CliniDial: A Naturally Emerged Multimodal Dialogue Dataset for Team Reflection During Clinical Operation

Anonymous ACL submission

Abstract

In clinical operations, teamwork can be the crucial factor that determines the final outcome. Prior studies have shown that there can be 58% more deaths than expected due to insufficient collaboration. To understand how the team practices teamwork during the operation, we collected ***CliniDial*** from simulations of medical operations. ***CliniDial*** includes the audio data and its transcriptions, the simulated physiology signals of the patient manikins, and how the team operates from two camera angles. We annotated behavior codes following an existing framework to understand the teamwork process. Experimental results show that ***CliniDial*** poses significant challenges to the existing models.

1 Introduction

In clinical settings, teamwork is crucial for a successful operation, and effective team collaboration can improve the safety and well-being of the patients (Catchpole et al., 2008; Weaver et al., 2010; Schmutz et al., 2019; Rosen et al., 2018). Failures in teamwork and communication among healthcare providers are a major contributing factor to the estimated 250,000 preventable deaths that occur in the U.S. each year (Rosen et al., 2018; Makary and Daniel, 2016). Breakdowns in areas like leadership, situation awareness, decision-making and communication frequently underlie the many forms of preventable patient harm, including hospital infections, falls, diagnostic errors and surgical mistakes (Baker et al., 2005; Herzberg et al., 2019; Keers et al., 2013). There can be 58% more deaths than expected due to insufficient collaboration (Knaus et al., 1986). Motivated by these statistics, in this paper we model the communication between team members as well as the data in the operation room to detect the effective steps and interactions needed for a successful procedure.

To understand how teamwork unfolds in the operating room, we collected ***CliniDial*** from simulations of medical operations. We collected the audio data, simulated physical signals from the patient manikins, as well as how the team operates from two camera angles. We then annotated behavior codes based on a team reflection behavior framework (Schmutz et al., 2021) to understand how the team members convey their objectives, strategies, and actions during the operation. We test various baseline methods with different setups. Experimental results show that ***CliniDial*** poses significant challenges to existing methods. In addition, we invite input from medical professionals to try to bridge the current NLP fields with the real-world applications they expect (Appendix E).

In summary, our contributions are two folds:

1. We present ***CliniDial***, a naturally emerged multimodal dialogue dataset for team reflection during clinical operation.

2. We evaluate our dataset against various existing methods with different setups and provide a detailed analysis of their results. Our experimental results reveal that our dataset poses significant challenges to existing methods, urging methodology innovation in our NLP community.

2 How is *CliniDial* Different?

Our real-world setting distinguishes ***CliniDial*** from existing datasets in various aspects. First, there are significant **label imbalances** in the collected data. Such label imbalances are less common in conventional NLP datasets where researchers have some levels of control over the data distribution by data filtering or downsampling. However, since our dialogues occur naturally in the operation room, the interlocutors are not tasked to generate dialogues but rather to perform the clinical operation and take care of the "patient" as a team. We do not pose any constraints on how the team communicate, and we observe that the amount of majority class labels sig-

