MONASH University

Analytics of Communicative Patterns in Educational Feedback

Jionghao Lin

Doctor of Philosophy

A Thesis Submitted for the Degree of Doctor of Philosophy at
Monash University in 2023
School of Information Technology

Copyright notice

Abstract

Educational feedback has been acknowledged as an effective method for enhancing learners' academic achievements. Feedback was initially conceptualised as providing information to learners. The recent literature has gradually shifted towards dialogic feedback, which emphasises the learning process. Dialogic feedback allows learners to understand feedback information related to their performance and use the information to improve their future work. Theoretically, dialogic feedback can be characterised from three dimensions: *i) cognitive dimension*, *ii) social-affective dimension*, and *iii) structural dimension*. Each dimension involves the transmission of information, which can be categorised into communicative patterns, i.e., referring to how interlocutors (e.g., educators and learners) interact with each other through feedback communication. To understand the construction of effective feedback, it is important to analyse effective communicative patterns from feedback practice based on these three dimensions.

The *structural dimension* of dialogic feedback can reflect how dialogic feedback can be delivered to learners. We argue that one-on-one tutoring dialogue is particularly pertinent to the design of dialogic feedback since the tutoring dialogue allows learners to make timely adjustments and exchange ideas with educators. Given its high relevance to dialogic feedback, the current PhD thesis mainly focuses on the investigation of one-on-one tutoring dialogue. Since one-on-one tutoring dialogue is impractical for large cohorts, assessment written feedback based on dialogic feedback design can be an alternative. To deliver effective written feedback, educators need to make efforts in the construction of feedback content and the design of the assignments (e.g., allowing learners to resubmit their improved work based on the given feedback). Although both types of feedback can enhance learners' academic performance, the effective construction of dialogic feedback in both is still limited. Thus, further efforts are required to investigate the feedback *social-affective dimension* and *cognitive dimension*.

To enhance the effectiveness of the feedback, it is necessary to explore how educators strengthen interpersonal relationships with learners during the feedback process, which involves the *social-affective dimension*. By scrutinising the literature, we identified a significant factor, i.e., linguistic politeness, which is about the use of language to show respect and care for interlocutors' feelings in communication. Educators' polite expressions consider learners' personal feelings and can motivate them to engage with feedback. However, excessive use of politeness in feedback may lead to unclear instruction and increase the learners' cognitive loads, thereby affecting the *cognitive dimension* of feedback. Therefore, this PhD thesis aimed to investigate how educators can effectively express politeness in feedback to enhance the *social-affective dimension* of feedback while ensuring that the *cognitive dimension* is not compromised.

The presence of the *cognitive dimension* in feedback can be manifested by the feedback content that educators encourage learners to reflect on their work and guide learners towards the desired learning achievement. However, the understanding of how the *cognitive dimension* of feedback

can be effectively delivered is still limited. In the case of one-on-one tutoring dialogue, previous studies have used educational dialogue acts to examine the conversational actions behind utterances made by educators and learners during feedback communication. The investigation of instructional strategies is still under-explored, especially in large-scale feedback sessions. In the written feedback, most prior works conducted feedback content analysis to investigate the effective design of the *cognitive dimension* of feedback. The insights on the design of dialogic feedback in assessment written assessments may be limited from most previous analysis results since the analysed feedback was given in the one-way form without allowing learners to submit their improved work based on the feedback they received.

This PhD thesis started with exploring politeness in one-on-one tutoring dialogue by using the LA approaches. We analysed an anonymised tutoring dialogue dataset, which contained 14,562 tutoring dialogue sessions. We employed linguistic theories of politeness to investigate the communicative patterns of politeness from the educators' and learners' utterances. Our analytical results demonstrated that there was no evident correlation between the educators' politeness and learner performance in non-instructional communication (e.g., greeting), but a correlation existed in instructional communication (e.g., asking questions). To gain a better understanding of the use of politeness in instructional communication, inspired by the previous studies, we adopted a widely-used tutoring dialogue act scheme to identify the dialogue acts in the tutoring dialogue communication and developed a dialogue act classifier to analyse the dialogue acts for all dialogue sessions. We further improved the classifier performance by incorporating dialogue contextual information and statistical active learning methods during the classifier training process. Through extensive analysis, we identified a set of effective communicative patterns involving the use of instructional strategies and politeness in instructional communication in the feedback provision process. In addition to the analytics of dialogue, we further investigated the communicative patterns (e.g., politeness and instructional information) from the assessment written feedback. Overall, the PhD thesis demonstrated recommendations for the practices of tutoring dialogue (e.g., the recommendations for effectively expressing politeness and delivering instructional strategies in tutoring dialogue), the development of dialogue-based ITS (e.g., the enhancement for identifying dialogue acts from tutoring dialogues), and the novel method of automatically analysing the quality of assessment feedback content based on the properties proposed in the well-known framework for learner-centred feedback.

