Act2P: LLM-Driven Online Dialogue Act Classification for Power Analysis

Wenbo Zhang[1]**, Yuhan Wang**[1]**,**
[1]Beijing University of Technology, Beijing, China
{Zhangwenbo, Wangyuhan}@emails.bjut.edu.cn

Abstract

In team communication, dialogue acts play a crucial role in helping team members understand each other's intentions and revealing the roles and communication patterns within interactions. Although existing studies have focused on using Dialogue Act classification to capture the speaker's intentions, few have explored the underlying power dynamics reflected by these dialogue acts. To this end, we present an online Dialogue Act Classification and Dynamic Power Analysis framework—Act2P, which is based on large language model. The framework combines the zero-shot learning capability of LLMs and introduces an online feedback classification method that allows for online classification with iterative feedback to previous stages, achieving efficient and accurate classification without the labeled data. Additionally, we also propose the PowerRank algorithm, which quantifies power dynamics through a graph-based structure. Through comparative experiments with existing methods, we demonstrate the significant superiority of Act2P in online scenarios and successfully visualize dialogue power in online, clearly presenting the distribution and dynamic transfer of power. This framework provides new scientific insights and practical tools for optimizing team collaboration.

1 Introduction

Effective communication is crucial in team-based tasks, influencing collaboration efficiency and task outcomes. Analyzing interaction patterns can reveal underlying relationships, optimizing teamwork. Dialogue Act Classification (Searle, 1969) plays a key role in Natural Language Processing by identifying user intent. However, existing researchs primarily focus on explicit utterance functions (Witzig et al., 2024; Colombo et al., 2020; Fu et al., 2025), overlooking the implicit power dynamics embedded in dialogue. Different dialogue acts often involve power exertion, acceptance, or resistance, significantly affecting team collaboration and decision-making.Power has been extensively studied in sociology, management, and linguistics, with traditional research emphasizing stable hierarchical structures. However, power in team interactions is inherently dynamic, continuously evolving throughout a conversation. Members' speech patterns, responses, and engagement influence power distribution. Traditional DAC methods rely heavily on manually labeled data, limiting their adaptability across domains. Inconsistencies in annotation schemes further reduce transferability and generalizability, leading to performance degradation in new contexts. These challenges hinder the integration of dialogue act classification with power quantification, restricting the ability to analyze online power shifts in team interactions.

Large Language Models(OpenAI, 2023; Dubey et al., 2024; Liu et al., 2024; Guo et al., 2025) offer a breakthrough in addressing these challenges. Their zero-shot learning capabilities enable dialogue act classification without requiring extensive annotations, allowing for greater adaptability across different datasets. Prompt engineering (Wei et al., 2022; Reynolds and McDonell, 2021) enhances LLMs' ability to classify dialogue acts efficiently, providing online analytical support. Additionally, LLM facilitate power shift detection, advancing the study of dynamic power quantification in conversation.

To address the aforementioned limitations, this paper proposes the Act2P framework, an online dialogue act classification and dynamic power analysis method based on Large Language Model(LLM). Act2P leverages the powerful language understanding capabilities of LLM to achieve zero-shot classification of dialogue acts. One of its key innovations is the introduction of an online feedback classification method, which iteratively optimizes the model based on online feedback, enabling rapid adaptation to datasets with limited or no human annotation and effectively improving classification

Findings of the Association for Computational Linguistics: ACL 2025, pages 20494–20504
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

accuracy and dynamic adaptability. Additionally, the Act2P framework introduces a power dynamic quantification algorithm based on dialogue acts, which can capture and reflect the flow and changes of power among team members online, providing new methods and perspectives for power analysis in team communication.

Therefore, the main contributions of this paper are as follows:

- **Propose a online feedback classification method.** This study designs an online dialogue act classification and dynamic power analysis framework based on Large Language Models (LLMs) called Act2P. Its core innovation is the online feedback classification method, which significantly improves the accuracy and adaptability of dialogue act classification through online feedback corrections.

- **Designing the PowerRank algorithm for power dynamic quantification.**This algorithm uses dialogue acts to construct a graph structure, precisely depicting the power dynamics and transfer mechanisms within a team. It also explores the role of different granularities of dialogue act labels in power quantification, providing effective methods and directions for optimizing power analysis in team communication.

In summary, Act2P not only overcomes the limitations of traditional dialogue act classification methods in terms of annotation dependency and online applicability but also explores the power dynamics reflected within dialogue acts.. This framework provides a novel theoretical and practical tool for investigating dynamic interactions within teams.

