocol, used proper brevity codes, and maintained professionalism throughout the interaction.
6. **Timing and Responsiveness** – Evaluated the promptness and appropriateness of the FO's responses to the FDC.

Each transcript received a composite score out of 30, calculated as the sum of the scores across these six items.

For each of the 45 synthetic transcripts, the agent generated five independent score reports, resulting in a total of 225 evaluations. The task prompt included the transcript to be scored, a detailed explanation of the scoring rubric with criteria and examples for each item, and instructions to provide a structured score report in a consistent format, including: a breakdown of scores for each rubric item; a composite score; and a comments section to provide a brief summary of the FO's performance, including strengths, areas for improvement, and any critical errors that impacted the mission. The FO agent retained access to the Call for Indirect Fire manual to ensure that scoring adhered to military standards and protocols. The agent was instructed to consult the manual for clarification if uncertainties arose about the proper format or procedural expectations.

Data Analysis

Analysis focused on evaluating the reliability, consistency, and validity of LLM evaluations. Composite scores were extracted from the transcripts' evaluation files and stored with metadata such as scenario difficulty, performance level, and replicate identifiers.

To assess the reliability of LLM evaluations, the intraclass correlation coefficient (ICC) was used to measure the consistency of scores within transcript replicates (Koo and Li 2016). In the particular, ICC(2,1) was used, which evaluates absolute agreement for single random raters under the assumption that each replicate represents an independent application of the grading rubric. To determine if the LLM could differentiate between high-, medium-, and low-performing FOs, a one-way analysis of variance (ANOVA) was conducted across the three performance levels, with post-hoc pairwise comparisons performed using Tukey's Honestly Significant Difference (HSD) test to identify specific differences between groups. The effect of scenario difficulty (Easy, Medium, Hard) on composite scores was examined to assess the LLM's sensitivity to contextual complexity. One-way ANOVA followed by Tukey's HSD was again used to analyze the scenario difficulty effects within performance levels. All data analysis was performed using Python, leveraging libraries such as pandas for data manipulation, SciPy for statistical tests, and statsmodels for ANOVA analyses. ICC calculations were conducted using the Pingouin library (Vallat 2018).

Results

Synthetic Data Generation

The following transcript illustrates a high-performing FO conducting a fire mission during an "Easy" scenario:

FO: "FDC, this is Alpha Three-One. Adjust fire, over."
FDC: "Alpha Three-One, this is FDC. Adjust fire, out."
FO: "Grid: Alpha Bravo 1234 5678, over."
FDC: "Grid: Alpha Bravo 1234 5678, out."
FO: "Target is a mechanized infantry platoon in the open. Request high explosive, fire when ready, over."
FDC: "Target is a mechanized infantry platoon in the open. Request high explosive, fire when ready, out."
FDC: "Message to observer, R6G adjusting, A8T in effect, three rounds per tube, target number AB 5612, over."
FO: "Message to observer, R6G adjusting, A8T in effect, three rounds per tube, target number AB 5612, out."
FO: "Direction 2680, over."
FDC: "Direction 2680, out."
FDC: "Shot, over."
FO: "Shot, out."
FDC: "Splash, over."
FO: "Splash, out."

FO: "Left 50, add 100, over."
FDC: "Left 50, add 100, out."
FDC: "Shot, over."
FO: "Shot, out."
FDC: "Splash, over."
FO: "Splash, out."
FO: "Right 20, drop 50, fire for effect, over."
FDC: "Right 20, drop 50, fire for effect, out."
FDC: "Shot, over."
FO: "Shot, out."
FDC: "Splash, over."
FO: "Splash, out."
FO: "End of mission, mechanized infantry platoon neutralized, estimate 80% casualties, over."
FDC: "End of mission, mechanized infantry platoon neutralized, 80% casualties, out."
FO: "Alpha Three-One, out."
FDC: "Roger, Alpha Three-One, out."

