CliniDial: A Naturally Occurring Multimodal Dialogue Dataset for Team Reflection in Action During Clinical Operation

Naihao Deng **Kapotaksha Das** **Rada Mihalcea**
Vitaliy Popov* **Mohamed Abouelenien***
University of Michigan, Ann Arbor University of Michigan, Dearborn
{dnaihao, zmohamed}@umich.edu

Abstract

In clinical operations, teamwork can be the crucial factor that determines the final outcome. Prior studies have shown that sufficient collaboration is the key factor that determines the outcome of an operation. 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 annotate behavior codes following an existing framework to understand the teamwork process for ***CliniDial***. We pinpoint three main characteristics of our dataset, including its label imbalances, rich and natural interactions, and multiple modalities, and conduct experiments to test existing LLMs' capabilities on handling data with these characteristics. Experimental results show that ***CliniDial*** poses significant challenges to the existing models, inviting future effort on developing methods that can deal with real-world clinical data. We open-source the codebase at https://github.com/MichiganNLP/CliniDial.†

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 provide initial analysis of our dataset, and lay out potential directions in using our dataset. We hope researchers can leverage our dataset creatively, and propose methods to handle real-world clinical data.

In this paper, we pinpoint three main characteristics of ***CliniDial***, including its label imbalances, rich and natural interactions, and multiple modalities. Corresponding to each feature, we design sets of experiments to investigate existing methods' ability to deal with such data, including the Large Language Models (LLMs) from GPT families and the open-source Llama families. Experimental results show that ***CliniDial*** poses significant challenges to these 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 F).

In summary, our contributions are two folds:

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

*Both senior authors contributed equally to this work.
†Due to ethical considerations, the text data and video representations will be provided upon reasonable requests.

Findings of the Association for Computational Linguistics: ACL 2025, pages 21781–21798
July 27 - August 1, 2025

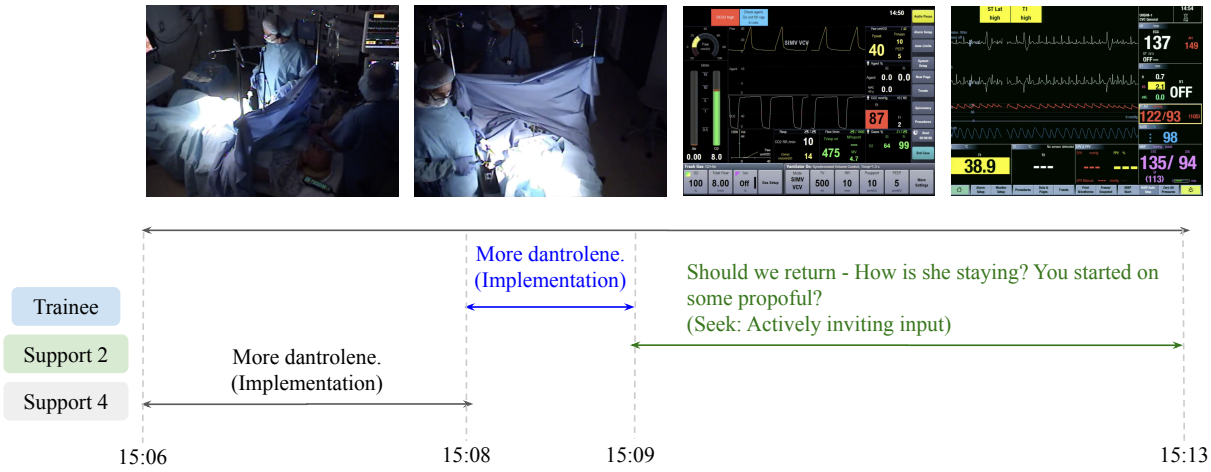


Figure 1: An example of the labeled dialogue in the simulated operation. Two cameras capture the scenes from two angles and two real-time monitoring systems provide the patient's physiological signals. We only include the trainee and the two supports in this example, as they are the only three people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

clinical operation.

2. We evaluate our dataset against various existing methods with different setups and provide an 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 significantly 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.

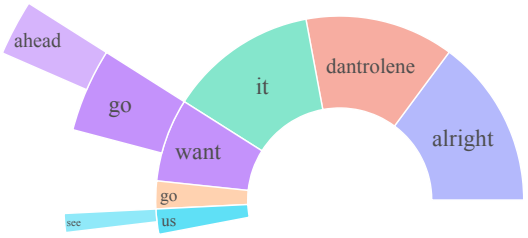
3 *CliniDial* Dataset

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 *. 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-

*The patient was diagnosed with acute cholangitis and is undergoing laparoscopic cholecystectomy



(a) Distribution of words uttered.

(b) Word cloud for frequent words.

Figure 2: Distribution of words uttered (a) and word clouds for frequent words (b) by the support role. Figures 8 and 9 present the plots for all three roles.

maker [†]. 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

Table 5 provides the definitions of each label and the corresponding examples. Following Schmutz et al. (2021), we include three labels of "Seek", "Evaluate" and "Plan". 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 additional details for each label. We describe the details of our annotation in Appendix C.2.