2 Related work

Dialogue Act Classification: Dialogue Act Classification (DAC) is an important task in natural language processing. Many studies utilize neural network architectures and attention mechanisms to capture contextual information. Early research (Kumar et al., 2018; Chen et al., 2018) primarily used RNN and CRF to capture the relationships between utterances. Wang et al. (2020) proposed the HUH graph convolutional network, which improved dialogue act classification through a denoising mechanism. Raheja and Tetreault (2019) combined context-aware self-attention with hierarchical RNNs to model dialogue act semantics.

Research has gradually focused on the impact of dialogue space modeling on classification. He et al. (2021) proposed a speaker-turn-aware method that combines speaker information with utterance representations. Ghosal et al. (2019) used graph structures to integrate contextual information at the speaker level. Song et al. (2023) and Sun et al. (2021) used graph structures to learn the representations of utterance nodes, improving utterance representation.

Some studies have used multimodal information for recognition, such as the online multimodal dialogue act classification framework proposed by Miah et al. (2023), which combines transcribed text and multimodal features for training.

This study proposes using Large Language Models (LLMs) for zero-shot dialogue act classification. Compared to existing methods, LLMs enable efficient classification in the absence of labeled data.

Power Analysis: Power dynamics have long been an important research topic in fields such as organizational behavior, psychology, and computational linguistics. Hofstede's Power Distance Index (PDI)(Hofstede, 1984) provides a theoretical foundation for understanding power distribution in organizations and cultures, measuring the degree of power inequality and its acceptance.

In language interactions, researchers focus on how speakers use language to manifest and maintain power. Danescu-Niculescu-Mizil et al. (2012) introduced the Linguistic Coordination Model, which shows that low-power individuals tend to imitate the language style of high-power individuals. Boghrati and Dehghani (2018) proposed the Syntactic Alignment Model, which demonstrates that low-power individuals imitate not only vocabulary but also syntactic structures. Choi et al. (2020) analyzed the language patterns of leaders and followers, revealing how role settings dynamically influence power.

In the email domain, Lam et al. (2018) introduced the Power Networks framework, which combines neural network prediction models with contextual modeling to accurately predict power relations in email communications. Raut et al. (2020) used supervised learning to classify power based on semantic and structural features, while Wen et al. (2025) analyzed power propagation paths by constructing email communication networks.

This study focuses on power dynamics reflected in dialogue acts and employs LLMs for their quantification and visualization.

3 Framework

Act2P is an framework based on Large Language Models, designed for online dialogue act classification and dynamic power analysis.The framework classifies dialogue acts while integrating power quantification algorithms and dynamic visualization techniques to capture and analyze power distribution and shifts in team communication online, offering an efficient tool for collaboration optimization, as illustrated in Figure 1.

3.1 Dialogue Act Classification Module

3.1.1 Task Description:

The goal of the Dialogue Act (DA) classification module is to predict the functional or intentional category of each utterance within a conversation, such as statements, questions, commands, or affirmations.This is essential for understanding the semantic structure of a dialogue and the speaker's communicative intent.

Formally, given a conversation C consisting of n utterances, it can be represented as:

$$C = \{u_1, u_2, \ldots, u_n\} \quad (1)$$

where u_i denotes the i-th utterance. Each utterance consists of a text component x_t^i and contextual metadata x_c^i (e.g., speaker identity).

The goal of the classification task is to map each utterance u_i to a predefined DA label y_i. Mathematically, the classification task can be defined as:

$$f : u_i \rightarrow y_i, \quad \forall i \in \{1, 2, \ldots, n\} \quad (2)$$

where y_i represents the DA label assigned to utterance u_i, drawn from a set of fixed labels.

3.1.2 Method Description:

We adopt a large language model (LLM) with zero-shot learning capabilities for dialogue act classification, enabling it to perform classification without the need for task-specific fine-tuning. Building upon this foundation, we systematically explore prompt engineering techniques by designing multiple prompt strategies to guide the model in better understanding dialogue context and category semantics. Furthermore, we propose an online feedback classification method that incorporates current prediction results to dynamically adjust previous classifications. This mechanism enables the model to continuously refine its understanding of the dialogue flow, improving coherence, robustness, and classification accuracy, especially in multi-turn conversations where contextual dependencies and ambiguous class boundaries are common. Detailed prompt templates are available at https://github.com/wangyhby/Act2P.