This transcript exemplifies a well-trained FO executing a CFF following the established communication protocol, demonstrating adherence to standard procedures and reflects the operational behavior expected of a skilled FO.

In contrast, the low-performing FO transcripts highlight several common errors and shortcomings that undermine effective communication during a CFF mission, including lack of clarity and confidence, incomplete and incorrect information, difficulty following protocol, inadequate target assessment, and excessive reliance on the FDC for guidance. These errors are representative of inexperienced operators struggling to follow protocol. By simulating such interactions, the transcripts provide valuable examples for performance assessment methods development.

Automated Grading

The agent reliably produced score reports in the desired format, and inspection revealed clear and systematic evaluation of FO performance. For example, in response to a low-performing FO, the evaluation stated:

1. Adherence to the Three-Transmission Format: *The FO adhered to the three-transmission format but showed hesitancy and lack of clarity in the transmissions. Score: 2/5 | Reason: The FO confirmed the grid eventually but with uncertainty and confusion in the target description and method of engagement.*
2. ***Clarity and Accuracy of Communication****: Communication was not consistently clear; the FO frequently used hesitation phrases like "uh," which hindered clear transmission. There was also uncertainty in the target grid initially given. Score: 3/5 | Reason: Deducted for lack of clarity and initial target grid uncertainty.*
3. Correctness and Completeness of the MTO: *MTO was largely correct and repeated accurately, but with filler words. Score: 4/5 | Reason: Almost complete but affected by the communication style.*
4. Execution of Adjust Fire: *The FO made correct corrections after the initial spotting but initially omitted proper direction reporting, which is needed before giving corrections. Score: 3/5 | Reason: Errors in timing and order of operations.*
5. Adherence to Protocol and Professionalism: *The FO did not fully maintain professionalism; the excessive use of fillers indicated a lack of confidence and deviation from communication protocol. Score: 2/5 | Reason: Detracted from protocol adherence due to unprofessional language.*
6. Timing and Responsiveness: *The FO did respond timely, although somewhat hesitantly at times, with confirmation and replies to the FDC. Score: 4 out of 5 | Reason: Mostly prompt but influenced by the lack of confidence in responses.*
Composite Score: *18/30*
Comments: *The FO demonstrated a basic understanding of procedures but struggled with clear communication and protocol adherence, affecting overall performance. Key areas of improvement include eliminating hesitation in communications, providing clear and confident descriptions during target engagement, and ensuring proper order and timing in transmissions. Additional training in focusing on clear, professional dialogue and reinforcing confidence in procedural execution is recommended. The initial uncertainty with grid transmission and lack of strong communication detracted from mission effectiveness.*

The example output demonstrates the agent's ability to provide a structured and detailed assessment of FO performance based on the rubric. The agent's scoring and comments closely align with the observed deficiencies in the low-performing transcript. The evaluation effectively breaks down performance into the components, providing both numeric scores and qualitative reasoning for each.

Reliability of LLM Scores

To evaluate the reliability of the agent scoring transcripts, the ICC was calculated across scoring replicates for each transcript, revealing a high degree of reliability (ICC(2,1)=0.934, 95% CI: [0.9, 0.96]; $p<0.001$). This result indicates excellent agreement across replicates, demonstrating the agent consistently evaluated transcripts despite the independence of scoring replicates.

ANOVA was used to determine whether agent scores differed significantly across the three performance levels. The results revealed a highly significant main effect of performance on scores ($p<0.001$), indicating that the agent was sensitive to variations in FO performance. Post-hoc results showed significant pairwise differences between all performance levels, which trend as expected (Figure 3A).