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature:

Print Name: Jionghao Lin

Date: 25/1/2023

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 4 original papers that were published in peer-reviewed publications and 2 papers that are under review. The core theme of the thesis is 'communicative patterns in educational feedback'. The ideas, implementations, and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the Faculty of Information Technology at Monash University under the supervision of Prof. Dragan Gašević, Dr. Guanliang Chen and Prof. Sharon Oviatt.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of Chapters 2, 3, 4, 5, and 6 my contribution to the work involved the following:

TABLE 1: Publications and the nature of contributions

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names(s) Nature and % of Co-author's contribution**	**Co-author(s) Monash student Y/N**
2	Investigating the Role of Politeness in Human-Human Online Tutoring	Published	Conceptualisation, experimentation, results analysis, manuscript writing 70%	1. David Lang, data curation, manuscript editing 3% 2. Haoran Xie conceptualisation 2% 3. Dragan Gašević supervision, manuscript writing 10% 4. Guanliang Chen conceptualisation, manuscript writing 15%	1. N 2. N 3. N 4. N

Thesis Chapter	**Publication Title**	**Status**	**Nature and % of student contribution**	**Co-author names(s) Nature and % of Co-author's contribution**	**Co-author(s) Monash student Y/N**
2	On the Role of Politeness in Online Human-Human Tutoring	Submitted	Conceptualisation, experimentation, results analysis, manuscript writing 70%	1. Mladen Raković, manuscript editing 5% 2. Haoran Xie manuscript editing 3% 3. David Lang data curation, manuscript editing 2% 4. Dragan Gašević supervision, manuscript writing 10% 5. Guanliang Chen conceptualisation, manuscript writing 10%	1. N 2. N 3. N 4. N 5. N
3	Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues.	Published	Conceptualisation, experimentation, results analysis, manuscript writing 70%	1. Shaveen Singh, manuscript writing 5% 2. Lele Sha data annotation 5% 3. Wei Tan experimentation 3% 4. David Lang data collection 2% 5. Dragan Gašević supervision, manuscript writing 5% 6. Guanliang Chen conceptualisation, manuscript writing 10%	1. N 2. Y 3. Y 4. N 5. N 6. N
4	Enhancing Educational Dialogue Act Classification with Discourse Context and Sample Informativeness.	Submitted	Conceptualisation, experimentation, results analysis, manuscript writing 55%	1. Wei Tan, experimentation 25% 2. Du Lan conceptualisation 5% 3. Wray Buntine manuscript editing 3% 4. David Lang data curation manuscript editing 2% 5. Dragan Gašević supervision, manuscript writing 5% 6. Guanliang Chen conceptualisation, manuscript writing 5%	1. Y 2. N 3. N 4. N 5. N 6. N

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author names(s) Nature and % of Co-author's contribution	Co-author(s) Monash student Y/N
5	Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues	Published	Conceptualisation, experimentation, results analysis, manuscript writing 75%	1. Mladen Raković, manuscript writing 5% 2. David Lang data curation manuscript editing 3% 3. Dragan Gašević supervision, manuscript writing 7% 4. Guanliang Chen conceptualisation, manuscript writing 10%	1. N 2. N 3. N 4. N
6	Learner-centred Analytics of Feedback Content in Higher Education	Published	Conceptualisation, experimentation, results analysis, manuscript writing 75%	1. Wei Dai annotation, manuscript editing 5% 2. Lisa-Angelique Lim annotation, manuscript editing 3% 3. Yi-Shan Tsai conceptualisation, manuscript editin 2% 4. Rafael Ferreira Mello manuscript editing 2% 5. Hassan Khosravi manuscript editing 2% 6. Dragan Gašević supervision, manuscript writing 6% 7. Guanliang Chen conceptualisation, manuscript writing 5%	1. Y 2. N 3. N 4. N 5. N 6. N 7. N

I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: Jionghao Lin

Student signature:

Date:25/1/2023

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Dragan Gašević

Main Supervisor signature:

Date:25/1/2023

Publications during enrolment

- Lin, J., Lang, D., Xie, H., Gašević, D., & Chen, G. (2020). Investigating the Role of Politeness in Human-Human Online Tutoring. In *International Conference on Artificial Intelligence in Education* (pp. 174-179). Springer, Cham.
- Lin, J., Raković, M., Lang, D., Gašević, D., & Chen, G. (2022). Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 282-293).
- Lin, J., Singh, S., Sha, L., Tan, W., Lang, D., Gašević, D., & Chen, G. (2022). Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems*, 127, 194-207.
- Lin, J., Dai, W., Lim, L., Tsai, Y., Mello, R., Khosravi, H., Gašević, D., & Chen, G. (2023). Learner-centred Analytics of Feedback Content in Higher Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. (pp. 100–110).
- Lin, J., Raković, M., Xie, H., Lang, D., Gašević, D., & Chen, G. (2023). On the Role of Politeness in Online Human-Human Tutoring. *British Journal of Educational Technology* (Under Review, Minor Revision).
- Lin, J., Tan, W., Du, L., Buntine, W., Lang, D., Gašević, D., & Chen, G. (2023). Enhancing Educational Dialogue Act Classification with Discourse Context and Sample Informativeness. *IEEE Transactions on Learning Technologies* (Under Review).

