On the Interpretability of Deep Learning Models for Collaborative Argumentation Analysis in Classrooms

Deliang Wang and Gaowei Chen
Faculty of Education, The University of Hong Kong
wdeliang@connect.hku.hk

Abstract

Collaborative argumentation holds significant potential for enhancing students' learning outcomes within classroom settings. Consequently, researchers have explored the application of artificial intelligence (AI) to automatically analyze argumentation in these contexts. Despite the remarkable performance of deep learning models in this task, their lack of interpretability poses a critical challenge, leading to teachers' skepticism and limited utilization. To cultivate trust among teachers, this PhD thesis proposal aims to leverage explainable AI techniques to provide explanations for these deep learning models. Specifically, the study develops two deep learning models for automated analysis of argument moves (claim, evidence, and warrant) and specificity levels (low, medium, and high) within collaborative argumentation. To address the interpretability issue, four explainable AI methods are proposed: gradient sensitivity, gradient input, integrated gradient, and LIME. Computational experiments demonstrate the efficacy of these methods in elucidating model predictions by computing word contributions, with LIME delivering exceptional performance. Moreover, a quasi-experiment is designed to evaluate the impact of model explanations on user trust and knowledge, serving as a future study of this PhD proposal. By tackling the challenges of interpretability and trust, this PhD thesis proposal aims to contribute to fostering user trust in AI and facilitating the practical implementation of AI in educational contexts.

1 Introduction

Collaborative argumentation refers to a dialogue-based activity in which participants engage in constructing, critiquing, and reconciling arguments through social interactions(Rapanta and Felton, 2022). Within classroom settings, empirical evidence consistently demonstrates that collaborative argumentation fosters critical thinking and knowledge construction by integrating learned facts and knowledge, reasoning, justifying, and negotiating (Asterhan and Schwarz, 2016; Gao et al., 2023). To fully harness its potential, teachers are encouraged to instruct students how to argue, facilitate students' engagement, and effectively manage collaborative argumentation (Asterhan et al., 2020; Rapanta and Felton, 2022). However, it has been observed that many teachers face challenges to master the necessary skills to effectively promote collaborative argumentation in their classrooms (Lugini, 2021; Oyler, 2019). To address this issue, some researchers propose recording and analyzing argumentative discussions utterance by utterance, employing an evaluation rubric to assess whether adjustments in teaching strategies and support interventions are needed for future classes (Lampert et al., 2010). However, for teachers who are already burdened with daily responsibilities, conducting such laborious manual analyses is not feasible.

To tackle this challenge, researchers have turned to the application of natural language processing (NLP) and artificial intelligence (AI) techniques to automate the analysis of classroom argumentation (McLaren et al., 2010; Nazaretsky et al., 2023; Wang et al., 2024b). Initially, conventional machine learning techniques were employed to examine various aspects of teachers' discourse and students' engagement (Olney et al., 2017; Reilly and Schneider, 2019). Subsequently, deep learning techniques were increasingly adopted to achieve more accurate analysis. For instance, Nazaretsky et al. (2023) utilized Transformer-based neural networks to train models that automatically analyze teachers' ability to attend to students' ideas. Despite these advancements, it has been observed that teachers are hesitant to trust the decisions made by such models (Nazaretsky et al., 2021, 2022). They express significant concerns regarding the lack of transparency and interpretability in these models, which undermines their trust (Nazaretsky

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 506–516

et al., 2022; Jackson and Panteli, 2023). Deep learning models often consist of complex structures with multiple layers interconnected by thousands or even millions of neurons, making them appear as “black boxes" that provide users with direct decisions without revealing the underlying process of prediction. The lack of understanding regarding the internal workings and individual decisions of these models likely leads to user distrust and underutilization of these tools, which can have a significant impact on the deployment of AI (Qin et al., 2020) and teacher instruction in this particular case.

To enhance user trust in AI-powered models and systems, researchers have proposed leveraging explainable AI (xAI) to unravel the working mechanisms and individual decisions, providing explanations of AI (Meske et al., 2022). As a result, various interpreting methods have been developed (Arrieta et al., 2020). A systematic review conducted by Haque et al. (2023) demonstrates that explanations provided by xAI effectively increase user trust and transparency in AI tools. Desipte the significant progress, the interpretability issue of deep learning models for collaborative argumentation analysis in the classroom context remains largely unexplored.

Hence, this PhD thesis proposal aims to investigate whether explainable AI methods can be effectively utilized to explain deep learning models for classroom collaborative argumentation analysis. Specifically, we train two deep learning models on authentic transcripts of classroom collaborative argumentation to automatically analyze argumentative moves (i.e., claim, warrant, and evidence) and specificity levels (i.e., low, medium, and high). Subsequently, we employ four interpreting methods — gradient sensitivity, gradient input, integrated gradient, and LIME — to explain the model predictions by quantifying the contributions of input. The experimental results demonstrate that all four interpreting methods effectively explain the model predictions, with the LIME method yielding the most competitive outcomes. Furthermore, we design a quasi-experiment to evaluate the impact of explanations on user trust in and knowledge of the AI-powered collaborative argumentation model. We aim to contribute to addressing the interpretability challenge in the field of AI-supported classroom teaching, potentially fostering user trust in AI and facilitating the practical application of AI in teaching contexts.