Prompt Engineering: In the task of dialogue act classification, the design of prompts is crucial for the performance of large language models (LLMs). By using different prompt design methods, such as direct classification, category description, and context augmentation, the model can better understand the context of the dialogue, improving classification accuracy and robustness. These methods effectively help the model distinguish between semantically similar categories and enhance its ability to recognize dialogue acts that depend on context, thereby improving the model's adaptability and generalization ability.

Hierarchical Classification Enhancement: The core idea of the hierarchical classification enhancement method is to optimize the computational efficiency and classification accuracy by dividing complex classification tasks into two stages: coarse-grained and fine-grained classification. In the coarse-grained classification stage, the model first performs an initial classification of the dialogue text, identifying broader categories. In the fine-grained classification stage, the model further refines the results based on the coarse classification to achieve more specific classification outcomes. This staged processing approach not only effectively reduces the computational load but also significantly improves classification accuracy, particularly in multi-class and highly ambiguous dialogue act classification tasks, demonstrating stronger robustness and adaptability.

Online Feedback Classification: The online feedback classification method improves online classification by using current results to correct previous classifications. Unlike traditional static classification, which relies solely on the current input, this method incorporates past predictions, improving accuracy and robustness.

The process can be described as follows: The online feedback classification method improves online classification by using current results to correct previous classifications. Unlike traditional static classification, which relies solely on the current

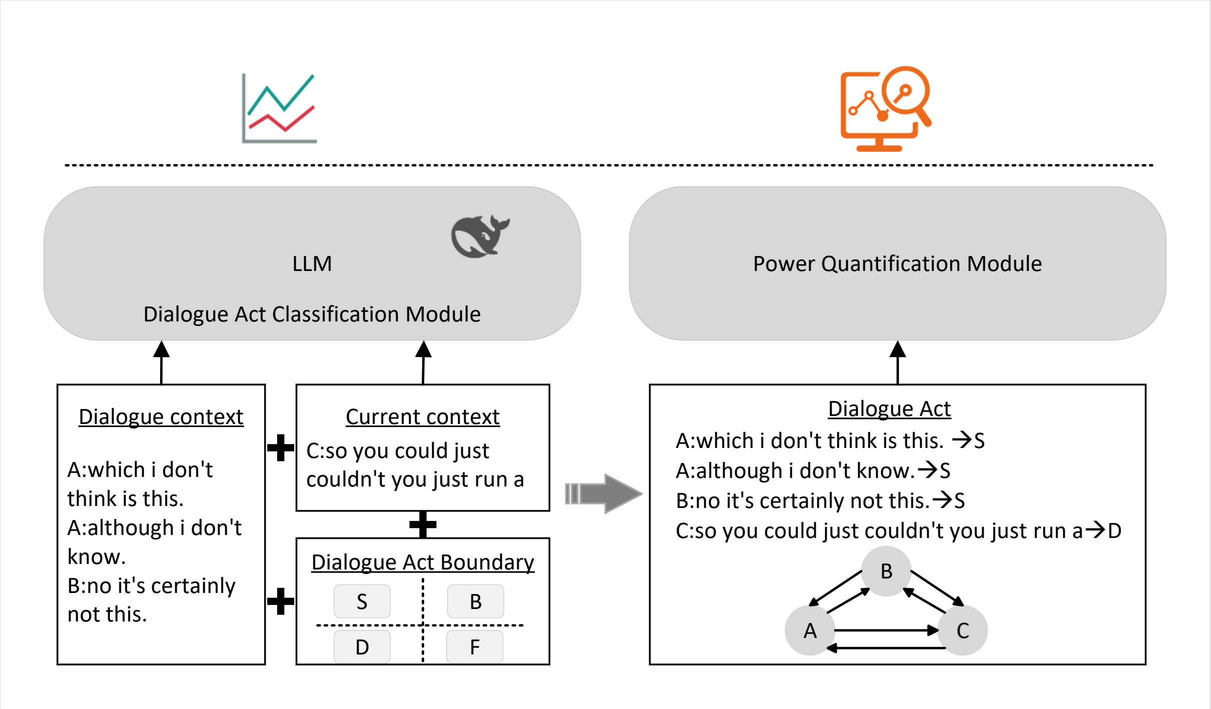
LLM
Dialogue Act Classification Module
Power Quantification Module
Dialogue context
A:which i don't think is this.
A:although i don't know.
B:no it's certainly not this.
Current context
C:so you could just couldn't you just run a
Dialogue Act Boundary
S
B
D
F
Dialogue Act
A:which i don't think is this. →S
A:although i don't know.→S
B:no it's certainly not this.→S
C:so you could just couldn't you just run a→D
B
A
C

Figure 1: Architecture of the Act2P Framework

input, this method incorporates past predictions, improving accuracy and robustness.

$$C_t = f_{pred}(X_t, C_{t-1}) \tag{3}$$

Where C_t is the current classification, f_{pred} is the classification function, X_t is the current input, and C_{t-1} is the previous classification. This allows the model to adjust based on prior predictions.