3.3 Dataset Statistics and Analysis

Table 1 provides the overall statistics of our collected dataset. In total, there are 2,279 utterances in our dataset uttered by the support role, 1,808 by the surgeon role, and 2,576 by the trainee role.

Dataset example. Figure 1 provides an example of the annotated dialogue in the simulated operation. As aforementioned, we have transcripts of different roles in the operation, together with camera views from two different angles, and the physiological signals of the patient mannequin. We provide additional examples in Figure 7 and dialogue snippets in Table 7 in Appendix D.

[†]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 distributions. Table 2 provides the total and role-specific label counts. Figure 3 provides the overall as well as role-specific label distributions. Such role-specific label distributions reveal the internal collaboration and role-specific contributions during the operation. We observe that, overall, the majority of labels are "seek" and "evaluate," rather than "implement" or "plan." This highlights the critical importance of communication and actively assessing the current situation during real-world operations. Breaking it down by role, the surgeon role is most associated with the "seek" task, which accounts for 30.4% of their labels. This indicates that, as the central figure in the operation, surgeons rely heavily on support and collaboration from other roles to fulfill their responsibilities effectively. Additionally, the support role has the highest proportion of "implement" labels (13.7%), which aligns with their primary function of providing assistance and executing essential procedures during the operation process.

Word analysis. Figure 2 provides the token distributions and word cloud for the support role. We provide these plots for all three roles in Figures 8 and 9 in Appendix D.

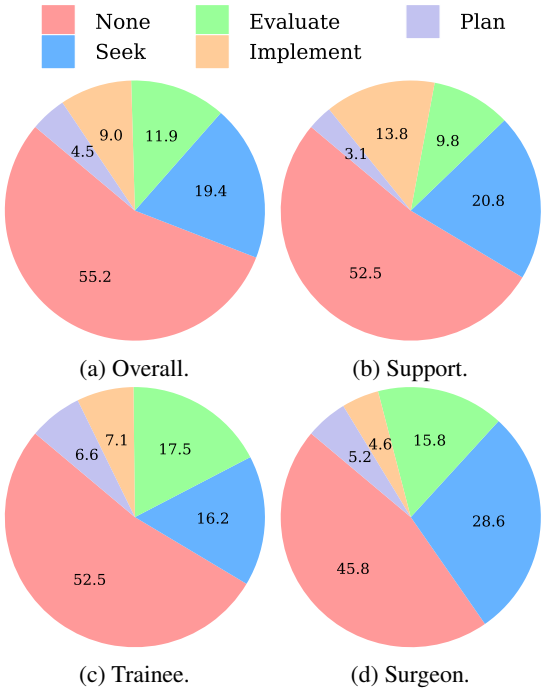
In terms of the word distributions, we observe that trainees and surgeons often use "thank you", while supports often use "alright" (Figure 8). This demonstrates the interactions happening during clinical operations, where there are such clues to acknowledge the actions conducted by others. In addition, such phrase usages reflect the role difference. Surgeons and trainees are the ones who need help from the support in the operation process, therefore they use "thank you" more often,

while the support's primary job is to support others, therefore there is more of "alright". Apart from these acknowledgment interactions, trainees often use the phrase "CO2 up" and surgeon often uses the phrase "gallbladder out", which relate to how they describe the situation or invite others' help during the operation process. In terms of the frequent words, we observe many terminologies used by these roles. For instance, both support and trainee roles frequently use the term "dantrolene" (Figure 9), a medication primarily used to relax muscles, particularly in emergency settings (Krause et al., 2004). Additionally, other medical terms appear frequently, such as "abdomen" and "gallbladder", which refer to anatomical structures; "septic", which relates to a patient's condition; and "antibiotic", which pertains to medication.

Camera view analysis. In addition, we provide a few qualitative observations based on videos captured from the two camera angles. We observe the *local focus on surgeon actions.* Most operational actions occur within a localized region. For example, as seen in the screenshots in Figure 1, the surgeon's body remains mostly stationary while manipulating tools during the operation. In addition, we observe *role-specific movement patterns.* Supports and trainees tend to move around the room, creating distinct communication scenarios. For instance, the support comes to the trainee to explain the background information in Figure 7a. In some scenes, the surgeon halts their operation to look around and communicate with colleagues. For instance, after talking to the trainee, the support moves to the doctor in the later scenes in Figure 7b. In other cases, the surgeon continues working while colleagues stop to seek information from them. These contrasting behaviors can be interesting clues to understanding role-specific dynamics, and we plan to include additional examples of such interactions in the final version. Our dataset poses *unique visual challenges in operation settings.* The people in the videos are different from those typically seen in everyday video datasets. As shown in Figure 1, participants wear uniforms, hats, and facial masks, which obscure facial expressions. For example, in Figure 1, a human observer might easily infer that the surgeon is frowning while staring at a monitor, even though most facial features are obscured. However, such subtleties can pose challenges to the VLMs. We would be happy to learn if you have suggestions for methods from the vision community that could help us further analyze the visual information in our dataset.

Label	None	Seek	Eval	Impl	Plan	All
Overall	3.7	1.3	0.8	0.6	0.3	6.9
Support	1.0	0.4	0.2	0.3	0.1	1.9
Trainee	1.2	0.4	0.4	0.2	0.1	2.2
Surgeon	0.7	0.5	0.3	0.1	0.1	1.6

Table 2: Total and role-specific label counts (in k).