2 Related Work

2.1 AI in classroom interaction

Many researchers have employed AI techniques to examine and analyze diverse facets of classroom interaction, with the aim of providing timely and valuable feedback to enhance teaching and learning. One fundamental approach involves using automatic speech recognition techniques to transcribe classroom recordings, encompassing teacher questions (Blanchard et al., 2015) and student speech (Evers and Chen, 2022). Additionally, researchers have investigated features of teacher discourse, including support types (Hunkins et al., 2022), uptake (Demszky et al., 2021), talk moves (Suresh et al., 2019), and instructional activities (Xu et al., 2020). Moreover, they have also examined characteristics of student utterances, such as speech acts (Shan et al., 2023), creativity (Chien et al., 2020), and sentiment (Huang et al., 2021). In the realm of classroom collaborative argumentation, researchers have explored modeling collaboration quality (Reilly and Schneider, 2019), knowledge graph (Zhen et al., 2021), and problem solving skills (Pugh et al., 2022).

Conventional machine learning algorithms, including random forest, naive Bayes, and support vector machine (SVM), have typically been employed for analyzing classroom interaction. Nonetheless, these algorithms necessitate manual selection of linguistic features and yield limited performance. Over the past decade, there has been an increasing adoption of deep learning algorithms, such as Transformer, Bert, and recurrent neural networks (Wang and Chen, 2024). In comparison to conventional machine learning algorithms, deep learning algorithms have demonstrated stronger performance across various tasks. However, as mentioned earlier, the opaque decision-making process of deep learning models engenders user distrust, thereby impeding their practical deployment and application (Wang et al., 2024b). Recently, large language models (LLMs) have exhibited remarkable capabilities in comprehending and processing natural language. Consequently, some studies have investigated their application in classroom interaction, such as detecting student talk moves (Wang and Demszky, 2023), evaluating teacher coaching (Wang et al., 2023b), and estimating instructional support (Hou et al., 2024; Whitehill and LoCasale-Crouch, 2023). However, there is still room for improvement in their performance.

2.2 Explainable AI

Researchers in explainable AI (xAI) propose a set of machine learning techniques that not only produce high-performing models but also enable humans to understand, trust, and manage the emerging AI tools effectively (Arrieta et al., 2020). xAI techniques can be categorized into ante-hoc and post-hoc explainability based on the degree of interpretability of AI models — how well humans can comprehend them (Burkart and Huber, 2021). Ante-hoc explainability pertains to self-explaining models that possess architectural interpretability (Alvarez Melis and Jaakkola, 2018), including logistic or linear regression, rule-based learning models, and general additive models. On the other hand, post-hoc explainability focuses on enhancing the interpretability of models that are not inherently transparent by employing external methods (Arrieta et al., 2020).

In addition, xAI techniques can also be classified as model-agnostic or model-specific, depending on the range of models they can explain. Model-agnostic methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), can be applied to all supervised learning models, while model-specific methods, such as LRP (Bach et al., 2015) and DeepLIFT (Shrikumar et al., 2017), are tailored to models with specific structures. Furthermore, xAI techniques can be divided into global and local methods (Lu et al., 2023). Global methods, such as knowledge instillation (Liu et al., 2018) and rule extraction (Bastani et al., 2017), aim to explain the inner workings of the entire model, whereas local methods, such as gradient sensitivity (Li et al., 2016) and LIME, provide interpretations of individual decision-making processes (Adadi and Berrada, 2018).

The utilization of these xAI techniques for providing explanations of AI models has been demonstrated to enhance user trust and understanding of AI models and systems across various domains (Haque et al., 2023), including the field of education (e.g., Conati et al., 2021; Lu et al., 2024; Ooge et al., 2022). In the context of classroom interaction, some studies have also explored the application of xAI techniques to unravel predictions of talk moves made by deep learning models (Wang et al., 2023a, 2024a). However, limited attention has been devoted to addressing the interpretability challenge of deep learning models in the analysis of collaborative argumentation within classrooms, which has the potential to significantly impact the quality of teaching and learning. Therefore, this study aims to investigate the feasibility of utilizing xAI techniques for this particular problem and designs an experiment to assess the effects of explanations on teachers and students, with the goal of facilitating future practical implementation.