All the papers are published/submitted to the conferences and journals that are on the quality list of the Faculty of Information Technology at Monash University.

Acknowledgements

I would like to begin by expressing my deepest gratitude to Prof. *Dragan Gašević* for his unwavering support, guidance, and encouragement throughout my PhD study journey. Prof. *Dragan Gašević* has been a beacon of light in times of darkness, offering positivity, joy, and hope when it was needed most. I am truly grateful to have had him as my supervisor.

I would also like to extend my heartfelt appreciation to Dr. *Guanliang Chen*, who has been a supervisor and friend, offering me guidance, support, and advice based on his experience and knowledge. He has always been willing to offer both professional and personal support, helping me to reach my goals and achieve my full potential.

My gratitude also goes to Prof. *Sharon Oviatt* for providing me with the opportunity to conduct research and recognising my research abilities. She has been a valuable supervisor who took the time to understand my strengths, weaknesses, and areas for growth and helped me to develop and grow in order to reach my full potential.

I would like to express my gratitude to all members of the *Center for Learning Analytics at Monash (CoLAM)* for their support and encouragement throughout my PhD study journey. Their valuable insights and guidance have been instrumental in shaping my research and academic growth. I would also like to extend my appreciation to *my friends* who have been with me through thick and thin, especially during my time at Monash. Your constant companionship and unwavering support have been a source of strength and motivation for me, and I am deeply grateful for your friendship.

I am also grateful to the *Faculty of Information Technology at Monash University* for offering me a life-changing learning experience. The knowledge and skills I acquired through my learning journey at Monash University have significantly influenced my understanding of the world and myself. The opportunity to learn and engage with diverse perspectives at Monash has broadened my view and inspired me to take action in meaningful ways. Thank you to all the faculty members and staff who contributed to making this experience possible.

Finally, I want to express my deepest gratitude to my beloved family, my mother *Shufen Jiang*, father *Wei Lin*, and grandmother *Shihua Xiong*. Throughout the most challenging years of my PhD study journey, they provided unwavering support and encouragement. Their love, patience, and understanding have been invaluable to me, and they always believed in me even when I doubted myself. I cannot thank them enough for everything they have done for me, and a simple expression of gratitude would not be enough to convey how much I appreciate them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A journey of a thousand miles begins with a single step.

– Lao Tzu

1.1 Background

1.1.1 Communicative Patterns in Educational Feedback

Educational feedback has been widely acknowledged as one of the most important instructional methods in enhancing a learner's academic achievement [2–8]. In teaching practices, human educators commonly use feedback to help learners minimise the learning gap between the learner's actual learning performance and the expected performance [9, 10]. Driven by the effectiveness of feedback on learning achievements, educational researchers have had a longstanding debate on the definition of feedback. In the existing feedback literature, the conceptualisation of feedback has gradually shifted from *feedback as information* to *feedback as a process* [11]. The earlier definition (i.e., *feedback as information*) of effective feedback focuses on the information and timeliness [10], whereas the recent understanding (i.e., *feedback as a process*) of feedback focuses more on learning process [11] where learners could make sense of the information pertaining to their learning performance, and use feedback information to further improve the quality of their future work [6, 11]. The reason behind this shift might be that providing feedback as a process is more beneficial to learners than as information [8, 11]. Many previous studies found that *feedback as a process* could lead to more effective feedback design to promote learners' achievements compared to the *feedback as information* [9, 12, 13].

Recent studies on *feedback as a process* [1, 7, 8, 14, 15] have started placing more emphasis on considering the feedback as a dialogue communication process, i.e., dialogic feedback.

According to Carless [14], dialogic feedback is defined as "*interactive exchanges in which interpretations are shared, meanings negotiated and expectations clarified.*" In other words, dialogic feedback encourages an interactive communication process that involves the transmission of information between educators and learners such as the exchange of performance-relevant information and the evolution of the relationship [1, 14–16]. In the feedback communication process, the transmission of information can be categorised into different types of ***communicative patterns*** which refer to how interlocutors (e.g., educators and learners) interact with each other through communication in a given context (e.g., educational domain) [17]. For example, educators can use the communicative pattern which presents supportive and motivational information (e.g., "*Well done! When you stuck with it and didn't give up, you succeeded!*") to help learners engage with their learning task [8]. When feedback is effectively communicated, both educators and learners can understand each other's viewpoints and work together to develop actionable solutions, which can be powerful to support learners in their learning process [1, 7, 14, 15]. Therefore, it is important for educational researchers to investigate the ***communicative patterns*** that constitute dialogic feedback.