(a) Overall. (b) Support. (c) Trainee. (d) Surgeon.

Figure 3: Overall label distribution and role-specific label distributions. The numbers on the pie charts represent percentage values.

Appendix C provides additional 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.

3.4 Potential Usage of *CliniDial*

In this paper, we present three case studies scrutinizing existing LLMs' capabilities on handling domain-specific data with specific characteristics in Sections 4 to 6. We include two potential usages of ***CliniDial*** and encourage future research in using our dataset in creative ways.

Testing the Effectiveness of Existing Methods. We present 6.9k examples of annotated examples

in ***CliniDial***. In this paper, we take the first step to test various methods including LLMs' capabilities in handling clinical data, especially on data with imbalanced class distribution, conversational nature, and multiple modalities. We highlight that our annotated dataset can be a valuable source to investigate the innate ability of LLMs in handling real-world clinical data.

Understanding the Interaction Mechanisms in Clinical Setups. As demonstrated in Section 3.3, our dataset presents a valuable source of interactions across different roles that happen in clinical operations. There is rich domain jargon involved such as "gallbladder", "dantrolene", and phrases that are specific to the clinical operation setup, such as "CO2 up", etc. Moreover, there are interesting phrase usage patterns during such interaction processes. For instance, surgeons and trainees use "thank you" often, while supports use "alright" often. Such language use patterns reflect what tasks each role carries out during the clinical operations, and how one role reacts to the requests or actions of the others. Therefore, the dialogue interaction in ***CliniDial*** can facilitate future research on understanding the interaction mechanisms in the clinical operation.

Incorporated in LLM's Training Loop. When we collect the dataset, we have included the timestamps for different modalities as shown in Figure 1. Such timestamps can map information across different modalities. For instance, given certain frames from the videos, we can pinpoint the sentence uttered by the surgeon and the supports correspondingly. This mapping can enable LLM training objectives such as masking information in one modality and then asking LLMs to predict the missing signals based on the information from all the remaining modalities. We highlight that despite the difficulty of the data collection process, ***CliniDial*** includes 6.5k turns and 49.9k words in total. Though such amount of data may not be sufficient for pre-training from the scratch, researchers may adopt our dataset for continual pre-training to facilitate models in clinical domains. In addition, we present 6.9k examples annotated with labels, which can serve as a valuable source in models' supervised fine-tuning stage.

In the following sections, we present three case studies revealing the challenges for existing methods including LLMs in dealing with our dataset.

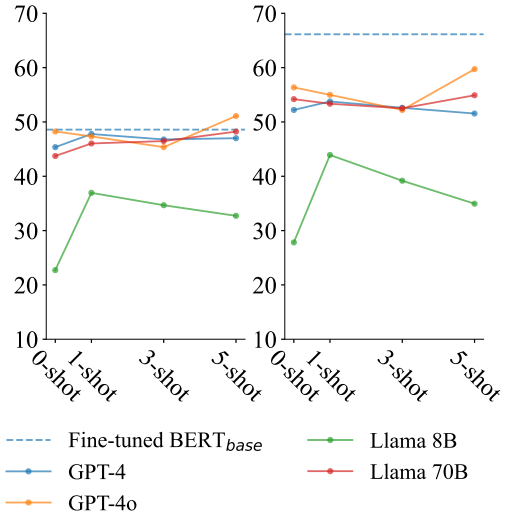


Figure 4: Comparison of macro F1 scores (F1 scores averaged by class, on the left) and micro F1 scores (F1 scores averaged by instances, on the right) versus number of demonstrations (number of shots). We compare both scores for the fine-tuned $\text{BERT}_{\text{base}}$ model, 0-shot and few-shot prompting for LLMs.

4 Charactersitic I: Imbalanced Class Distribution

Here we constrain our study within the text domain for handling label imbalance.

4.1 Evaluation Setups

We directly fine-tune a BERT base (Devlin et al., 2019) model to learn directly from the skewed data. In addition, we prompt Llama 3 8B and 70B models (abbreviated as Llama in figures) and GPT-4 and GPT-4o with and without demonstrations. Specifically, we use the following prompt to provide the list of all possible labels:

Prompt

In the classification task, there are 5 labels:
[Seek, Evaluate, Plan, Implement, None].
Here are the details for each label:
[Description of each label]
Fill in the blanks:
Output in the format of {
"Sentence": <SENTENCE>,
"Label": ,
}

In the few-shot settings, we provide corresponding examples along with the label definitions to the models. Table 5 in Appendix A provides labels, their corresponding definitions, and examples.

We adapt the definition from the corresponding information provided in the annotator guidelines. "Examples" provides the demonstrations of examples sourced from the dataset. We use this prompt for the prompting experiments in Sections 4 to 6. Appendix E provides additional baseline models and their results. We extract the label from the JSON output of the model to calculate the macro F1 scores (F1 scores averaged by class) and the micro F1 scores (F1 scores averaged by instances).