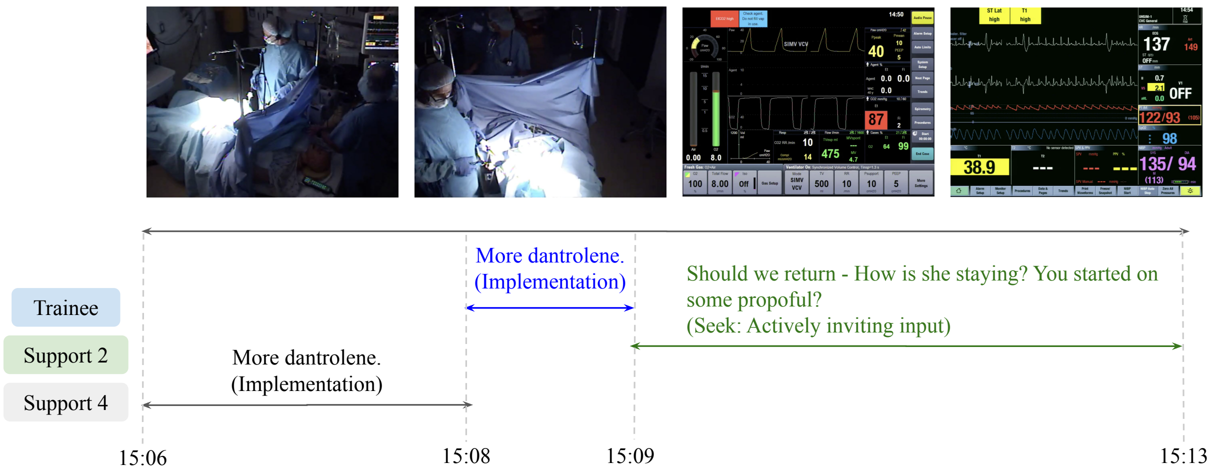

Figure 1: An example of the labeled dialogue in the simulated operation. Two cameras capture the scenes from two angles and two real-time monitoring systems provide the patient's physiological signals. We only include the trainee and the two supports in this example, as they are the only three people speaking during this time frame.

nificantly outmatches the minority class labels. Second, there are **rich and natural interactions** between the team members. Compared to the conventional dialogue benchmarks (Budzianowski et al., 2018) which typically contain 30 turns at most, the dialogue in our collected dataset contains 311 turns on average. Third, there are **rich modalities** in the collected data. Compared to the conventional NLP datasets with text modality (Chen et al., 2021) or the conventional multimodal datasets which focus on vision and text modalities (Tapaswi et al., 2016; Lei et al., 2018; Castro et al., 2022), the data we collect includes not only the dialogue, but also the corresponding audio, the operation views from two camera angles, and the physiological signals from the "patient" aligned for each timestamp.

3 Dataset

3.1 Data Descriptions

Scenarios. A team of board certified anesthesiologists together with support staff is tasked with the intraoperative management of a 36-year-old female who is undergoing a minimally invasive surgery [1]. This scenario takes place in a simulated operating room where we present a mannequin as the female patient and simulate her physiological signal changes from the backend. Specifically, the patient develops malignant hyperthermia (MH; a rare complication of general anesthesia that could develop in any patient) as the simulated scenario progresses. Many healthcare providers lack sufficient clinical exposure to MH, potentially hindering their ability to recognize, treat, and manage these rare but severe cases effectively (Isaak and Stiegler, 2016). We want to stress that this is not a real operation, and the intent is to train medical trainees in "near-life" surgical operations.

Roles. In the simulated operation, a confederate plays the role of the surgeon. The trainee who serves as the anesthesiologist is the main decision-maker [2]. The support participants are also trainees who support an anesthesiologist. Appendix B provides additional details of the simulated operation and the roles of the team members.

3.2 Labels

Following Schmutz et al. (2021), we include three labels of "Seek", "Evaluate" and "Plan" detailed in Table 3 in Appendix A. As our data is sourced from clinical operations, we are interested in not only how the teams engage in reflection or diagnostic behaviors, but also how the team progresses from diagnostic actions to interventions or implementation actions. Therefore, we assign an extra label "Implement" to such behaviors. Appendix A provides more details for each label. We describe the details of our annotation in Appendix C.2.

3.3 Dataset Statistics

Figure 1 provides an example of the annotated dialogue in the simulated operation. Table 1 provides the statistics of our collected dataset. Table 2 pro-

[1]The patient was diagnosed with acute cholangitis and is undergoing laparoscopic cholecystectomy

[2]This is because malignant hyperthermia is a body's adverse reaction to an anesthetic.

General	# Sessions	22
	# Participants / Session	6
Language	# Turns	6.5k
	# Words	49.9k
	# Turns / Session	311
	# Words / Session	2.3k
Others	Duration (min) / Session	19
	# Camera Angles	2
	# Physiological Signals	9

Table 1: Statistics of our collected dataset.

Label	None	Seek	Eval	Impl	Plan	All
Num	3.7k	1.3k	0.8k	0.6k	0.3k	6.9k

Table 2: Label distributions.

vides the label distributions. Appendix C provides more information for the dataset as well as the physiological signals included.

We apply ten-fold cross-validation on our dataset and report the average macro and micro F1 scores in the following setups. For each fold, we use 17, 2, and 3 sessions for training, validation, and testing, respectively.

4 Charactersitic I: Imbalanced Class Distribution

Evaluation Setups. Here we constrain our study within the text domain and apply different methods to handle label imbalance. We directly tune a BERT base (Devlin et al., 2019) model to learn directly from the skewed data. In addition, we prompt the 8B and 70 B versions of the open-source Llama 3 model (abbreviated as Llama in figures) and closed-source GPT-4 and GPT-4o with and without demonstrations. Appendix D provides additional baseline models and their results.