To understand the construction of dialogic feedback, a well-known feedback study by Yang and Carless [1] proposed a dialogic feedback framework *Feedback Triangle* shown in Figure 1.1. The framework characterises dialogic feedback from three dimensions: the organisation and management of feedback (*structural dimension*), the social and interpersonal negotiation of feedback (*social-affective dimension*), and the content of feedback (*cognitive dimension*). The *Feedback Triangle* framework [1] has been widely acknowledged in the existing feedback research [7, 8, 18–20]. Many prior studies (e.g., [18, 20, 21]) employed the framework [1] to analyse feedback from the three feedback dimensions (i.e., *structural dimension*, *social-affective dimension*, and *cognitive dimension*). Driven by the widespread use of the *Feedback Triangle* framework [1], we decide to delve into the details of the three dimensions of dialogic feedback.

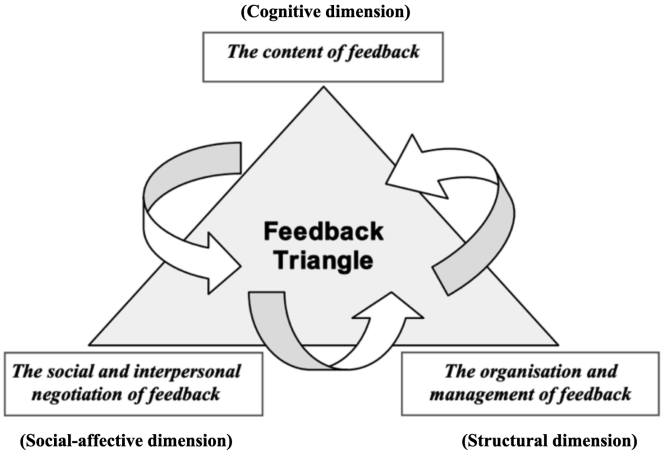


FIGURE 1.1: Feedback Triangle (adopted from Yang and Carless [1])

The organisation and management of feedback (i.e., *structural dimension*) are about how feedback is delivered [1]. Specifically, a recent work [22] proposed a conceptual framework on dialogic feedback, which identified several components related to the *structural dimension*, including the timing of feedback (e.g., feedback is provided before the submission), assessment design (e.g., allow learners resubmit their works), and personalised information (e.g., feedback content is tailored to each learner's needs) [22]. These components informed how the dialogic feedback could be crafted in practice [22]. For example, we argue that the one-on-one tutoring dialogue might indicate the high potential of being dialogic feedback since the feedback in the dialogue form can allow learners to make adjustments timely before the submission, gain more opportunities to exchange ideas with educators and seek clarification on the problem [18, 22–26]. While tutoring dialogues hold a high potential for being dialogic feedback, it is impractical to provide feedback through dialogue between educators and a large group of learners [8]. An alternative way for feedback provision is assessment written feedback which is a widely-used form of feedback [27]. The assessment written feedback is often released as a deliverable after learners have submitted their assessments which is less timely compared with the dialogue based on the framework by [27]. To facilitate the dialogic process for written feedback, educators typically need to invest efforts in crafting feedback content (e.g., providing explanations and suggestions on students' works) and the design of the assignments (e.g., allowing students to resubmit their improved work based on the given feedback) [22]. Both tutoring dialogues and assessment written feedback have the potential of being dialogic feedback. To support learners' work on learning tasks, the feedback should convey effective content (e.g., the comments on learning tasks and actionable suggestions on the further step) [10]. According to the *Feedback Triangle* framework [1], the feedback content is positioned within the *cognitive dimension* of feedback, which is designed to help learners tackle learning problems effectively and promote learners' independent learning ability.

The *cognitive dimension* of feedback involves providing instructional content (e.g., criteria for learning goals, strategies for guiding learners' thinking, and suggestions for improving learners' working progress) [1, 18]. To provide effective feedback, the content of feedback should guide learners in taking actions to tackle the assessment and encourage learners to reflect on their own thinking processes and identify areas for improvement [1]. However, the effectiveness of the provided feedback might not always align with the educator's expectations. For example, learners, as the recipients of feedback, might not always engage with the guidance of feedback content, which might be related to the social and interpersonal relationship [1]. In the feedback provision process, educators are always in a more powerful position where the educators have the authority to grade and evaluate the learner's work and determine their performance [1, 28]. Due to the imbalanced relationship between educators and learners, educators might not realise that the words or comments in their feedback are not perceived as supportive or motivational by the learners [1, 29], which can negatively impact learners' engagement and learning outcomes

[7, 30]. To address this issue, it is important for educators to consider the social and interpersonal aspects of their feedback and ensure that the feedback is communicated in a supportive and constructive manner.