4.2 Discussions

Tuning-based method. Figure 4 compares the F1 scores averaged by class (macro F1 scores) and F1 scores averaged by instances (micro F1 scores). Though the fine-tuned $\text{BERT}_{\text{base}}$ model can achieve the highest micro F1 score of 66.6%, it yields the macro F1 score of 48.6%, which is much lower compared to its micro F1 score, and is comparable to GPT-4o's macro F1 score at 0-shot (48.3%) or 5-shot (51.1%). 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.

Prompting-based method. There is a significant performance boost for Llama 8B from 0-shot, achieving a macro F1 score of 22.7% to 1-shot, achieving a macro F1 score of 37.0%, suggesting even a single example can guide smaller LLMs to better reason. However, when we increase the number of demonstrations, the Llama 8B model experiences a performance decline, from a macro F1 score of 37.0% at 1-shot to 34.7% at 3-shot and 32.7% at 5-shot. In contrast, there is slight performance improvement for the Llama 70B model when we increase the number of demonstrations, from a macro F1 score of 43.7% at 0-shot, to 46.0% at 1-shot, 46.5% at 3-shot and 48.2% at 5-shot. We attribute such a phenomenon to the limited innate capabilities of Llama 8B model, as the smaller scale model may not capture the underlying knowledge from a few demonstrations, instead it may be distracted by the longer input when we increase the number of demonstrations. Moreover, there is only slight performance improvement for Llama 70B and GPT models when we increase the number of demonstrations. For the Llama 70B model, its macro F1 score improves from 46.0% at 1-shot to 48.2% at 5-shot. For the GPT-4 model, its macro F1 score remains around 47% when we increase the shot number from 1 to 5, while for GPT-4o model, its macro F1 score improves from 47.3% at 1-shot to 51.1% at 5-shot. We hypothesize that the real-world nature of our dataset leads to diverse dialogue patterns, making a few demonstrations insufficient for the model to cover all scenarios.

5 Charactersitic II: Conversational Nature

As discussed in Section 3.3, ***CliniDial*** involves rich interactions among people where they actively communicate information in the operation process. Hence, an ideal model would leverage the context information of the interaction.

5.1 Evaluation Setups

We take the best performed closed-source LLM, GPT-4o, and the best performed open-source LLM, Llama 70B from Figure 4. We then prompt them with one turn both before and after the current round (context size of 3 in Figure 5) or two turns before and after the current turn (context size of 5 in Figure 5). Specifically, we insert the following additional prompt into the prompt we use in Section 4 after the label descriptions:

Additional Prompt

For the dialogue:
<CONTEXT BEFORE>
<ROLE>: <SENTENCE>
<CONTEXT AFTER>

In the prompt, <CONTEXT BEFORE> and <CONTEXT AFTER> correspond to the turns before the current utterance and the turns after, and the model needs to assign a label for <SENTENCE>. In both situations, we report the performance by providing no demonstration (0-shot) or a single demonstration (1-shot).

5.2 Discussions

Figure 5 reports the performance comparison across different settings. For GPT-4o, we observe a performance boost when we include the interactions. For instance, under the 1-shot setup, the GPT-4o model's macro F1 score improves from 47.3% to 49.8% and micro F1 scores improve from 55.0% to 58.0% when we increase the context size from 1 to 3. However, when we further increase the context size to 5, it suffers a performance decline compared to the context size of 3, but still outperforms the case when the context size is 1. This

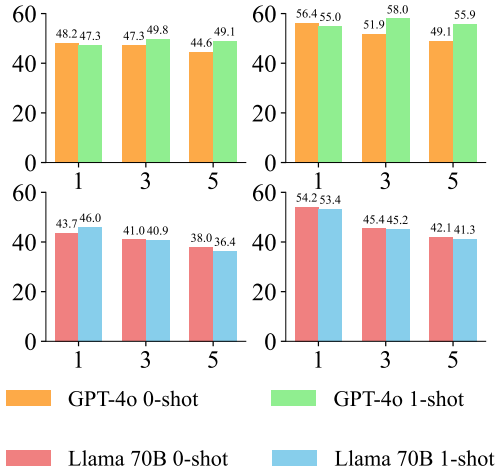


Figure 5: Comparison of macro F1 scores (F1 scores averaged by class, on the left) and micro F1 scores (F1 scores averaged by instances, on the right) versus the context size (x-axis). For instance, "3" on x-axis represents a context of size "3", where we include one turn both before and after the current turn in our prompt to the LLM.

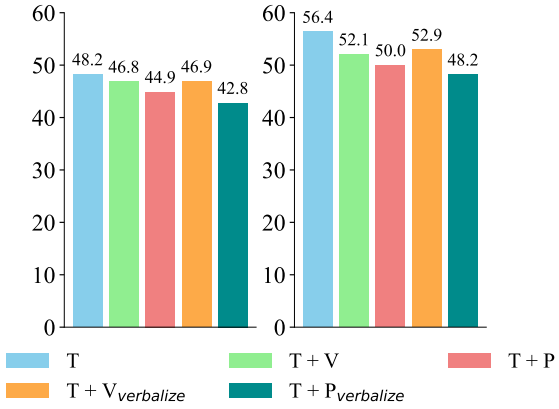


Figure 6: Comparison of macro F1 scores (F1 scores averaged by class, on the left) and micro F1 scores (F1 scores averaged by instances, on the right) when we pass in different modalities. "T" standars for text-only, "V", "P" standard for visual signals and physiology signals, respectively. "T + $V_{verbalize}$" and "T + $P_{verbalize}$" stand for verbalizing the content by GPT-4o first, and then pass the text description with the other instructions to the GPT-4o model.