3 Method

3.1 Data

We selected a publicly accessible corpus known as *Discussion Tracker* (Olshefski et al., 2020) to construct and elucidate deep learning models for analyzing collaborative argumentation in classroom environments. This corpus comprises 108 meticulously transcribed multi-party discussions conducted in American high school English language arts classes, collected between 2019 and 2022 (Lugini, 2021). The student discourse has been segmented into turns, which represent the sequential order in which individuals participate in the conversation. Turns containing collaborative argumentation have been further divided into argument discourse units, each annotated using a well-established coding scheme for argument moves and specificity. The argument moves are labeled as claim, evidence, and warrant. Specificity encompasses the presence of four key elements: (1) specificity towards a particular character or scene, (2) notable qualifications or elaborations, (3) usage of content-specific terminology (e.g., text quotes), and (4) a series of supporting reasons (Lugini et al., 2019; Olshefski et al., 2020). The specificity levels are classified as low, medium, or high. A comprehensive overview of the definition, examples, and quantities of argument moves and specificity within the corpus can be found in Table 1 and 2. The selection of argument moves and specificity for AI modeling is based on their significant impact on enhancing students' learning outcomes (Lee, 2006). For instance, automatically identifying students' argument moves during discussions can offer insights into their argumentative structures. By intervening when their arguments are poorly structured, teachers can enhance the quality of their argumentation. Similarly, the specificity of argument moves is closely linked to the quality of the discussion (Chisholm and Godley, 2011). During the construction of deep learning models for analyzing argument moves and specificity, we employed a random selection process to allocate 90% of the

data from the corpus for model training purposes, while the remaining 10% was set aside for model testing.

Table 1: Argument moves in the *Discussion Tracker* corpus (Lugini et al., 2019; Olshefski et al., 2020).

Label	Definition	Example	Number
Claim	An arguable statement that puts forth a specific understanding of a text or subject matter.	Also, at that same point, I feel like guilt overall was another one of the Nazis' tactics or goals at the end.	8,207
Evidence	Facts, records, textual citations, or testimonies employed to substantiate or validate a claim	I pulled out a quote that said, "His last words had been my name. He had called out to me and I don't answer".	3,043
Warrant	Rationales that explain how a particular instance of evidence bolsters a specific assertion.	This was nice because it wasn't like, "The Jewish kid running next to me", like that kid had a name. So, that was great.	1,385

Table 2: Specificity in the *Discussion Tracker* corpus (Lugini et al., 2019; Olshefski et al., 2020).

Label	Definition	Example	Number
Low	A statement that does not include any of the four components	It makes us think about what he said.	5,853
Medium	A statement that achieves any one of the four elements.	Like she's not even caring about them, she's caring about Willy.	4,250
High	A statement that clearly fulfills at least two elements of specificity.	They honestly don't really have a characterization because I feel like they don't really have like personalities or connections with other people.	2,532

3.2 Model

According to the systematic review conducted by Wang et al. (2024b), Bert has emerged as the most widely utilized deep learning model for analyzing classroom interaction. Therefore, for this study, we opted to adopt BertForSequenceClassification (Devlin et al., 2018) as the baseline model to construct and explain deep learning models specifically designed for analyzing argument moves and specificity within collaborative argumentation in the classroom.

Specifically, we set the student utterances as the input for both models, while the output consisted of predicted labels for argument moves or specificity, along with their corresponding probabilities. During the training of the models, we utilized AdamW as the optimizer, with 8 epochs, a batch size of 32, and a learning rate of 4e-4. The implementation of the code was carried out in Python 3.8, utilizing the PyTorch and HuggingFace libraries.

Given the focus of this study was not on training a deep learning model with state-of-the-art performance, we did not conduct parameter optimization or cross-validation. Following the training process, the model for argument move analysis achieved an accuracy of 0.7910 and an F1 score of 0.7503, while the model for specificity analysis attained an accuracy of 0.7152 and an F1 score of 0.6820.

3.3 Interpreting method

To explain the deep learning models developed for analyzing argument moves and specificity, we employed four local and generic interpretation methods: gradient sensitivity (GS) (Li et al., 2015), gradient input (GI) (Kindermans et al., 2019), integrated gradient (IG) (Sundararajan et al., 2017), and LIME (Ribeiro et al., 2016). The inclusion of these local and generic methods was driven by two key considerations. First, given the diverse range of deep learning models utilized for collaborative argumentation, model-specific xAI methods can be applied to other models regardless of their internal structures. Second, the convergence of multiple local explanations enables a comprehensive understanding of the overall functioning of the entire model.

Formally, let us consider a student's argumentative utterance denoted as u, which consists of n tokens. We represent the embedding of the utterance as v, with each token's embedding indicated as v_i ($v_i \in R^m$), where i denotes the token's position. The well-trained deep learning model f predicts the label of argument move or specificity, denoted as l, along with its corresponding probability $f_l(v)$. The methods of gradient sensitivity, gradient input, integrated gradient, and LIME differ in their approaches to calculating the contribution of each token towards the predictions.