The social and interpersonal negotiation of feedback (*social-affective dimension*) is intended to influence an individual's self-efficacy, self-esteem, and motivation [1]. To be effective, feedback in this dimension should be presented in a respectful and supportive manner, which can help strengthen the social relationship between the educators and learners in the feedback process [1, 7, 8]. Particularly for low-performing learners, the social-affective dimension of feedback is important because these learners may feel vulnerable to unsupported comments (e.g., criticism) [31]. Educators are suggested to enhance the social relationship with learners, which can motivate learners to act upon the feedback [7, 31]. Thus, to help learners benefit from the feedback on their learning tasks, it is important to consider both the *cognitive* and *social-affective dimensions* of feedback when crafting the feedback.

1.1.2 Learning Analytics in Educational Feedback

The presence of feedback information from the *cognitive* and *social-affective dimensions* may be closely connected to the communicative patterns of feedback. As discussed, communicative patterns reflect how educators communicate with learners regarding the exchange of content information and the evolution of the relationship between educators and learners [17]. By investigating the effective communicative patterns from feedback, researchers can gain insights into how feedback can be communicated effectively and used to strengthen the educator-learner relationship. Driven by this, many recent studies [32–39] have employed learning analytic approaches to investigate the communicative patterns in feedback to enhance feedback effectiveness. As defined in 2011 for the first *International Conference on Learning Analytics and Knowledge* (LAK) [40], learning analytics is "*the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs*". By applying learning analytics approaches to feedback research, Cavalcanti et al. [34] investigated communicative patterns in feedback (e.g, the emotional components of feedback by using LIWC [41] and feedback readability using Coh-Metrix [42]) to identify effective feedback practices proposed by [43] in assessment written feedback. Derham et al. [4] investigated the correlation between communicative patterns (e.g., cognitive, emotional, and relational components of feedback by using LIWC [41]) and learner academic performance.

Though demonstrated certain insights on the feedback practice, prior studies [32, 33, 35, 44–46] suggested that more efforts should be invested to identify the effective communicative patterns in feedback so as to enhance the feedback quality provided by educators. Additionally, as the

learner-educator ratio keeps increasing yearly, the provision of quality feedback for large learner cohorts becomes a challenge [32, 33, 39]. The reason is that training expert educators to provide feedback is non-trivial, which is always time-consuming and cost-demanding [47, 48]. An alternative is to develop automated feedback systems for feedback provision such as Intelligent Tutoring Systems (ITSs) (e.g., AutoTutor [49] and Beetle II [50]) and learning analytics feedback systems (e.g., OnTask [51] and SRES [52]). Although achieved certain successes, the current feedback systems still require further enhancements in order to provide feedback that is as effective as educator-crafted feedback [53]. To support the practice of delivering dialogic feedback and the design of developing automated feedback systems, the current PhD thesis identified three research gaps detailed as below.

1.2 Research Gaps

Gap 1: Lack of analysing communicative patterns in one-on-one online tutoring dialogues.

Many recent feedback studies [18, 22–26] have placed more emphasis on considering dialogic feedback as a dialogue. We argue that the one-on-one online tutoring dialogue is particularly pertinent to the design of dialogic feedback since the tutoring dialogues could present adequate interaction between educators and learners, which can allow learners to make adjustments timely, gain more opportunities to exchange ideas with educators and seek clarification on the problem [18, 22–26]. However, the provision of tutoring dialogue is challenged due to the high demand for expert educators [8]. To provide tutoring dialogues for large learner cohorts, researchers have developed dialogue-based Intelligent Tutoring Systems (ITSs), learning environments where computer programs replaced human educators [54–57]. Many dialogue-based ITSs developed to date have demonstrated empirical benefits in supporting thousands of learners to study a variety of subjects, e.g., AutoTutor assisted learners in learning the fundamentals of operating systems and many other learning subjects [49, 55–57], Rimac proactively supports learners to learn physics [58], and Beetle II boosted learner understanding of concepts in electronics and electricity [50]. Despite many successful implementations of dialogue-based ITSs for providing tutoring dialogues, the communicative patterns of existing ITSs still need to be improved to become comparably effective as human educators as reported in a recent systematic literature review [59]. One of the major improvements proposed in prior research is to communicate politely with a learner [56, 60]. Many prior studies documented the benefit of polite feedback to learners' learning performance [60–64]. However, when educators excessively used politeness in the tutoring dialogues, it might have negative effects on the learning process [65–67]. For instance, some human educators might spend too much time working on polite expressions which in turn can hinder the feedback communication process [65–67]. Moreover, the

previous research demonstrated that learners with high prior knowledge preferred to receive instructional information that was directly expressed [68]. Therefore, both benefits and hindrances of guiding learners politely coexist in the feedback communication process, which motivated us to investigate how human educators express politeness in the tutoring dialogue communication process.