indicates that context can help models better reason the target sentence, but when too much context is provided, the information may be diluted and is less helpful. In contrast, providing demonstrations and increasing context size negatively impact Llama 3's performance. Under the 1-shot setup, Llama 3's macro F1 score drops significantly, from 46.0% to 40.9% when the context size increases from 1 to 3, and further to 36.4 when the context size increases from 3 to 5. We attribute this performance decline to the increased input length. On average, including context information and one demonstration results in an input length of approximately 1,000 tokens per example, utilizing one-eighth of Llama 3's 8k context window. We hypothesize that Llama 3, with its smaller context window, struggles to process such long inputs effectively, consistent with findings by He et al. (2024). In contrast, GPT-4o, equipped with a much larger context window of 128K, is better suited to handle input lengths of this magnitude.

6 Charactersitic III: Multimodality Beyond Text and Vision

6.1 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 camera views or the physiological signals by prompting:

Prompt for Verbalization

The video frames show the [operation scene for/ physiological signal changes from] this patient. Please describe what happens in a few sentences.

After we acquire the verbalized descriptions, we feed the following description after the label definitions we use in the prompt in Section 4.

Description

Here is the description for what happens on the scene: <DESCRIPTION>

6.2 Discussion

From Figure 6, we can see that GPT-4o fails to leverage the visual or the physiological signals effectively. When we add the visual scene input directly, the model's macro F1 score decreases from 48.2% (text-only) to 46.8%, and when we add the screenshot of the physiological signals directly, the model's macro F1 score decreases to 44.9%. This demonstrates that domain-specific data from modalities other than text pose significant challenge to advanced LLMs like GPT-4o. Specifically, for the op-

$V_{verbalized}$	The video frames show a surgical operation taking place in an operating room. In both frames, the surgical team, including at least two members dressed in blue sterile gowns and gloves, is actively engaged in a procedure. The patient is lying on the operating table, partially covered by surgical drapes. A bright surgical light illuminates the operative field, and an anesthesia machine and monitoring equipment are visible nearby. The scene appears to be well-organized, with a focus on maintaining a sterile environment.
$P_{verbalized}$	The patient is on mechanical ventilation (SIMV VCV mode) with normal ventilator settings. However, there is a warning for high $EtCO_2$ (87 mmHg), indicating hypoventilation or CO_2 retention. Oxygen saturation is 99%, and heart rate is 64 bpm—stable but concerning for rising CO_2 levels.

Table 3: An example of $V_{verbalized}$ and $P_{verbalized}$ corresponding to the scenes in Figure 1. The model incorrectly assigns the value of 64 to heart rate, which is the Et value instead.

eration scenes and physiological signals in Figure 1, while a trained medical professional may understand the situation and read the monitor correctly, we hypothesize that the dark scene of the camera and the strong light focusing on the patient's abdomen area may pose challenges to LLMs in their reasoning process. Moreover, since there would not be too many scenes online corresponding to physiological signals, GPT-4o may have not encountered such data in its pre-training process, leading to its limited capabilities to process it. When the physiological signals are verbalized instead of presented as images, GPT-4o experiences an additional 2% performance drop. This highlights the challenges of incorporating physiological data into model's reasoning process, as the errors accumulate during the two-step process of first verbalizing the data and then reasoning over the verbalized context. In contrast, when we verbalize the visual scenes, GPT-4o performs comparably (46.9% versus 46.8% for macro F1 score) to when we pass the visual scenes directly. This indicates that GPT-4o can better handle text-based representations of visual information than it can with text-based representations of physiological signals. We attribute this to GPT-4o's lack of domain-specific knowledge, unlike visual scenes, interpreting physiological signals requires more specialized expertise.

Case Study. We present an example of $V_{verbalized}$ and $P_{verbalized}$ corresponding to the scenes in Figure 1 in Table 3. For $V_{verbalized}$, the model describes the scenes accurately. However, the model hallucinates in the $P_{verbalized}$. In Table 3, 64 is not the heart rate, instead, it is the Et value for the gases. We believe that this is an important direction for future research to make models correctly reason the situations, especially in a high-stakes domain such as clinical operations.