If new context affects the previous classification, it is updated as:

$$C_t^{new} = f_{update}(C_t, C_{t-1}, X_t) \tag{4}$$

Where C_t^{new} is the updated classification. This method enables the model to adapt dynamically and improve classification performance online.

3.2 Power Quantification Module:

3.2.1 Task Description:

Given a conversation $C = \{u_1, u_2, \ldots, u_n\}$ consisting of n utterances, where each utterance u_i has been labeled with a corresponding dialogue act label y_i by the dialogue act classification module, and each utterance is associated with a set of speakers $S = \{s_1, s_2, \ldots, s_m\}$, where each s_i represents a speaker. The task of the Power Quantification Module is as follows:

- Assign a corresponding power weight w_i to each dialogue act based on its pragmatic function.
- Calculate the power value changes for each speaker by considering the sequence of the dialogue and the interaction patterns between participants.
- Generate power dynamics curves and visualizations to intuitively reflect the flow of power throughout the dialogue.

3.2.2 Powerrank

The PowerRank algorithm is based on the traditional PageRank (Berkhin, 2005) algorithm, which evaluates the importance of power by calculating node relationships and interactive behaviors. To better reflect the real-time nature of the algorithm and its insensitivity to certain categories, we use the LLM to dynamically adjust the power distribution between participants in the conversation, further enhancing real-time responsiveness. This ensures that power distribution is adjusted promptly during the conversation, accurately reflecting dynamic changes. The individual power value P_i of participant s_i is updated iteratively as follows:

$$
\begin{aligned}
P_i(t+1) &= (1-\alpha)\cdot P(s_i,s_j) \\
&\quad + \alpha\cdot\sum_{j\in N(i)}\frac{P_j(t)\cdot w_{ij}}{d_{\text{out}}(j)}
\end{aligned}
\tag{5}
$$

Here, $P_i(t+1)$ represents the power value of node i at time $t+1$. The parameter α is a damping factor, typically set to 0.85, which balances the weight between personalized preferences and the network structure. w_{ij} denotes the weight between nodes i and j, quantifying the influence or strength of the connection between them. The set of neighboring nodes $N(i)$ includes all nodes that are directly connected to node i, while $d_{\text{out}}(j)$ represents the out-degree of node j, which is the number of edges emanating from node j.

$P(s_i,s_j)$ represents the personalized preference value enhanced by LLM. More specifically, for the power flow preference vector $P(s_i,s_j)$, if the conversation involves only two participants, the LLM triggers the power enhancement mechanism. The LLM assesses that the power values of the two speakers are stronger, scoring the current power of each speaker to obtain $P_{\text{LLM}}(s_i)$ and $P_{\text{LLM}}(s_j)$, resulting in:

$$
P(s_i,s_j)=\begin{cases} P_{LLM}(s_i) & \text{if} S=s_i \\ P_{LLM}(s_j) & \text{if} S=s_j \\ 0 & \text{otherwise} \end{cases}
\tag{6}
$$

The Powerrank algorithm is as follows:

Algorithm 1 PowerRank Algorithm

```
Input: Graph $G(V,E)$, initial power values $b_v$, damping factor $\alpha$, convergence threshold $\epsilon$, maximum iterations $max\_iter$, and power flow preference vector $p(s_i,s_j)$
Output: Return the final PowerRank scores $r$, where $r[i]$ is the power score for node $v_i$.
$t \leftarrow 0$
Calculate the normalized matrix $\tilde{W}$ of $W$ to make $\sum_{i=1}^{n}\tilde{w}_{ij}=1, \forall v_j \in V$
while $\|r(t+1)-r(t)\|_1 \geq \epsilon$ and $t < max\_iter$ do
    Initialize $r^0[i] \leftarrow b_v[i]$ for all $v_i \in V$
    For all $v_i \in V$, update
        $r_i^{(t+1)} = (1-\alpha)\cdot p(s_i,s_j) + \alpha\cdot\sum_{j\in\text{In}(i)} W[j,i]\cdot r_j^{(t)}$
    $t \leftarrow t+1$
return $r_i$ where $r[i]$ is the power score for node $v_i$
```

4 Experimental Setup

4.1 Datasets

We conduct experiments and report results based on the Meeting Recorder Dialog Act (MRDA) dataset (Shriberg et al., 2004). MRDA is a publicly available benchmark dataset for multi-party conversation audio, widely used in research on online dialog act (DA) classification.We provide the statistics of the datasets in Table 1.