Driven by the above discussion on communicative patterns in the tutoring dialogue, we proposed three potential directions for exploring effective communicative patterns. Firstly, though the effectiveness of politeness in supporting learners' achievement has been discussed in many empirical studies noted above, further efforts should be devoted to investigating how educators express politeness in the tutoring dialogue communication process. Secondly, as noted in prior studies [65–67], educators might need to be careful to rephrase the instructional information (e.g., corrective feedback, thought-provoking questions, and suggestions) into polite expressions because the instructional information of feedback content should be direct, explicit, and clear [3]. To better understand how the instructional information is presented in the tutoring dialogue and supports learners' academic performance, prior works [44, 69] employed dialogue acts to reflect the instructional information behind the conversational utterances in the tutoring dialogue. However, the understanding of effective instructional information on supporting learners' academic performance is still insufficient. For example, most existing studies [44, 69, 70] focused on the analysis of successful tutoring dialogues (i.e., those in which learners successfully solved problems) and the identification of effective instructional information. We argue that, to provide learners with the necessary help, educators should also learn from unsuccessful tutoring dialogues where learners failed to solve the problems and understand how to prevent the factors that led to these failures. Therefore, it is worthwhile to further investigate the instructional information in the form of dialogue acts in tutoring dialogue. Thirdly, tutoring dialogues always involve the interplay between the use of politeness and instructional information [66, 67]. As both benefits and hindrances of guiding learners politely coexist in instructional communication, to maintain the effectiveness of one-on-one online tutoring dialogues, it is worth investigating the extent to which expert educators express politeness in instructional communication, which still remains unknown.

Considering the importance of guiding human educators in providing effective tutoring dialogues and contributing to the design of dialogue-based ITS, we argue that it is worthwhile to investigate the communicative patterns (i.e., politeness and instructional information) from real-world educational dialogues.

Gap 2: Lack of effort in facilitating the analysis of tutoring dialogues.

As noted in Gap 1, understanding how educators communicate instructional information with learners is important. A typical approach to understanding communication between educators

and learners in tutoring dialogues is through the use of dialogue acts that can reflect what educators and learners do during the tutoring dialogue communication process [44, 70]. Early works of tutoring dialogue analysis [69, 71–74] investigated the dialogue acts based on the limited number of dialogue sessions (e.g., 48 programming dialogue sessions [75] and 222 algebra dialogue sessions [76]), which might not be sufficient to fully understand the use of dialogue acts in the tutoring dialogue communication process. Driven by this, a recent research trend is to analyse the dialogue acts in large tutoring dialogue corpora. Prior studies [44, 70, 77, 78] employed the traditional supervised machine learning models (e.g., Support Vector Machine, Naive Bayes, and Decision Tree) to train on the annotated dialogue sessions and used the trained dialogue act classifiers to automatically identify the dialogue acts from the entire dialogue corpus. The dialogue acts identified automatically by the classifiers can be applied to analyse the use of dialogue acts in the entire dialogue corpus. Therefore, automating the recognition of dialogue acts in a large-scale data corpus can facilitate tutoring dialogue analysis. According to the *Design Recommendations for Intelligent Tutoring Systems* [56], the identification of dialogue acts from the tutoring dialogue is a significant part of designing the dialogue-based ITSs. However, accurately identifying dialogue acts remains an ongoing challenge that needs improvement [56]. Given the critical role of dialogue acts in analysing tutoring dialogues, this PhD study aimed to enhance the performance of classifying dialogue acts and further facilitate tutoring dialogue analysis.

Gap 3: Lack of analysis of communicative patterns in assessment feedback.

In addition to the analytics of communicative patterns in tutoring dialogues, we argue that it is necessary to investigate effective communicative patterns from the assessment written feedback due to its widespread use for supporting learners, especially in higher education [27]. Informed by [3, 8], effective written feedback should provide high-quality content and demonstrate the effect on the learners' improvements (e.g, improvement in learning outcomes). To understand how feedback content affects learners' improvement, researchers have employed learning analytics approaches to investigate the relationship between communicative patterns in feedback and the learners' learning outcomes. For example, a recent study by [32] used learning analytics approaches to analyse the correlation between the communicative patterns (e.g., N-grams, sentence types) and learners' academic performance on the task. However, we argued that their analysis results might provide limited insights into designing dialogic feedback in the form of assessment written feedback. The reasons can be ascribed to insufficient discussions on 1) how the assignments were designed (e.g., whether allow learners to submit their improved work based on the given feedback to the same or similar subsequent assignment) and 2) how the effective feedback content can be constructed to support learners. As informed by [22], the assessment written feedback can enhance its potential of being dialogic feedback by making more efforts on the construction of feedback content (e.g., presenting explanations and suggestions on

students' works) and the design of the assignments (e.g., allowing students to resubmit their improved work based on the given feedback). However, most prior works [32–35] are insufficient to demonstrate insights into the design of dialogic feedback in assessment written assessments since the analysed feedback was given in the one-way form without allowing learners to submit their improved work based on the feedback they received. Driven by this, we argue that it is worthwhile to further investigate communicative patterns from assessment written feedback.