7 Related Work

Multimodal Datasets. Recent years have witnessed significant advancement of multimodal large language models (MLLMs) (Achiam et al., 2023; Liu et al., 2024) that typically involves vision and text capabilities. These MLLMs have demonstrated impressive performance on various visual benchmarks such as visual recognition (Zhang et al., 2024), video understanding (Xu et al., 2021), 3D understanding (Hong et al., 2023) and beyond. Researchers have proposed various vision and text benchmarks to investigate the capabilities of these MLLMs, including captioning tasks such as MS-COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) for image captioning, and MSVD (Chen and Dolan, 2011) for video captioning, question answering tasks such as VQA (Antol et al., 2015) for image question answering and MSVD-QA (Xu et al., 2017) for video question answering. Recently, there is a shift of interest in proposing more nuanced and culturally diverse benchmarks. For instance, WildQA proposes video QA dataset on scenes in the wild (Castro et al., 2022), Ego4D proposes various visual tasks from the egocentric viewpoint (Grauman et al., 2022), CVQA investigates into the culturally diverse multilingual visual question answering (Romero et al., 2024). In addition, researchers have proposed datasets involving other modalities, such the Touch and Go dataset on tactile (Yang et al., 2022), MMAU on audio understanding (Sakshi et al., 2024). To the best of our knowledge, we are the first to propose a dataset that includes the physiological signals. Moreover, we provide timestamps for the utterances, which allows researchers to align the text data with video frames and the physiological signals.

Datasets in Clinical Domains. There has been interdisciplinary research between NLP and clinical or medical domains (Spasic et al., 2020). For in-

stance, researchers have leveraged natural language generation methods to generate medical reports or summaries (Song et al., 2020; Papadopoulos Korfiatis et al., 2022; Ben Abacha et al., 2023), understanding the medical consultant process (Chen et al., 2023). However, most of these existing datasets focus on the consultant process in the clinical setup. We highlight that ***CliniDial*** focuses on the conversation during clinical operations, which possess significant domain-specific features as discussed in Section 3.3. We hope ***CliniDial*** can facilitate future NLP research into understanding the complex clinical operation scenarios.

8 Conclusion

In this paper, we introduced ***CliniDial***, a naturally emerged multimodal dialogue dataset collected from clinical operations. Unlike existing benchmarks, ***CliniDial*** addresses real-world complexities such as imbalanced label distributions, rich team interactions, and multiple data modalities. Through studies on three key characteristics of our dataset, we found that the best-performing model achieves a macro F1 score of only 51.09, indicating significant room for improvement. This performance suggests that existing methods struggle on CliniDial, particularly in handling imbalanced class distributions, leveraging conversational context, and integrating domain-specific multimodal signals. These limitations highlight the gap between existing NLP methods and the demands of real-world clinical applications. We hope ***CliniDial*** can bridge the gap between advancements in our community and real-world clinical applications. We encourage future research efforts to develop domain-adaptive NLP techniques, improve multimodal fusion strategies, and eventually address the challenges of real-world applications.

Limitations and Future Directions

Simulation Setup. The clinical operation described in the study is simulated, so it is likely that the dialogue between the anesthesiologists and support staff lacks the sense of urgency present in a real medical setting. In real-life clinical environments, time pressure, high-stress situations, and the need for quick decision-making usually shape the communication dynamics. This distinction is important to consider when analyzing the dialogue, as the lack of urgency might influence both the content and the tone of communication.

However, to the best of our knowledge, we are the first to study such a medical operation scenario, even in a simulated operation process. In fact, it would be nearly impossible to collect the real emergency operation recordings due to ethical and legal considerations. These simulations are the typical training that medical professionals rely on, and to the best of our knowledge, the best possible way to collect such data.

Scope of the Data. We want to emphasize the difficulty of setting up the real-world clinical operation environment, recruiting people to participate, collecting the data. Although the dataset is collected mainly on 22 clinical operation sessions, we note that there are 6.5k turns and 49.9k words in total in ***CliniDial***.

Scope of the Analysis. We provide various analyses on our dataset in Section 3 and highlight how our dataset is different from the existing benchmarks in Section 2. In addition, we discuss the potential future directions that researchers may explore in Section 3. Furthermore, we study the three characteristics of our dataset and provide the performance of popular NLP methods with respect to each of them. However, due to the scope of this study, we cannot evaluate every possible method and would like to invite future efforts on a comprehensive evaluation of NLP methods on clinical data. For instance, our dataset can be leveraged to answer questions such as "are there sequences of labels that occur frequently in the corpus?" We encourage future efforts on a more in-depth exploration that might reveal underlying structures or recurring communication patterns in the dialogues between anesthesiologists and support staff, and provide a richer understanding of the linguistic dynamics.

Scope of the Experiments. We encourage future efforts to investigate the low F1 scores for the existing LLMs. For instance, prompting methods such as chain of thought (CoT) prompting could be tested to check whether they could enhance the LLM's performance and lead to a higher F1 score, which can lead to a more reliable approach for analyzing clinical dialogues. In this paper, we did not include analyses of the audio setting. Audio characteristics can provide additional insights into the emotional state, stress, urgency, or intent behind the spoken words, offering a better understanding of what's really going on. We leave the exploration

of models fine-tuned with medical expertise to future study. To the best of our knowledge, there are no specific LLMs targeting clinical operation setup.

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.

Acknowledgments

This study is based upon work partially supported by the National Science Foundation under Grant IIS-2202451 and grant IIS-2306372. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433.

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.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pages 190–200.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. Bioinformatics, 39(1):btac817.

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.

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.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18995–19012.

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.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems, 36:20482–20494.

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.

Th Krause, MU Gerbershagen, M Fiege, R Weisshorn, and F Wappler. 2004. Dantrolene–a review of its pharmacology, therapeutic use and new developments. Anaesthesia, 59(4):364–373.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. Advances in neural information processing systems, 36.