To evaluate whether scenario difficulty (Easy, Medium, Hard) influenced scores within each performance level (High, Medium, Low), one-way ANOVAs were conducted for each performance level separately. The results revealed significant effects for the high and medium performance groups, but no significant effect for the low performance group (Figure 3B). ANOVA for the high group revealed a statistically significant effect of scenario difficulty on scores (p=0.023). Post-hoc pairwise comparisons showed a significant difference between easy and hard scenarios (mean difference = -0.48, p=0.035), with lower scores observed in the Hard scenario. For the medium group, the ANOVA also revealed a significant effect of scenario difficulty on scores (p=0.029), with post-hoc tests showing significant differences between easy and hard scenarios (mean difference = -1.8, p=0.35). The results indicate that scenario difficulty influenced scores differently across performance levels, with medium performers showing a more pronounced sensitivity to scenario difficulty.

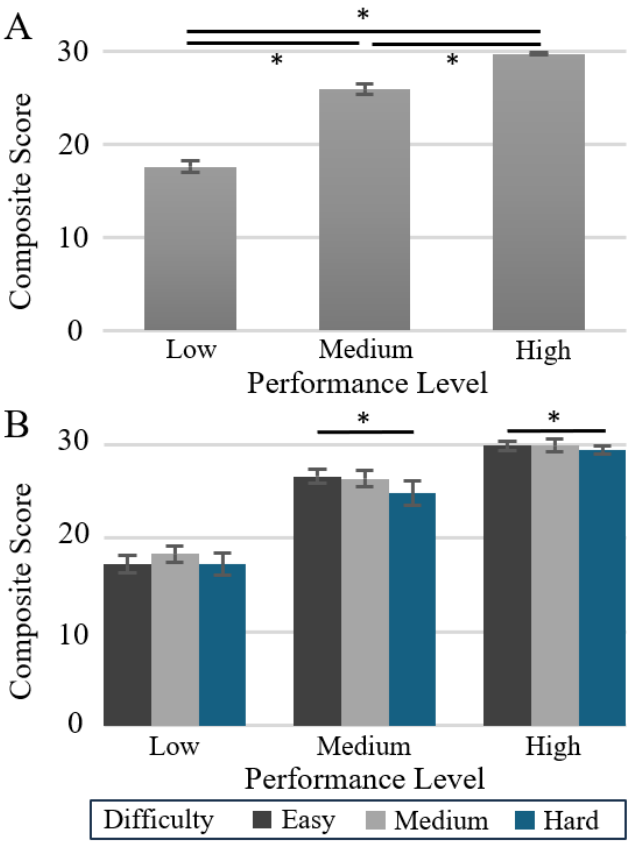

Figure 3: A. Mean composite scores consistently increase with simulated performance level (* adjusted p < .05, Tukey-HSD). B. Medium and high performing FOs score significantly lower on hard scenarios when compared with easy scenarios (* adjusted p < .05, Tukey-HSD). Error bars are 95% CI.

Discussion

This study demonstrates the feasibility of using an LLM agent to evaluate structured communication tasks. The agent successfully generated realistic transcripts representing varying performance levels and evaluated these transcripts using a rubric. The agent-produced scores exhibited excellent reliability. The agent effectively distinguished between FOs of varying performance levels, with significant score differences aligning with expected trends.

While we achieved our primary objective of demonstrating the overall feasibility of this approach, there are several notable limitations. While the synthetic transcripts were generally realistic, certain nuances of real-world communication, such as environment stressors, were not fully captured. This is particularly evident in the lack of performance degradation for low performers across scenario difficulty levels, potentially due to a floor effect in how transcripts were generated and/or scored. Scenario complexity could be improved by introducing dynamic and adaptive scenario generation to better simulate real-world challenges. The grading rubric was designed to illustrate the feasibility of using LLM agents for automated assessment of structured communication tasks. While the rubric provided a systematic framework, its design is inherently arbitrary and not validated against human expert evaluations. Future work should involve refining the rubric in collaboration with subject matter experts to ensure greater alignment with real-world performance metrics. Finally, our study focused solely on FO communications. While this provides a starting point, the broader interactions with a FiST involve additional complexities, including decision-making tasks that are less inherently structured than the CFF. By expanding on this proof-of-concept and incorporating additional FiST roles, this would enable evaluation of team-level performance and coordination.