3.3.1 Gradient sensitivity (GS)

The gradient sensitivity (GS) method (Li et al., 2015) assumes that if a feature holds importance for the model's prediction, even a slight change in that feature will lead to significant differences in the prediction. Consequently, this method consid-

ers the gradients of the features as their respective contributions to the predictions, as illustrated in Equation 1, where j denotes the j-th dimension in v_i. The contribution of the i-th token in the input utterance is then determined by summing up all the feature gradients in v_i, as shown in Equation 2.

$$C_{GS}(v_{ij}) \approx \frac{\partial f_l(v)}{\partial v_{ij}} \tag{1}$$

$$C_{GS}(v_i) \approx \sum_{j=1}^{m} \frac{\partial f_l(v)}{\partial v_{ij}} \tag{2}$$

3.3.2 Gradient input (GI)

Building upon the GS method, Kindermans et al. (2019) propose an alternative perspective on feature contribution, suggesting that it can be viewed as the product of sensitivity (i.e., feature partial derivative) and saliency (i.e., feature value), as demonstrated in Equation 3. Alternatively, the gradient input (GI) method can be regarded as a simplified version of first-order decomposition (Bach et al., 2015). In Equation 4, the non-linear prediction $f_l(v)$ is approximated by the linear sum of token contributions, where the dot product between the embedding v_i of the i-th token and its derivative $\frac{\partial f_l(v)}{\partial v_i}$ serves as the token's contribution.

$$C_{GI}(v_i) \approx \sum_{j=1}^{m} \frac{\partial f_l(v)}{\partial v_{ij}} v_{ij} \tag{3}$$

$$f_l(v) \approx \sum_{i=1}^{n} \frac{\partial f_l(v)}{\partial v_i} \cdot v_i \tag{4}$$

3.3.3 Integrated gradient (IG)

The integrated gradient (IG) method involves selecting an additional reference sample $\hat{u}$.We assume that the embedding and predicted probability of label l for this reference sample are denoted as $\hat{v}$ and $f_l(\hat{v})$, respectively. The IG method posits that the difference in predictions between these two samples can be attributed to differences in the input embeddings, as illustrated in Equation 5, where $C_{IG}(v_i)$ represents the contribution of token v_i to the prediction. By considering the straight-line path from the baseline embedding $\hat{v}$ to the input embedding v, and calculating gradients at each point along the path (Sundararajan et al., 2017), $C_{IG}(v_i)$ is obtained by accumulating these gradients, as shown in Equation 6. For this study, a reference sample with all-zero tokens was employed for both models.

$$\sum_{i=1}^{n} C_{IG}(v_i) = f_l(v) - f_l(\hat{v}) \tag{5}$$

$$C_{IG}(v_i) \approx \sum_{j=1}^{m} (v_{ij} - \hat{e}_{ij}) \times \int_{\beta=0}^{1} \frac{\partial f\left(\hat{e} + \beta \times (v - \hat{v})\right)}{\partial v_{ij}} d\beta \tag{6}$$

3.3.4 LIME

LIME, which stands for Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016), calculates feature contributions of the sample u by selecting neighboring samples and constructing an interpretable model to approximate the predictions of the deep learning model f. Specifically, given an input utterance u consisting of n tokens represented as $(u_1, u_2, ..., u_n)$, LIME generates a set of perturbed samples (e.g., u') in the proximity of or distant from the original sample u. This is achieved by randomly preserving some tokens in u while omitting others. For instance, considering a binary vector $s = (s_1, s_2, ..., s_n)$ where $s_i \in \{0, 1\}$, if $s_i = 1$, token u_i will be included in the perturbed sample u', while if $s_i = 0$, it will be absent. Subsequently, LIME employs the deep learning model f to predict the labels (e.g., $f_l(u')$) for these perturbed samples. Based on these neighboring perturbed samples and their corresponding predictions, LIME selects an interpretable model g that fits the data while endeavoring to closely approximate the predictions of the deep learning model. The predictions of the interpretable model g on these samples (i.e., $g_l(u')$) aim to closely match the predictions of the deep learning model f (i.e., $f_l(u')$). In Equation 7, the loss function measures the discrepancy between $f_l(u')$ and $g_l(u')$, while also considering the distance between u and u' as a weight denoted by $\pi_u(\mathbf{u})$. In our task, the weight is computed using cosine distance. For computing token relevance, a linear regression model is selected as g, as depicted in Equation 8. Additionally, the number of perturbed samples is set to be 500. For further technical details, refer to the work by Ribeiro et al. (2016).

$$loss = \sum_{u'} \pi_u(\mathbf{u}) \left(f_l(u') - g_l\left(u'\right)\right)^2 \tag{7}$$

$$g_l(u) \approx \sum_{i=1}^{n} C_{LIME}(v_i) \cdot v_i \tag{8}$$

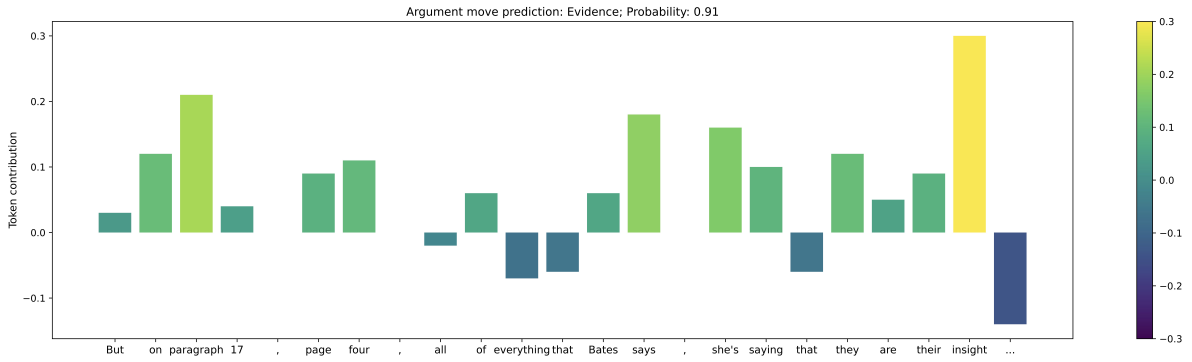


Figure 1: A visualized explanation for a prediction from the Bert model for argument move analysis using the LIME method.