The MRDA dataset contains 75 multi-party meetings, each considered as an independent conversation. The average length of each conversation is 1442.5 utterances. The dataset provides both manually annotated transcription text and corresponding audio signals, offering robust support for online DA classification tasks. We partition the dataset into 51 training sets, 12 validation sets, and 12 test sets.The MRDA dataset adopts a labeling system consisting of 52 dialog act labels , which can be divided into multiple hierarchical levels based on different granularities. Specifically, these dialog act labels are clustered into 12 general labels and 5 basic labels. We discuss whether the different granular label divisions can reveal behavioral patterns and power dynamics in finer-grained dialogues, providing a more comprehensive perspective for power quantification research.

4.2 Evaluation Metric

In the dialogue act classification task, we choose accuracy as the primary evaluation metric, following previous studies for comparison. In power quantification analysis, due to the lack of relevant research for comparison, we have defined our own evaluation criteria. These criteria assess the model's effectiveness and prediction accuracy through two dimensions.

4.2.1 Power Distribution Validity Verification:

This dimension evaluates whether power curves at different granularities (5, 12, and 52 categories) reflect participants' actual power distribution, focusing on identifying dominant participants, especially the professor. By analyzing power rankings, we ensure the label system aligns with actual power distribution and the model accurately reflects each participant's power position.

Dataset	|C|	|L|	Dialogs		Utterances	
			Train	Test	Train	Test
MRDA	5/12/52	1442.5	51	12	75K	16.4K

Table 1: different granularities of DA labels |C|, utterances per dialog |L|, and number of dialogs and utterances in each split

4.2.2 Power Ranking Prediction Accuracy Evaluation:

The second dimension evaluates the gap between the power rankings predicted by the LLM and the actual DA labels. We quantify the deviation by comparing the power rankings predicted by the LLM with those calculated from the real DA labels, using the following three evaluation metrics to assess the prediction results.

Rank Accuracy (RA): In calculating Rank Accuracy (RA), we use the following formula to quantify the match between the predicted rankings and the true rankings for each turn in the dialogue. The formula computes the accuracy by counting the items where the predicted rankings match the true rankings, as expressed below:

$$RA = \frac{1}{turns} \sum_{i=1}^{turns} \sum_{j=1}^{n} 1(LO_i[j] = RO_i[j]) \quad (7)$$

Where: $turns$ denotes the total number of dialogue turns. n denotes the total number of participants. LO_i is the order of speakers predicted by the LLM for each turn. RO_i is the order of speakers based on the true labels for each turn.

Dominant Speaker Accuracy(DSA) Dominant Speaker Accuracy (DSA) measures whether the model correctly identifies the dominant speaker in each turn. In a conversation, the dominant speaker typically leads the discussion, decision-making, and topic guidance. Accurately predicting the dominant speaker is crucial for capturing the power dynamics, as their speech and actions often influence the direction of the entire dialogue. The formula is as follows:

$$DSA = \frac{1}{turns} \sum_{i=1}^{turns} 1(preD_i = realD_i) \quad (8)$$

Where: DSA denotes the accuracy of predicting the dominant speaker (the one with the highest power). $turns$ represents the total number of dialogue turns. $preD$ is the dominant speaker predicted by the model in the i-th turn. $realD$ is the actual dominant speaker according to the true labels in the i-th turn.

Spearman Rank Correlation: Spearman Rank Correlation (Zar, 2005) measures the "relative order" between predicted and actual power rankings. It focuses on rank relationships rather than exact matches, allowing for a finer assessment of differences, especially when there are subtle changes in the power ranking. This metric provides a comprehensive evaluation of power ranking differences. The formula is as follows:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \quad (9)$$

Where: d_i is the rank difference between the two variables in each observation group. n is the number of observations.

4.3 Implementation Details

We chose to conduct experiments using the APIs of large language models, which eliminates the need for GPU resources.

5 Results and Analysis

We evaluated the performance of the proposed LLM-based zero-shot classification framework in online domain adaptation classification tasks (online DA classification) and compared it with current related research.In our experiments, we selected the average of 10 trial results for evaluation.