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

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A dataset of primary care mock consultations. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. arXiv preprint arXiv:2406.05967.

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.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. arXiv preprint arXiv:2410.19168.

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.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In Proceedings of the 28th International Conference on Computational Linguistics, pages 717–729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. JMIR medical informatics, 8(3):e17984.

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.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In ACM Multimedia.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021. VLM: Task-agnostic video-language model pre-training for video understanding. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4227–4239, Online. Association for Computational Linguistics.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. 2022. Touch and go: Learning from human-collected vision and touch. arXiv preprint arXiv:2211.12498.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

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.

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 4: 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 4 provides an overview of the behavior subcodes for each label. Table 5 provides the comprehensive definition and examples corresponding to 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.

Label	Seek
Definition	All statements that request information from the team about the current event and invite team members to provide information and share ideas; Inquiring for further information. Or expressions of uncertainty with an implicit invitation to share information. Tone of voice; Content of what is being said (questioning the information; Unsure). In response to clarify something.
Examples	Is there anything we are missing? Is there anything else we should be doing? What is the plan afterwards?
Label	Plan
Definition	Laying out the course of action for the next few minutes. Needs to contain at least 2 actions to show a sequence of actions.
Examples	Once MH is recognized: Going to stop the agent and go up on flows.
Label	Evaluate
Definition	Clear formulation of a working hypothesis or diagnosis about the current situation, or various pieces of information are brought together and a summary is provided; Recapping lab results (does not have to be new information); Providing an explicit judgment for something, give value to a certain process, information or strategy ...
Examples	That's a nasty gallbladder.
Label	Implement
Definition	Stating one's ongoing actions or designating tasks.
Example	Yes. Yes. Pushing. Pushing. Pushed.
Label	None
Definition	None of the other labels apply here.
Examples	Okay.

Table 5: Examples of the labels, their definitions, and corresponding utterances in our dataset. We omit part of the definitions for the label "Evaluate". Appendix A provides additional details of each label.

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 on the six files. The two researchers then independently annotated the remaining dataset.

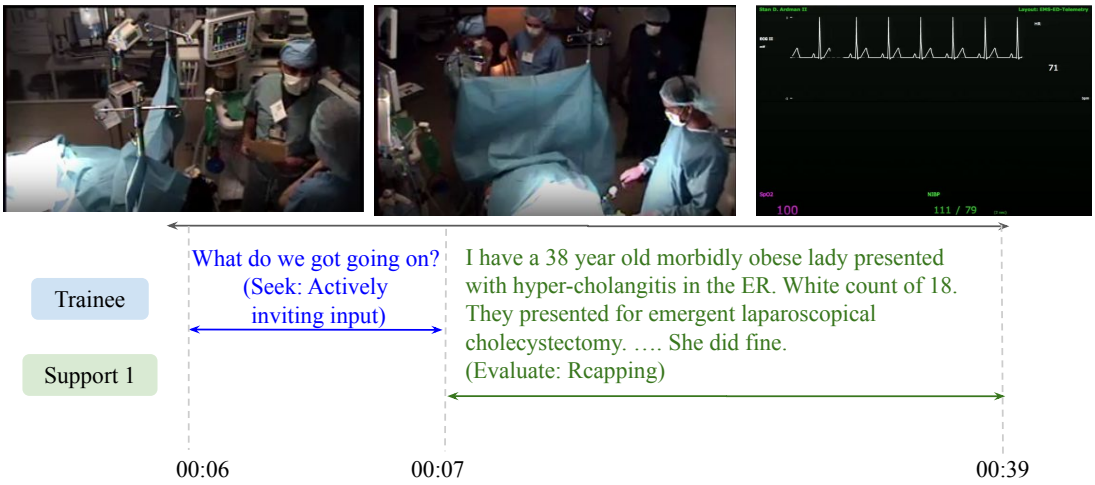
D Dataset Analysis

Figure 7 provides two additional examples of the dialogues in our dataset. Table 7 provides dialogue snippets in our dataset. Figure 8 provides the word distributions for the three roles, respectively. Figure 9 provides the most frequent words uttered by the three roles, respectively.