3.4 Interpreting example

By employing the proposed four interpreting methods, we are able to derive the contribution of each token in a student's utterance towards the predictions made by the deep learning models developed for argument move and specificity analysis. However, the resulting explanations are presented in numerical form, which may pose challenges for comprehension, particularly for teachers and students who are the primary users of these models and explanations. To address this issue, we have designed the explanations in a visualized format. As depicted in Figure 1, we utilize bar charts to represent the token contributions. Additionally, to ensure accessibility for individuals with color-blindness or color-weakness, we employ yellow, green, and purple colors to highlight positive and negative contributions that correspond to support or objection, respectively.

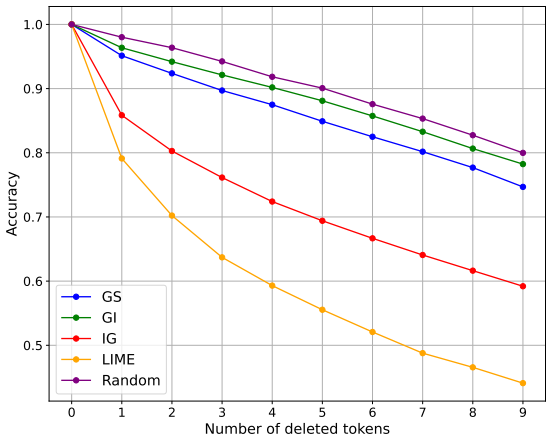
4 Computational Experiment

Prior to providing visualized explanations for users, we carried out a computational experiment to assess whether the obtained token contributions accurately represent their significance to the model prediction. In particular, we chose student utterances for which the argument move and specificity labels were correctly predicted by deep learning models. Based on the decreasing order of token contributions computed by the four interpreting methods, we removed the most critical words in a step-wise manner until nine words were eliminated. If the token contributions truly signify their importance in the prediction of deep learning models, the removal of the most importance ones would result in a substantial change in prediction accuracy. Taking into account that random deletion could also lead to a change in prediction accuracy, we conducted a random deletion experiment for comparison purposes. In our experiment, we separately selected 9,547 utterances for the Bert-based argument move model and 8,417 utterances for the Bert-based specificity model, all of which had a length greater than 10.

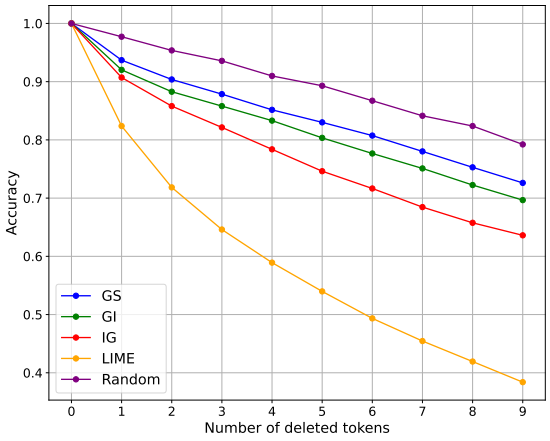
As depicted in Figure 2, the removal of words from initially accurately predicted utterances based on their contributions results in a substantial decrease in prediction accuracy compared to the elimination of words at random. For example, for the Bert-based argument move model, eliminating nine words according to contributions computed by LIME and IG causes the prediction accuracy to decline from 1.0 to 0.44 and 0.59, respectively, while random deletion only leads to a drop in prediction accuracy to 0.80. Similarly, for the Bert-based specificity model, removing nine words based on contributions calculated by LIME and IG results in a decrease in prediction accuracy from 1.0 to 0.38 and 0.63, respectively, whereas random deletion only causes the prediction accuracy to reduce to 0.79. The experimental results suggest that the four interpreting methods can explain argument move and specificity analysis by effectively identifying crucial words within argumentation, with the LIME method demonstrating the most exceptional performance in model explanation. Thus, we will use LIME to provide model explanations in the subsequent user experiment.

5 User Experiment Design

Following the successful validation of the explanations, we designed an experiment aimed at evaluating the impact of these explanations on user trust



(a) Bert-based argument move model



(b) Bert-based argument move model

Figure 2: Accuracy change when deleting words from initially correctly predicted utterances based on their contributions computed by gradient sensitivity (GS), gradient input (GI), integrated gradient (IG), and LIME methods.

in and knowledge of the deep learning models for argument move and specificity. We will implement it in future practice as a critical empirical study of the PhD thesis.