Discussions. Figure 2 compares the F1 scores averaged by class (macro F1 scores) and F1 scores averaged by instances (micro F1 scores). For the tuning-based methods such as $\text{BERT}_{\text{base}}$, though it can achieve the highest micro F1 score, the macro F1 score remains much lower and is comparable to 0-shot or few-shot performances by GPT-4. This suggests that tuning-based methods bias the model to better learn the majority class, while the LLMs with a few demonstrations from each class do not suffer from the performance disparity between the macro and micro F1 scores. There is a significant performance boost for Llama 8B from 0-shot to 1-shot, suggesting even a single example can guide

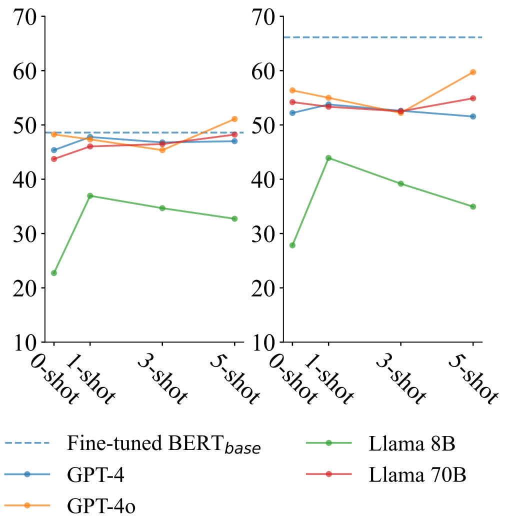

Figure 2: Comparison of macro F1 scores (F1 scores averaged by class, on the left) and micro F1 scores (F1 scores averaged by instances, on the right) versus number of demonstrations (number of shots). We compare both scores for the fine-tuned $\text{BERT}_{\text{base}}$ model, 0-shot and few-shot prompting for LLMs.

smaller LLMs to better reason. In contrast, there is no significant performance boost if we increase the number of demonstrations in the few-shot setting. We hypothesize that since our dataset includes dialogues happening in the real world, there is a diverse forms of dialogue patterns. Therefore, a few demonstrations may be insufficient for the model to assess all the possible situations.

5 Charactersitic II: Conversational Nature.

As shown in Figure 1, people are interacting with each other to actively communicate information in the operation process. Hence, an ideal model would leverage the context information of the interaction to better assess the current situation.

Evaluation Setups. We take the best performed closed-source LLM, GPT-4o, and the best performed open-source LLM, Llama 70B from Figure 2. We then prompt them with one turn both before and after the current round (context size of 3 in Figure 3) or two turns before and after the current turn (context size of 5 in Figure 3). In both situations, we report the performance by providing no demonstration (0-shot) or a single demonstration (1-shot).

Discussions. Figure 3 reports the performance comparison across different settings. For GPT-4o,

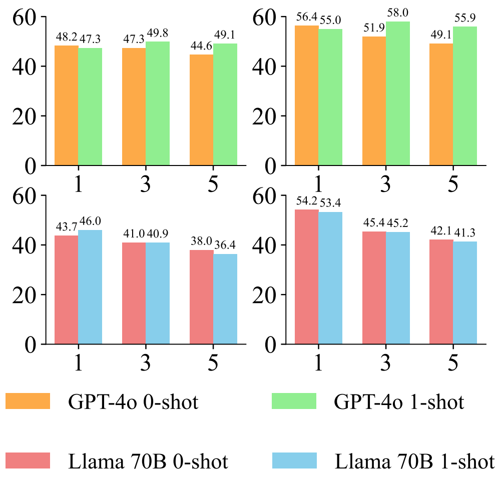

Figure 3: Comparison of macro F1 scores (F1 scores averaged by class, on the left) and micro F1 scores (F1 scores averaged by instances, on the right) versus the context size (x-axis). For instance, "3" on x-axis represents a context of size "3", where we include one turn both before and after the current turn in our prompt to the LLM.