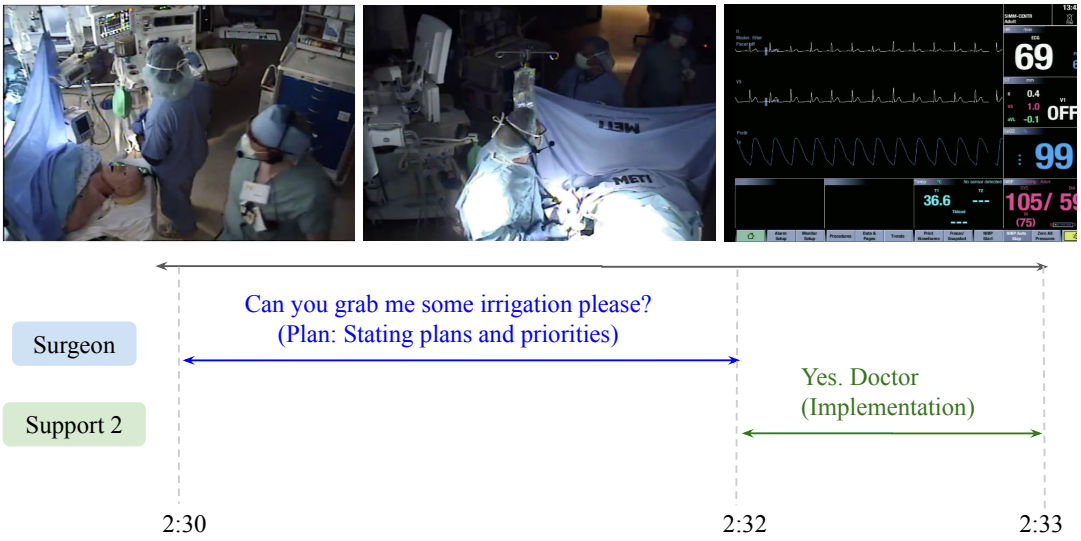
Specific word use. Table 6 provides the word counts for "dantrolene", we find that its frequency varies significantly. In one session, it appears 31 times, while in another, it occurs only once throughout the entire operation. This variability suggests that certain medical terms naturally appear more frequently in specific cases rather than uniformly across all procedures. Therefore, the presence of "dantrolene" does not indicate a dataset bias toward specific procedures but rather reflects real-world variations in clinical practice.

E 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



(a) An example happens at the beginning of the operation, when the trainee seeks background information and the support provides such information to the trainee.



(b) An example happens when the surgeon is conducting the operation, when the surgeon pauses and asks for support, and the support helps the surgeon to grab the irrigation.

Figure 7: Additional examples of the labeled dialogue in the simulated operation.

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.

F 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:

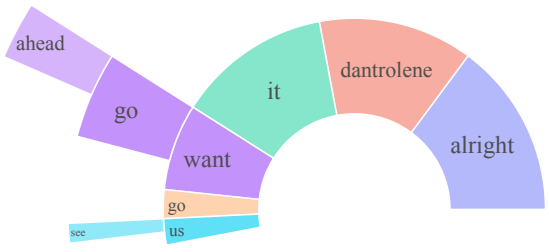
1. They see a great opportunity to apply these LLMs on behavioral evaluation in the medical domain. They point out that the current evaluation practices in medical domains have significant limitations (Kolbe and Boos, 2019; Klonek et al., 2019; Stevenson et al., 2022), which typically are labor-intensive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

Session ID	Counts
1	8
2	12
3	10
4	10
5	8
6	31
7	9
8	18
9	14
10	16
11	3
12	4
13	13
14	**31**
15	20
16	17
17	13
18	3
19	6
20	19
21	**1**
22	12

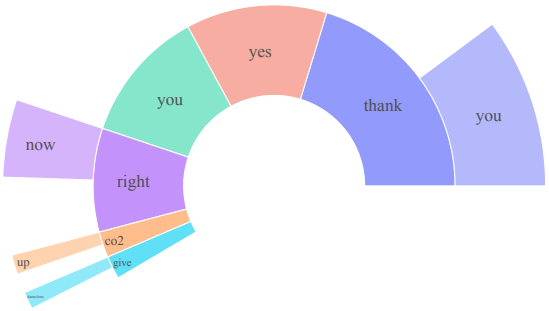
Table 6: Word counts for "dantrolene" in the 22 sessions. We note that all the sessions last around the same. For instance, session 21 ("dantrolene" appears once) lasts for 20:03 minutes, while session 14 ("dantrolene" appears 31 times) lasts for 19:06 minutes.

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.



(a) Support.



(b) Trainee.

(c) Surgeon.

Figure 8: Distributions of words uttered from different roles in ***CliniDial***.

Dialogue Snippet 1	
...	
Support 3	So currently their recommendations are to not change the machine. Do you want to put the patient back on the vent and free up your hands?
Trainee	Sure. Okay. If that's the recommendation. Absolutely. Great so next thing insulin, glucose, calcium.
Support 2	The ICU is calling.
Trainee	Okay. Okay
Support 1	They're kinda just getting started.
...	
Dialogue Snippet 2	
...	
Support 3	Do you want to monitor her end-tidal with this?
Trainee	That would be great?
Support 3	I gave 250 milligrams of the dantrolene.
Trainee	Okay. Good.
Support 2	Here's the cold saline. I gotta go get the ice, okay?
Surgeon	Thank you, Matt.
Trainee	Dantrolene is in. We're going to cool the patient. The other thing is if we could.
...	
Dialogue Snippet 3	
...	
Trainee	You've given how much neo?
Support 1	I've given probably- this really started maybe a few minutes ago- probably getting like 500mics.
Trainee	Was she responding to it?
Support 1	She's responded a little bit. It's just kinda kept her around here but I think just because of the nausea, vomiting, and sepsis issue.
Trainee	Okay.
Support 1	36 year old lady who presented to the ER today with abdominal pain, nausea, vomiting. She was diagnosed with acute cholecystitis. They're afraid she's becoming septic.
...	

Table 7: Examples of the dialogues in ***CliniDial***.

(a) Support.

(b) Trainee.

(c) Surgeon.

Figure 9: Word cloud plots for frequent words uttered by different roles involved in ***CliniDial***.