5.1 Participants

Given that the deep learning models were developed within the context of high school English lessons, our target participants for the experiment will be 60 high school English teachers who are interested in receiving AI analysis for their classroom teaching. We will randomly assign them to either an intervention group (N = 30) or a control group (N = 30), taking into account variables such as age, gender, and teaching experience. This randomization process will ensure that there are no significant differences in demographic information across the three variables mentioned. Both groups will receive automated analysis pertaining to the argument move and specificity of collaborative argumentation in their classrooms. The key distinction between the intervention group and the control group lies in the provision of explanations. Specifically, the intervention group will receive explanations accompanying the automated analysis, while the control group will not receive any explanations.

5.2 Experiment procedure

The experiment procedure, as designed in Figure 3, encompasses five distinct stages. In stage 1, teachers from both the intervention and control groups will be required to record two videos of collaborative argumentation within their classrooms. These videos will then be uploaded to the *classroom discourse analyzer* (CDA) system (Chen et al., 2015), an automated platform specifically designed to facilitate classroom dialogue analysis for teachers. Leveraging automatic speech recognition software and deep learning models developed in this study, the CDA system will transcribe and automatically analyze the argument move and specificity exhibited in the collaborative argumentation videos. Moving to stage 2, teachers will be invited to attend a workshop where they will analyze the first collaborative argumentation video using the AI-powered CDA system. Importantly, the system will provide argumentation analysis directly, without any accompanying explanations. Transitioning to stage 3, teachers will be required to complete a questionnaire aimed at assessing their trust in and knowledge of the AI-powered system, particularly concerning the AI analysis, based on their interaction with the system.

Proceeding to stage 4, teachers will be invited to analyze the second collaborative argumentation video utilizing the AI-powered CDA system. However, while the intervention group will receive argumentation analysis accompanied by explanations, the control group will continue to receive AI analysis without explanations. Finally, in stage 5, teachers from both groups will complete a questionnaire to report their trust in and knowledge of the system based on their interaction with AI during stage 4. Moreover, a subset of ten teachers from the intervention group will be randomly selected for an

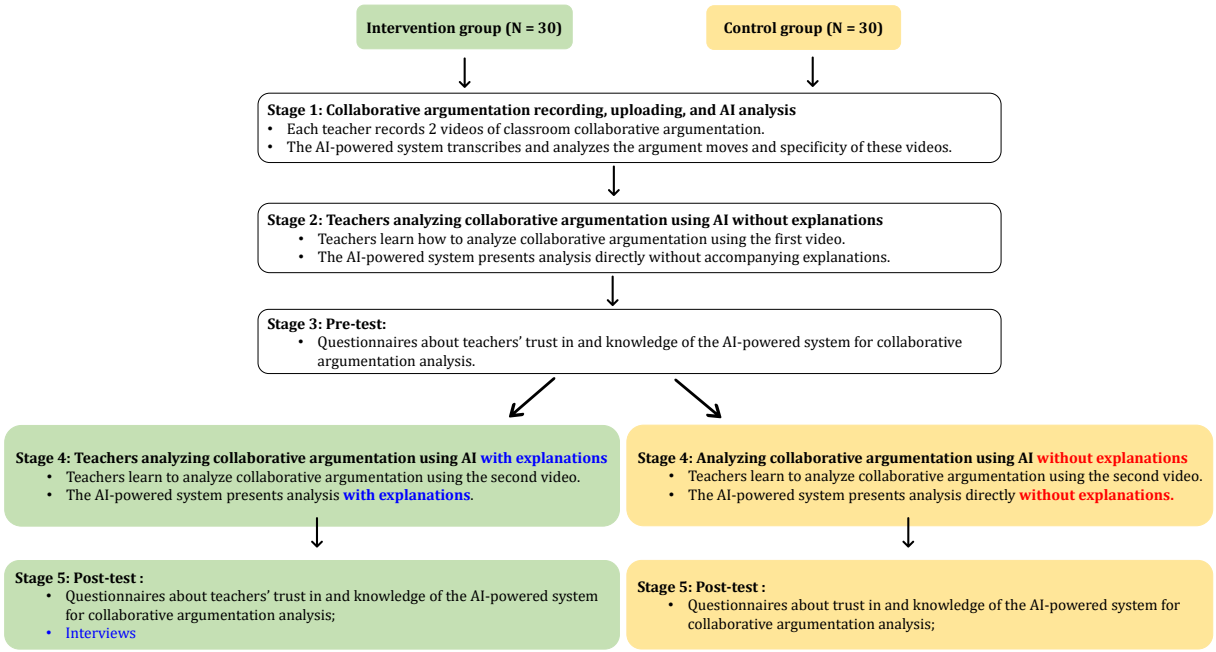


Figure 3: The procedure of the user experiment.

interview to explore their experiences and perceptions regarding the utilization of AI and explanations for collaborative argumentation analysis.

5.3 Instruments

To assess the level of trust among teachers in the AI-powered system, specifically regarding the deep learning model for collaborative argumentation analysis, we will adapt a trust scale initially developed by Jian et al. (2000). Originally designed to evaluate user trust in automated systems, this scale has been widely utilized to measure human trust in AI-powered tools. It encompasses factors such as perceived fidelity, loyalty, reliability, security, integrity, and familiarity with the AI tools. The questionnaire consists of 11 items and employs a 7-point Likert scale to capture participants' responses accurately.