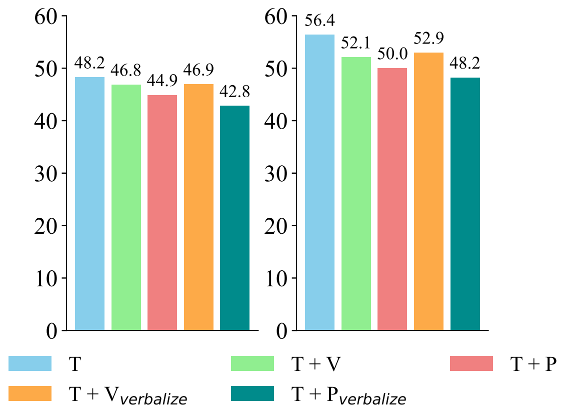

Figure 4: Comparison of macro F1 scores (F1 scores averaged by class, on the left) and micro F1 scores (F1 scores averaged by instances, on the right) when we pass in different modalities. "T" standars for text-only, "V", "P" standard for visual signals and physiology signals, respectively. "T + $V_{\text{verbalize}}$" and "T + $P_{\text{verbalize}}$" stand for verbalizing the content by GPT-4o first, and then pass the text description with the other instructions to the GPT-4o model.

it may leverage the demonstration as well as the context information better, and acquires the best results when we feed it with one demonstration and context size of 3. In contrast, the demonstration and the context hurt Llama 3's performance. One possible reason could be because of the long input prompt. On average, there is around 1,000 tokens per example if we feed the context information and one demonstration, while LLMs with a smaller context window size like Llama 3 may struggle with such long context information, similar to the findings by He et al. (2024).

6 Characteristic III: Multimodality Beyond Text and Vision.

Evaluation Setups. We evaluate the GPT-4o model, a multimodal end-to-end LLM with different modalities as the input, including feeding pure text (T), text and the operation video from two angles (+V), text and the physiology signals (+P). In addition, we try to let GPT-4o first verbalize what happens in the video, then pass the verbalized version of the information to GPT-4o together with the dialogue and instructions.

Discussion. From Figure 4, we can see that GPT-4o may still fail to leverage the visual or the physiological signals effectively. Moreover, when we verbalize the information for the physiological signals, GPT-4o suffers a 2% performance drop for both macro and micro F1 scores. This indicates that though GPT-4o is capable handling a variety of vision tasks (Deng et al., 2024; Zhou et al., 2024b), reasoning over frames that require medical domain knowledge remains challenging. In addition, adding visual modality also hurts GPT-4o's performance on our task. As we sample the frames corresponding to the timestamps when the dialogue happens, some cases may correspond to ten or more frames. For such cases, GPT-4o may struggle to leverage information from the video effectively, which aligns with the finding by Zhou et al. (2024a).

7 Conclusion

In this paper, we introduced ***CliniDial***, a naturally emerged dialogue dataset from the clinical operation with distinguished characteristics from the existing benchmarks. Through experiments, we showed that existing methods do not work well on ***CliniDial***. We hope the described characteristics in ***CliniDial*** could invite future effort to close the gap between NLP methods and the real-world applications.

Limitations

Due to the difficulty of setting up the environment and the data collection process, the dataset is collected mainly on 22 clinical operation sessions.

However, we note that there are 6.5k turns and 49.9k words in total in ***CliniDial***. We demonstrate the distinct characteristics of our data from the existing benchmarks and provide the performance of popular NLP methods. However, due to the scope of this study, we cannot evaluate every possible method and would like to invite future effort on a comprehensive evaluation of NLP methods on clinical data.

Ethics Statement

We note that the study was approved by the Institutional Review Board. Since the data from the two cameras may reveal the identity of the team, we may not release the camera data. We are considering to release an anonymized version of the dialogue transcription to facilitate future research on clinical NLP. We expect researchers to continue building new algorithms and methods on top of this clinical dataset.

References

Joseph Allen, Hayley Hung, Joann Keyton, Gabriel Murray, Catharine Oertel, and Giovanna Varni. 2021. Insights on group and team dynamics. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 855–856.

David P Baker, Sigrid Gustafson, Jeff Beaubien, Eduardo Salas, and Paul Barach. 2005. Medical teamwork and patient safety: the evidence-based relation. *AHRQ publication*, 5(53):1–64.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. 2022. In-the-wild video question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5613–5635, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

K Catchpole, A Mishra, A Handa, and P McCulloch. 2008. Teamwork and error in the operating room: analysis of skills and roles. *Annals of surgery*, 247(4):699–706.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. *Preprint*, arXiv:2402.12424.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Simone Herzberg, Matt Hansen, Amanda Schoonover, Barbara Skarica, James McNulty, Tabria Harrod, Jonathan M Snowden, William Lambert, and Jeanne-Marie Guise. 2019. Association between measured teamwork and medical errors: an observational study of prehospital care in the usa. *BMJ open*, 9(10):e025314.