Conclusion

This study provides a foundation for leveraging LLM agents to evaluate structured communication tasks in high-stakes domains. The demonstrated reliability and sensitivity to performance levels highlight the potential for LLM agents to transform training evaluations. By addressing limitations, this approach has the potential to significantly enhance readiness assessment and training effectiveness.

Acknowledgements

The authors acknowledge the assistance of OpenAI's ChatGPT 4o in helping phrase sections of the manuscript and assisting with code to analyze data.

References

Chen, M., Papangelis, A., Tao, C., Kim, C. Rosenbaum, A., Liu, Y., Yu, Z., and Hakkani-Tur, D. 2023. "PLACES: Prompting Language Models for Social Conversation Synthesis." arXiv.

Cushing, S. 1994. *Fatal Words: Communication Clashes and Aircraft Crashes*. University of Chicago Press.

Department of the Army. 2017. "Observed Fires."

Downing, S. M. 2004. "Reliability: On the Reproducibility of Assessment Data." *Medical Education* 38 (9): 1006–12.

Golchin, S., Garuda, N., Impey, C., and Wenger, M. 2024. "Grading Massive Open Online Courses Using Large Language Models." arXiv.

Ishida, T., Liu, T., Wang, H. and Cheung, W. K. 2024. "Large Language Models as Partners in Student Essay Evaluation." arXiv.

Katuka, G. A., Gain, A., and Yu, Y. 2024. "Investigating Automatic Scoring and Feedback Using Large Language Models." arXiv.

Koo, T. K. and Li, M. Y. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–63.

Lin, Y., and Chen, Y. 2023. "LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models." arXiv.

Lingard, L., Espin, S., Whyte, S., Regehr, G., Baker, G. R., Reznick, R., Bohnen, J., Orser, B., Doran, D., and Grober, E. 2004. "Communication Failures in the Operating Room: An Observational Classification of Recurrent Types and Effects." *BMJ Quality & Safety* 13 (5): 330–34.

Patterson, E. S., Roth, E. M., Woods, D. D., Chow, R., and Gomes, J. O. 2004. "Handoff Strategies in Settings with High Consequences for Failure: Lessons for Health Care Operations." *International Journal for Quality in Health Care* 16 (2): 125–32.

Roque, A., Leuski, A., Rangarajan, V., Robinson, S., Vaswani, A., Narayanan, S., and Traum, D. 2006. "Radiobot-CFF: A Spoken Dialogue System for Military Training." In *Interspeech 2006*, paper 1828-Mon2FoP.8-0. ISCA.

Roque, A., and Traum, D. 2006. "An Information State-Based Dialogue Manager for Call for Fire Dialogues." In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, edited by Jan Alexandersson and Alistair Knott, 88–95. Sydney, Australia: Association for Computational Linguistics.

Salas, E., Sims, D. E., and Burke, C. S. 2005. "Is There a 'Big Five' in Teamwork?" *Small Group Research* 36 (5): 555–99.

Stensrud, B., Taylor, G., and Crossman, J. 2006. "IF-Soar: A Virtual, Speech-Enabled Agent for Indirect Fire Training." In *25th Army Science Conference, Orlando, FL*.

U.S. Marine Corps. 2015. "Call for Indirect Fire B2C2497 Student Handout."

Vallat, R. 2018. "Pingouin: Statistics in Python." *Journal of Open Source Software* 3 (31): 1026.

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., and Fu, Q. 2024. "Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs." arXiv.

Yoon, S, He, Z., Echterhoff, J., and McAuley, J.. 2024. "Evaluating Large Language Models as Generative User Simulators for Conversational Recommendation." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by Kevin Duh, Helena Gomez, and Steven Bethard, 1490–1504. Mexico City, Mexico: Association for Computational Linguistics.