Regarding the questionnaire for knowledge assessment, it aims to evaluate teachers' understanding of the basic functionalities of the AI-powered system and their comprehension of the deep learning model for collaborative argumentation analysis, including how the model makes predictions. This evaluation is crucial in demonstrating the effectiveness of the developed AI model and its accompanying explanations. The design of the knowledge questionnaire will be undertaken by two researchers who are responsible for the development and integration of the AI-powered collaborative argumentation model into the CDA system.

6 Conclusion

Recognizing the significance of collaborative argumentation in teaching and learning, this study employs Bert (i.e., a widely adopted deep learning approach) and authentic discussion transcripts to develop two models for automated analysis of argument moves (i.e., claim, evidence, and warrant) and specificity levels (i.e., low, medium, and high) within collaborative argumentation. Given that the "black box" nature of deep learning models may raise trust concerns among users, four explainable AI methods are proposed to unpack model analysis and provide explanations. These methods include gradient sensitivity, gradient input, integrated gradient, and LIME. The computational experiments demonstrate the effectiveness of these methods in explaining model predictions by computing word contributions, with LIME exhibiting the most exceptional performance. Consequently, this study aims to apply the developed model and the LIME method for collaborative argumentation analysis and explanation. A quasi-experiment is designed to evaluate the influence of model explanations on user trust and knowledge, representing a future extension of this PhD proposal. By addressing the challenges of interpretability and trust, this PhD thesis proposal contributes to the field of AI-supported classroom teaching, potentially fostering user trust in AI and facilitating the practical implementation of AI in educational contexts.

This proposal also has several limitations that should be addressed before the formal implementation of the quasi-experiment. First, the study utilizes only one dataset, leaving uncertainty about the applicability of the explainable AI methods to models on other datasets of classroom collaborative argumentation. Second, although the explanations for collaborative argumentation analysis are designed in a visual format, it is unclear whether this is the preferred format for teachers and how it might impact their perception of the explanations. Therefore, further research should focus on evaluating the proposed method across multiple datasets and conducting a preliminary experiment to identify the optimal visualization of explanations. This will help avoid confounding the effects of explanations on users' overall trust.

Acknowledgments

This work was supported by Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221).

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Christa SC Asterhan, Christine Howe, Adam Lefstein, Eugene Matusov, and Alina Reznitskaya. 2020. Controversies and consensus in research on dialogic teaching and learning. *Dialogic Pedagogy*, 8.

Christa SC Asterhan and Baruch B Schwarz. 2016. Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist*, 51(2):164–187.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*.

Nathaniel Blanchard, Michael Brady, Andrew M Olney, Marci Glaus, Xiaoyi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney D'Mello. 2015. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17*, pages 23–33. Springer.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Gaowei Chen, Sherice N Clarke, and Lauren B Resnick. 2015. Classroom discourse analyzer (cda): A discourse analytic tool for teachers. *Technology, Instruction, Cognition and Learning*, 10(2):85–105.

Yu-Cheng Chien, Ming-Chi Liu, and Ting-Ting Wu. 2020. Discussion-record-based prediction model for creativity education using clustering methods. *Thinking skills and creativity*, 36:100650.

James S Chisholm and Amanda J Godley. 2011. Learning about language through inquiry-based discussion: Three bidialectal high school students' talk about dialect variation, identity, and power. *Journal of Literacy Research*, 43(4):430–468.

Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial intelligence*, 298:103503.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Katerina Evers and Sufen Chen. 2022. Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, 35(8):1869–1889.

Lei Gao, Xiaoran Li, Yanyan Li, and Wanqing Hu. 2023. Capturing temporal and sequential patterns of socio-emotional interaction in high-and low-performing collaborative argumentation groups. *The Asia-Pacific Education Researcher*, 32(6):817–831.

AKM Bahalul Haque, AKM Najmul Islam, and Patrick Mikalef. 2023. Explainable artificial intelligence (xai) from a user perspective: A synthesis of prior literature and problematizing avenues for future

research. *Technological Forecasting and Social Change*, 186:122120.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. *arXiv preprint arXiv:2404.15310*.

Changqin Huang, Zhongmei Han, Ming Li, Xizhe Wang, and Wenzhu Zhao. 2021. Sentiment evolution with interaction levels in blended learning environments: Using learning analytics and epistemic network analysis. *Australasian Journal of Educational Technology*, 37(2):81–95.

Nicholas Hunkins, Sean Kelly, and Sidney D'Mello. 2022. "beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. In *Lak22: 12th international learning analytics and knowledge conference*, pages 230–238.

Stephen Jackson and Niki Panteli. 2023. Trust or mistrust in algorithmic grading? an embedded agency perspective. *International Journal of Information Management*, 69:102555.

Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280.