Hayley Hung, Litian Li, Jord Molhoek, and Jing Zhou. 2024. The discontent with intent estimation in-the-wild: The case for unrealized intentions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Robert Scott Isaak and Marjorie Podraza Stiegler. 2016. Review of crisis resource management (crm) principles in the setting of intraoperative malignant hyperthermia. *Journal of anesthesia*, 30:298–306.

Richard N Keers, Steven D Williams, Jonathan Cooke, and Darren M Ashcroft. 2013. Causes of medication administration errors in hospitals: a systematic review of quantitative and qualitative evidence. *Drug safety*, 36:1045–1067.

Florian Klonek, Fabiola Heike Gerpott, Nale Lehmann-Willenbrock, and Sharon K Parker. 2019. Time to go wild: How to conceptualize and measure process dynamics in real teams with high-resolution. *Organizational Psychology Review*, 9(4):245–275.

William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. 1986. An evaluation of outcome from intensive care in major medical centers. *Annals of internal medicine*, 104(3):410–418.

Michaela Kolbe and Margarete Boos. 2019. Laborious but elaborate: the benefits of really studying team dynamics. *Frontiers in psychology*, 10:433269.

Nale Lehmann-Willenbrock and Hayley Hung. 2023. A multimodal social signal processing approach to team interactions. *Organizational Research Methods*, page 10944281231202741.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Martin A Makary and Michael Daniel. 2016. Medical error—the third leading cause of death in the us. *Bmj*, 353.

Michael A Rosen, Deborah DiazGranados, Aaron S Dietz, Lauren E Benishek, David Thompson, Peter J Pronovost, and Sallie J Weaver. 2018. Teamwork in healthcare: Key discoveries enabling safer, high-quality care. *American Psychologist*, 73(4):433.

Jan B Schmutz, Zhike Lei, and Walter J Eppich. 2021. Reflection on the fly: development of the team reflection behavioral observation (turbo) system for acute care teams. *Academic Medicine*, 96(9):1337–1345.

Jan B Schmutz, Laurenz L Meier, and Tanja Manser. 2019. How effective is teamwork really? the relationship between teamwork and performance in healthcare teams: a systematic review and meta-analysis. *BMJ open*, 9(9):e028280.

Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2009. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE transactions on systems, man, and cybernetics-part A: systems and humans*, 40(1):185–197.

Christina Stevenson, Avneesh Bhangu, James J Jung, Aidan MacDonald, and Brodie Nolan. 2022. The development and measurement properties of the trauma non-technical skills (t-notechs) scale: a scoping review. *The American Journal of Surgery*, 224(4):1115–1125.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Sallie J Weaver, Michael A Rosen, Deborah DiazGranados, Elizabeth H Lazzara, Rebecca Lyons, Eduardo Salas, Stephen A Knych, Margie McKeever, Lee Adler, Mary Barker, et al. 2010. Does teamwork improve performance in the operating room? a multilevel evaluation. *The Joint Commission journal on quality and patient safety*, 36(3):133–142.

Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve WJ Kozlowski, and Hayley Hung. 2018. The i in team: Mining personal social interaction routine with topic models from long-term team data. In *23rd International Conference on Intelligent User Interfaces*, pages 421–426.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024a. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024b. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*.

Labels	Behavior Subcodes
Seek	Actively inviting input Expressing uncertainty
Evaluate	Stating a working hypothesis Recapping Explicitly assessing the situation Reasoning
Plan	Stating plans and priorities
Implement	Stating one's ongoing actions Designating tasks

Table 3: Behavior subcodes corresponding to each of our labels. We follow the definition from Schmutz et al. (2021) to determine the subcodes for "Seek", "Evaluate", and "Plan". We add another label of "Implement" given the characteristics of our data source.

A Label Details

Table 3 provides an overview of the behavior subcodes for each label.

Seek includes:

- the action of actively inviting the team members to provide information and share ideas about the current event.
- expressing uncertainty with an implicit invitation to share information.

Evaluate includes:

- a clear formulation of a working hypothesis or diagnosis about the current situation.
- bringing together various pieces of information and providing a summary.
- providing an explicit judgment, giving value to a certain process, information, or strategy. This can be the process of evaluating information that has been gained through seeking information.
- explaining why certain things are more important, or why a specific behavior needs to be done.

Plan refers to laying out the course of action for the next few minutes that needs to contain at least two actions.

Implementation refers to stating the member is conducting the task or delegates a task to another team member.