Building upon the dialogic feedback framework [1], a recent study [8] proposed a learner-centred feedback framework for analysing assessment written feedback. Ryan et al. [8] proposed a set of effective feedback attributes related to communicative patterns in assessment feedback. It should be noted that the learner-centred feedback framework [8] expands the *Feedback Triangle model* [1] into more detailed attributes for the design of assessment written feedback. Given the potential impact of learner-centred feedback framework [8] on assessment feedback analysis, we posit that more efforts should be devoted to using the learner-centred feedback framework to analyse the communicative patterns (e.g., how educators strengthen the relationships with learners and how educators promote learner independence in the feedback process) in the assessment feedback, which can help improve the practice of feedback design and the design of an automated assessment feedback system.

To address the above three gaps in the literature, we aimed to employ learning analytics approaches to investigate communicative patterns in both one-on-one online tutoring dialogue and assessment written feedback. Additionally, we aimed to facilitate the analysis of tutoring dialogues by enhancing the identification of dialogue acts from educators and learners in the tutoring dialogues.

1.3 Research Questions and Contributions

In the following, we present the research questions investigated in Chapters 2–6. Each chapter focuses on tackling one of the three research gaps, as detailed in Table 1.1. To gain a better understanding, we also demonstrate an overview of how each chapter tackled different research gaps in Figure 1.2. Chapters 2, 3 and 5 are focused on the investigation of communicative patterns (i.e., politeness and dialogue acts) in one-on-one online tutoring dialogues (i.e., Gap 1), Chapter 4 is dedicated to the enhancement of automatic classification of dialogue acts (i.e., Gap 2), and Chapter 6 is in the investigation of communicative patterns (e.g., politeness and instructional information) in assessment feedback (i.e., Gap 3).

To better understand the communicative patterns of dialogic feedback in the dialogue form, in **Chapter 2**, we tackled Gap 1 by focusing on investigating the role of politeness in one-on-one online tutoring dialogue, which can guide educational researchers and practitioners about the use

TABLE 1.1: An overview of the research gaps addressed in each chapter.

Chapter	Title	Research Gaps		
		Gap1	**Gap**2	**Gap**3
Chapter 2	Investigating the Role of Politeness in the Tutoring Dialogues	√		
Chapter 3	Mining Effective Instructional Strategies from the Tutoring Dialogues	√		
Chapter 4	Enhancing the Identification of Instructional Strategies from the Tutoring Dialogues		√	
Chapter 5	Exploring the Politeness of Instructional Strategies from the Tutoring Dialogues	√		
Chapter 6	Analysing the Communicative Patterns in Assessment Feedback			√