5.1 Dialogue Act Classification results

We implemented DAC using Deepseek-v3 and explored the accuracy of different methods, including Prompt Optimization, Hierarchical Classification, and Online Feedback, on the MRDA dataset.From Table 2, we can observe that,optimizing the prompts significantly improved performance. Initially, we used simple prompts, but later added category descriptions, and hierarchical recognition to enhance clarity. Notably, our designed Online Feedback Classification strategy improved the model's accuracy from 70.30% to 84.53% with Basic_label, with similar improvements observed for General_label and Full_label. The reason for choosing Deepseek for this experiment is detailed in Appendix A, where comparisons with other large models are provided.

In the online feedback classification, only F and D are easily confused within the main categories and require contextual responses for accurate judgment. Therefore, we incorporated online feedback correction for Disruption(D) and FloorGrabber(F). After the correction, as shown in the figure 2, the F1 scores of both categories have slightly improved, which validates the effectiveness of our method's online feedback correction capability.

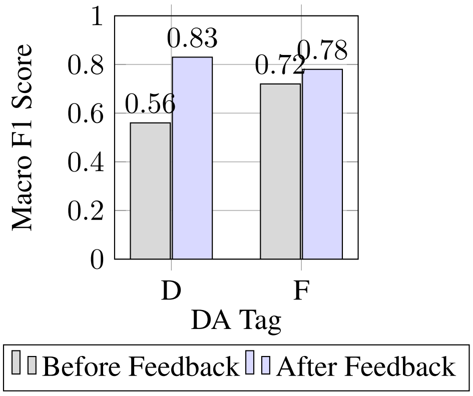

Figure 2: Comparison of F1 Scores for D and F Categories

Currently, most research on DA classification experiments is based on the Basic_label of the MRDA dataset and primarily uses supervised learning models, relying on large amounts of manually labeled data to train and fine-tune models to improve classification performance. In contrast, this paper explores a zero-shot DA classification method based on LLM, which does not rely on manual labeling but instead achieves accurate DA classification through prompt engineering combined with an online feedback classification. In Table 3,although our model has not fully surpassed supervised learning models, compared to these methods, Our zero-shot approach significantly contributes to manually labelled data reduction(MLDR).

5.2 Power Quantification Results

This section presents the experimental results of power quantification using the Deepseek-v3 model under different label granularities. We compared the impact of different label granularities on power dynamics and explored the model's performance in capturing and quantifying power flow in conversations. In the experiment, we used the PowerRank algorithm to quantify the power distribution of each speaker in the dialogue and visualized the changes in power.

5.2.1 Power Distribution Validity Verification:

The MRDA dataset comes from academic discussion meetings, where professors typically hold more power than other students. This provided a reference for power judgment in the model. We validated the effectiveness of different label granularities in capturing power by predicting whether the dominant speaker was a professor. In the test data, 10 meetings included a professor role, and we used the LLM to predict power dynamics under different label granularities, quantifying the final dominant role using the Pagerank algorithm. We found that the power validity corresponding to the General_label is 90%, while the validity for other labels is around 80%. And the details of the weights and graphical design can be found in Appendix B.

5.2.2 Power Ranking Prediction Accuracy Evaluation:

In addition to assessing power flow effectiveness, we evaluate the discrepancy between predicted and true power rankings. Finer label systems, like Full_label, capture subtle power shifts but increase complexity, lowering classification accuracy. Simpler labels, like Basic_label, improve accuracy but may miss detailed power dynamics. The choice of label granularity must balance detail with accuracy to avoid errors in power quantification. By evaluating three metrics, we assess the differences between predicted and true power, helping us choose the best label granularity for improved model performance. The experimental results are shown in figure 3.

Based on the evaluation results, this framework recommends General_label as the standard label granularity for power quantification. The rationale behind this choice is that General_label strikes a good balance between capturing power flow effectiveness and ranking accuracy. Although it slightly lags behind Full_label in Rank Accuracy (RA), it excels in Dominant Speaker Accuracy (DSA) and is more stable across different contexts. Furthermore, compared to Full_label, General_label simplifies the classification task, improving accuracy and reducing errors due to excessive label granularity. Overall, General_label effectively captures power dynamics while maintaining model

Model	Basic_label	General_label	Full_label
Deepseek-v3	70.30	60.65	29.12
Deepseek-v3 (Prompt Optimization)	80.34	70.23	36.43
Deepseek-v3 (Hierarchical Classification)	82.83	73.56	44.00
Deepseek-v3 (Online Feedback)	**84.53**	**75.97**	**45.53**

Table 2: Model Performance on Different Label Granularities

Model	Accuracy	MLDR
He et al. (2021)	92.2	0%
Chapuis et al. (2020)	**92.4**	0%
Miah et al. (2023)	91.8	0%
Our model	84.53	**99.9%**