B Scenario Details

The role of primary anesthesiologist was played by one of the course participants. The surgeon and secondary anesthesiologist (assistant) were played by other course participants. The role of surgeon served as a confederate along with the course instructors. The scenario begins with the primary anesthesiologist taking over the case from one of the course instructors. The patient is receiving general anesthesia and the procedure has already begun. The procedure is complicated by surgical difficulties resulting in the surgeon requesting additional muscle relaxants and increased insufflation pressures. There is also concern that the patient is developing sepsis given the significant gallbladder infection. The patient develops malignant hyperthermia (MH) as the simulated scenario progresses. The primary anesthesiologist must recognize this and begin appropriate treatment. Treatment algorithms for MH are well-known and broadly available (Hopkins et al., 2020; Rosenberg et al., 2020). Definitive treatment includes stopping the triggering agents, administering dantrolene, and supportive care.

C Dataset Information

The total number of anesthesiologists studied was 22; 15(68%) males and 7(32%) females. As part of the Maintenance of Certification in Anesthesiology (MOCA©), anesthesiologists who were board certified after 2000 were required to participate in a simulation course at a simulation center. The participants were board certified anesthesiologists who attended a simulation course at a midwestern academic medical center over a 5 year period. Date of initial certification was obtained from the American Board of Anesthesiologists (ABA) Physician Directory. The study was approved by the Institutional Review Board.

C.1 Physiological Signals

The physiological signals in our dataset include:

SpO2 refers to Peripheral Oxygen Saturation which measures the oxygen saturation level in the blood. Such signal is typically measured through a pulse oximeter.

ECG II refers to Electrocardiogram Lead II which represents the electrical activity of the heart as measured by electrodes placed on the body.

APB refers to Arterial Blood Pressure which represents the pressure exerted by blood on the walls of the arteries during the cardiac cycle.

HR refers to Heart Rate which indicates the number of heartbeats per minute.

NIBP refers to Non-Invasive Blood Pressure which measures blood pressure without the need to insert instruments into the body.

Temperature represents the body's temperature and is often measured using a thermometer.

Respiratory Waveform represents the pattern of inhalation and exhalation.

CO_2 means Carbon Dioxide which typically refers to end-tidal CO_2, which represents the concentration of carbon dioxide at the end of an exhaled breath.

IBP refers to Invasive Blood Pressure which measures blood pressure using invasive techniques, typically involving a catheter inserted into an artery or vein.

C.2 Annotation Details

Two researchers coded six out of 22 randomly selected data files. The researchers discussed findings and resolved discrepancies through the process of social moderation. They achieved a Cohen's kappa score of 0.73. The two researchers then independently annotated the remaining dataset.

D More Details about the Methods

In addition to the methods in Section 4, we have a majority vote baseline model which always predicts the major class. As expected, it reaches a decent micro F1 score (55.63) due to the class imbalance, while a much lower macro F1 score (14.01). In addition, we test two non-deep learning methods such as RUSBoost (Seiffert et al., 2009) and SMOTE (Chawla et al., 2002) algorithm which is specifically designed to address class imbalance. However, these pre-deep learning methods attains 24.21 and 32.32 macro F1 scores, much worse than simply tuning $\text{BERT}_{\text{base}}$ model or prompting LLMs.

E What Do Medical Professionals Expect from NLP?

We are also interested to see how the medical professionals would view the results we get by employing these current NLP methods. Therefore, we invite feedbacks from a medical professional who has been working in the domain for over a decade. Here are what we get:

1. They see a great opportunity to apply these LLMs on behavioral evaluation in the medical domain. They point out that the current evaluation practices in medical domains have significant limitations (Kolbe and Boos, 2019; Klonek et al., 2019; Stevenson et al., 2022), which typically are labor-intensive and prone to personal biases and errors. They expect NLPers to develop consistent, reliable evaluation protocol to give feedback to the healthcare professionals.

2. They expect a protocol that can take multimodal input into consideration including the team dialogue, patient vitals, and procedure videos. We note that this is one of the characteristics for ***CliniDial***. They also hope the NLP system could pinpoint specific teamwork deficiencies in the process.

3. They also point out the related NLP methods that they find useful in their domain. For instance, intent classification, dialogue summarization, and multimodal reasoning works from NLP can provide quantifiable insights into teamwork dynamics and communication patterns in multimodal clinical data (Zhang et al., 2018; Allen et al., 2021; Lehmann-Willenbrock and Hung, 2023; Hung et al., 2024). We note that ***CliniDial*** contain rich conversational data.