Magdalene Lampert, Heather Beasley, Hala Ghousseini, Elham Kazemi, and Megan Franke. 2010. Using designed instructional activities to enable novices to manage ambitious mathematics teaching. *Instructional explanations in the disciplines*, pages 129–141.

Carol D Lee. 2006. 'every good-bye ain't gone': analyzing the cultural underpinnings of classroom talk. *International Journal of Qualitative Studies in Education*, 19(3):305–327.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in nlp. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691.

Xuan Liu, Xiaoguang Wang, and Stan Matwin. 2018. Improving the interpretability of deep neural networks with knowledge distillation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 905–912. IEEE.

Yu Lu, Deliang Wang, Penghe Chen, Qinggang Meng, and Shengquan Yu. 2023. Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*, 33(3):519–542.

Yu Lu, Deliang Wang, Penghe Chen, and Zhi Zhang. 2024. Design and evaluation of trustworthy knowledge tracing model for intelligent tutoring system. *IEEE Transctions on Learning Technologies*.

Luca Lugini. 2021. *Analysis of collaborative argumentation in text-based classroom discussions*. Ph.D. thesis, University of Pittsburgh.

Luca Lugini, Diane Litman, Amanda Godley, and Christopher Olshefski. 2019. Annotating student talk in text-based classroom discussions. *arXiv preprint arXiv:1909.03023*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Bruce M McLaren, Oliver Scheuer, and Jan Mikšátko. 2010. Supporting collaborative learning and e-discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education*, 20(1):1–46.

Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, 39(1):53–63.

Tanya Nazaretsky, Mutlu Cukurova, and Giora Alexandron. 2022. An instrument for measuring teachers' trust in ai-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference*, pages 56–66.

Tanya Nazaretsky, Mutlu Cukurova, Moriah Ariely, and Giora Alexandron. 2021. Confirmation bias and trust: human factors that influence teachers' attitudes towards ai-based educational technology. In *CEUR Workshop Proceedings*, volume 3042.

Tanya Nazaretsky, Jamie N Mikeska, and Beata Beigman Klebanov. 2023. Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 122–132.

Andrew M Olney, Patrick J Donnelly, Borhan Samei, and Sidney K D'Mello. 2017. Assessing the dialogic properties of classroom discourse: Proportion models for imbalanced classes. *International Educational Data Mining Society*.

Christopher Olshefski, Luca Lugini, Ravneet Singh, Diane Litman, and Amanda Godley. 2020. The discussion tracker corpus of collaborative argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1033–1043.

Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining recommendations in e-learning: Effects on adolescents' trust. In *27th International Conference on Intelligent User Interfaces*, pages 93–105.

Joe Oyler. 2019. Exploring teacher contributions to student argumentation quality. *Studia paedagogica*, 24(4):173–198.

Samuel L Pugh, Arjun Rao, Angela EB Stewart, and Sidney K D'Mello. 2022. Do speech-based collaboration analytics generalize across task contexts? In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 208–218.

Fen Qin, Kai Li, and Jianyuan Yan. 2020. Understanding user trust in artificial intelligence-based educational systems: Evidence from china. *British Journal of Educational Technology*, 51(5):1693–1710.

Chrysi Rapanta and Mark K Felton. 2022. Learning to argue through dialogue: A review of instructional approaches. *Educational Psychology Review*, pages 1–33.

Joseph M Reilly and Bertrand Schneider. 2019. Predicting the quality of collaborative problem solving through linguistic analysis of discourse. *International Educational Data Mining Society*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Dapeng Shan, Deliang Wang, Chenwei Zhang, Ben Kao, and Carol KK Chan. 2023. Annotating educational dialog act with data augmentation in online one-on-one tutoring. In *International Conference on Artificial Intelligence in Education*, pages 472–477. Springer.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9721–9728.

Deliang Wang, Cunling Bian, and Gaowei Chen. 2024a. Using explainable ai to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load. *British Journal of Educational Technology*.

Deliang Wang and Gaowei Chen. 2024. Are perfect transcripts necessary when we analyze classroom dialogue using aiot? *Internet of Things*, 25:101105.

Deliang Wang, Dapeng Shan, Yaqian Zheng, and Gaowei Chen. 2023a. Teacher talk moves in k12 mathematics lessons: Automatic identification, prediction explanation, and characteristic exploration. In *International Conference on Artificial Intelligence in Education*, pages 651–664. Springer.

Deliang Wang, Dapeng Shan, Yaqian Zheng, Kai Guo, Gaowei Chen, and Yu Lu. 2023b. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519. International Educational Data Mining Society.

Deliang Wang, Yang Tao, and Gaowei Chen. 2024b. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, 123:102275.

Rose E Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*.

Jacob Whitehill and Jennifer LoCasale-Crouch. 2023. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *arXiv preprint arXiv:2310.01132*.

Shiting Xu, Wenbiao Ding, and Zitao Liu. 2020. Automatic dialogic instruction detection for k-12 online one-on-one classes. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, pages 340–345. Springer.

Yuanyi Zhen, Lanqin Zheng, and Penghe Chen. 2021. Constructing knowledge graphs for online collaborative programming. *IEEE Access*, 9:117969–117980.