Table 3: Comparison of model accuracy and manually labelled data reduction

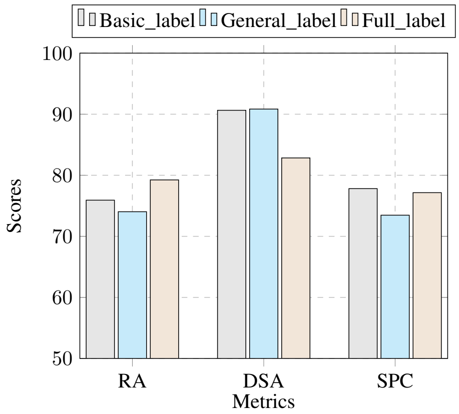

Figure 3: Comparison of Evaluation Metrics for Different DA Labels

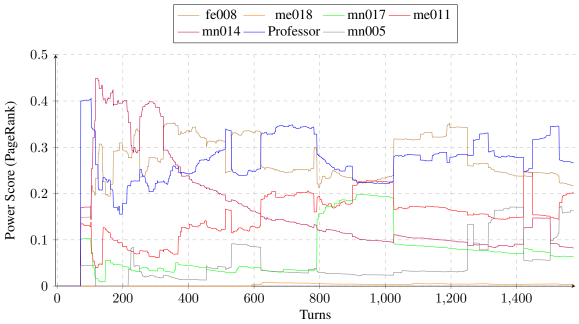

Figure 4: Power Awareness in Turns for Different Speakers

efficiency, making it the most suitable label granularity for real-time and accurate power analysis in practical applications

5.2.3 Visualization:

In this section, based on the results from Sections 5.2.1 and 5.2.2, we only present the results for General_label. We use charts to display the power trend and the power share of each participant under General_label, providing an intuitive presentation of power quantification results. These visualizations allow us to clearly observe the impact of label granularity on the ability to capture power distribution,which can be found in figure 4.

In the Figue 4, we can observe that as the conversation progresses, the power values change in real-time, with participants' power fluctuating significantly over time. While the professor's power may not always be the highest during certain discussion phases, overall, the professor's power remains at a relatively high level, especially during key moments such as decision-making and topic guidance. This suggests that, although the professor may not dominate in some discussions, overall, they remain the dominant power figure, with their authority exhibiting strong stability and influence throughout the conversation. In contrast, the power values of other participants fluctuate more, reflecting their supporting roles in the discussion. Therefore, while power distribution in the conversation fluctuates, the professor's power remains dominant in the overall discussion, reflecting their leadership and guiding role in academic discussions.

6 Conclusion

We propose an online dialogue act classification and dynamic power analysis framework, Act2P, based on large language models (LLM), aimed at effectively capturing and quantifying power dynamics in real-time team communication. We demonstrate that the framework, through the design of efficient prompts and online feedback classification, can quickly adapt to different conversational scenarios and perform accurate classification in a zero-shot learning setting. By incorporating power quantification mechanisms, we can monitor and analyze power shifts in real-time, providing in-depth insights into team communication patterns. Future work could explore ways to improve dialogue act classification accuracy under different granularities of labels, further enhance the precision of power quantification, and attempt to integrate other collaborative features such as speech information to strengthen the framework's real-time capability and adaptability.

7 Limitations

Lack of Support for Speech Features: The current framework is based solely on text-based large language models for dialogue act classification and power quantification analysis, without incorporating speech features such as emotion, tone, and speech rate. However, these non-verbal features in speech play a crucial role in conveying intent and power dynamics. Therefore, the lack of support for speech features may limit the model's performance in complex conversational scenarios, especially in situations where tone, emotional shifts, and speaker intentions need to be analyzed.

Lack of Existing Research on Power Quantification Based on Dialogue Acts: This study combines dialogue act classification with power quantification, but there is currently a lack of in-depth research on how to closely integrate dialogue acts (DA) with power analysis. Due to the absence of sufficient reference frameworks, power analysis cannot be compared against baselines. Future research needs to further explore methods for power quantification based on dialogue acts to enhance the depth and comparability of research in this field.

8 Ethical Considerations

This work involves the use of Large Language Models (LLMs) for dialogue act classification, which raises potential ethical concerns. While LLMs offer significant advantages in automating classification tasks, they can be misused for malicious purposes, such as generating fraudulent content or spreading misinformation. Additionally, LLMs may produce hallucinations, leading to incorrect or biased classifications. These challenges highlight the need for careful consideration in deploying LLM-based systems, ensuring they are used responsibly and that safeguards are in place to mitigate potential risks. It is essential to validate and monitor the performance of LLMs to prevent misuse and ensure they contribute positively to real-world applications.

References

Pavel Berkhin. 2005. A survey on pagerank computing. *Internet mathematics*, 2(1):73–120.

Reihane Boghrati and Morteza Dehghani. 2018. Follow my language! effect of power relations on syntactic alignment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. volume EMNLP 2020.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*, pages 1514–1525.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7594–7601.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Changzeng Fu, Yikai Su, Kaifeng Su, Yinghao Liu, Jiaqi Shi, Bowen Wu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2025. HAM-GNN: A hierarchical attention-based multi-dimensional edge graph neural network for dialogue act classification. *Expert Syst. Appl.*, 261:125459.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. pages 154–164. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zihao He, Leili Tavabi, Kristina Lerman, and Mohammad Soleymani. 2021. Speaker turn modeling for dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2150–2157. Association for Computational Linguistics.

Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*, volume 5. sage.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Michelle Lam, Catherina Xu, and Vinodkumar Prabhakaran. 2018. Power networks: A novel neural architecture to predict power relations. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 97–102.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Md Messal Monem Miah, Adarsh Pyarelal, and Ruihong Huang. 2023. Hierarchical fusion for online multimodal dialog act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7532–7545.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vipul Raheja and Joel R. Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3727–3733. Association for Computational Linguistics.

Purva Raut, Rohit Chawhan, Tejas Joshi, and Pratik Kasle. 2020. Classification of power relations based on email exchange. In *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 486–489. IEEE.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Rui Song, Fausto Giunchiglia, Lida Shi, Qiang Shen, and Hao Xu. 2023. Sunet: Speaker-utterance interaction graph neural network for emotion recognition in conversations. *Engineering Applications of Artificial Intelligence*, 123:106315.

Yang Sun, Nan Yu, and Guohong Fu. 2021. Integrating rich utterance features for emotion recognition in multi-party conversations. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pages 51–62. Springer.

Dong Wang, Ziran Li, Haitao Zheng, and Ying Shen. 2020. Integrating user history into heterogeneous graph for dialogue act recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4211–4221.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Tao Wen, Yu-wang Chen, Tahir Abbas Syed, and Darminder Ghataoura. 2025. Examining communication network behaviors, structure and dynamics in an organizational hierarchy: A social network analysis approach. *Information Processing & Management*, 62(1):103927.

Philine Witzig, Rares Constantin, Nikola Kovacevic, and Rafael Wampfler. 2024. Multimodal dialog act classification for digital character conversations. In *ACM Conversational User Interfaces 2024, CUI 2024, Luxembourg, July 8-10, 2024*, page 12. ACM.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

A Model Selection Based on Benchmark Dataset Performance

To select the most accurate model for recognizing dialogue acts (DA), we conducted tests on several large models that performed exceptionally well across multiple metrics. The models were evaluated based on their performance in DA classification tasks, considering both accuracy and robustness in handling various dialogue scenarios. As shown in the table 4, after analyzing the results, we selected the Deepseek-v3 model, which achieved the highest accuracy, proving to be the most effective model for our specific needs.

LLM	Basic	General	Full
Llama3.1-405b	56.04	43.79	24.32
Gpt-4o	62.95	53.15	25.66
Qwen2.5-Max	64.44	51.48	23.87
Deepseek-v3	**82.83**	**73.56**	**44.00**

Table 4: Model Performance on DA Recognition with Different Granularities

B Weight and Relationship Graph Design

Our framework assigns different weights to dialogue act types and guides graph construction, providing a flexible and adaptive approach to power quantification analysis. To automate the weight assignment process and minimize human intervention, we replaced manual weight assignment with a large language model (LLM). The table 5 below shows the detailed weight information for the general_label, illustrating how the model adjusts its weight distribution to more accurately represent the power relationships and dynamics present in the conversation. This method allows for more robust analysis and can be easily adapted to different types of dialogues, demonstrating the effectiveness of LLM in handling complex, dynamic interactions.

Label	Weight	Power Flow
b	0.5	Current → Previous
fh	0.2	Self power increase
fg	0.5	Previous → Current
qy	0.3	Previous → Current
qw	0.3	Previous → Current
qr	0.3	Previous → Current
qrr	0.3	Previous → Current
qo	0.3	Previous → Current
qg	0.3	Previous → Current
h	0.2	Self power increase
%	0.5	Current → Next

Table 5: Weight and Power Flow for Different Dialog Act Labels