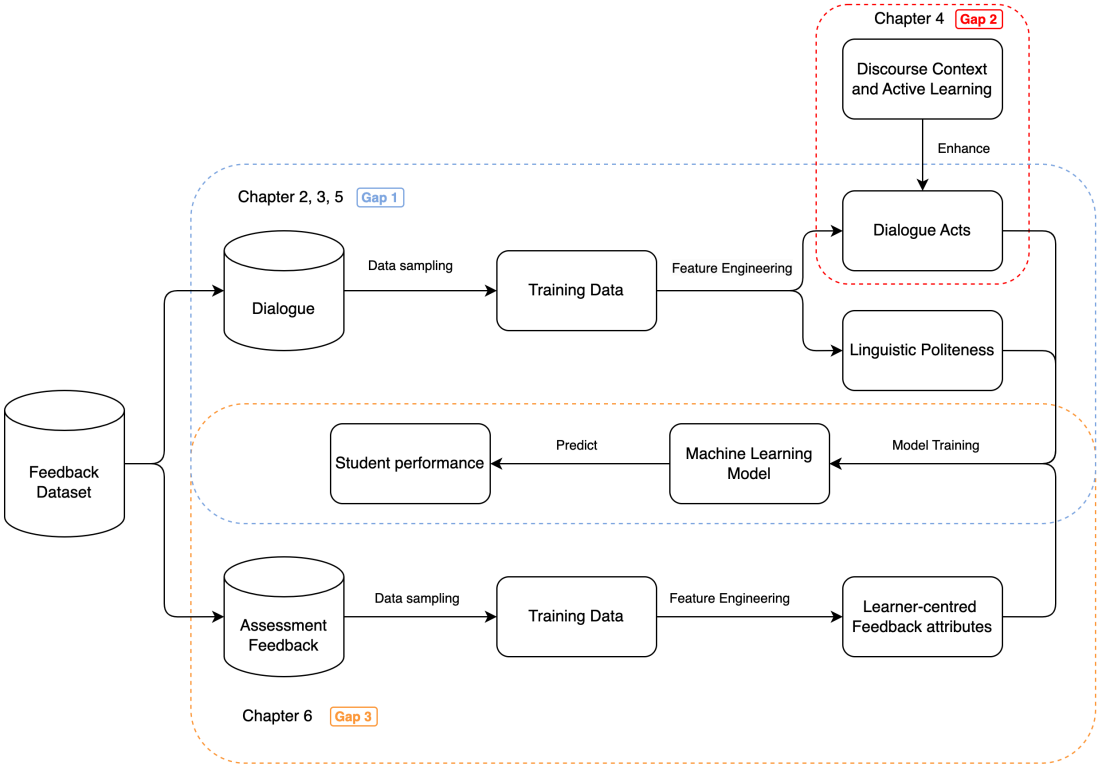


FIGURE 1.2: An overview of how each chapter tackled different research gaps.

of politeness in tutoring dialogue and inform the design of dialogue-based ITS. Though prior studies documented that polite expressions from educators were more effective than non-polite expressions in promoting learners' learning gains in different educational contexts [60–64, 68], it still remained unclear how politeness could be adopted in the one-on-one online tutoring dialogues and how the use of politeness from educators correlated with learners' academic performance. Driven by this, the current PhD study aimed to contribute to the feedback practice on politeness investigation in real-world tutoring dialogues, and further shed light on the design of dialogue-based ITS. Specifically, we investigated the following **R**esearch **Q**uestions:

RQ 1 *What is the manifestation of politeness in the one-on-one tutoring dialogue process?*

- **RQ 1.1** What politeness features characterise tutoring dialogues?

- **RQ 1.2** To what extent does the politeness level of educators' utterances correlate with the learners' problem-solving performance?
- **RQ 1.3** To what extent do the politeness features predict the learners' problem-solving performance?

To answer RQ1, we used the tutoring dialogue dataset provided by the *Yup* online tutoring service. The research on this dataset was approved by the *Human Research Ethics Committee of Monash University* under Project ID 26156. The dataset contained 15,723 tutoring dialogue sessions, which involved 5,165 K-12 learners and 116 qualified educators. For each tutoring dialogue, a learner often initiated a session by sending an unsolved problem (presented in an image or a textual message) as a request for help. Then, an experienced human educator was allocated to guide the learner to solve the problem by using textual messages and images. It should be noted that the educators were required to carefully guide learners to solve problems by themselves and were not allowed to directly share answers with the learners. Thus, the tutoring dialogue dataset contained detailed processes of how the learners and educators worked collaboratively on solving problems in different STEM subjects such as mathematics, chemistry and physics (among these subjects, 92% of tutoring subjects were related to maths tutoring). In the stage of data preparation, we also included one of the most important factors (i.e., learner prior knowledge) in the analysis. As suggested by Yang and Li [79], educators who support learners on the learning tasks should consider the learner's prior knowledge which was positively associated with learners' self-efficacy and learning performance. It should be noted that we had no access to the information indicating learners' prior knowledge of the mastery levels of a specific learning subject (e.g., mathematics). Instead, we used the learner's prior progress (i.e., the progress that learners made on the problems consulted with the educators before joining a tutoring session) as the proxy of the learner's prior knowledge. This is in line with the definition of learner prior knowledge (i.e., learner experiences on the learning contents before they have a tutoring session [80]).

To investigate the role of politeness in tutoring dialogues, we answered RQ1 based on the well-known politeness theory proposed by [81] to extract the features related to politeness from the tutoring dialogues by using two well-established tools developed by Danescu-Niculescu-Mizil et al. [82] and Niu and Bansal [83]. Informed by Brown and Levinson's (1987) *Politeness Theory*, we considered the analysis of politeness from two perspectives, namely, politeness strategy and politeness level. First, politeness strategy can be identified by the use of certain politeness markers, e.g., the marker `please`, "*Could you please check the value of x first?*". A study by [82] developed the **Politeness Strategy Tool**[1] to identify the markers of politeness strategy by applying natural language processing techniques. This tool can extract 21 politeness strategies (e.g., `HasPositive`, "*Good job*"). Then, to quantify the level of politeness in a piece of text,

[1] available in the `ConvoKit` toolkit `https://convokit.cornell.edu/`

Niu and Bansal [83] developed a tool (i.e., **Politeness Scoring Tool**[2]), which can generate scores ranging from 0 to 1 to represent politeness level from non-polite to polite. To answer RQ1.1, we extracted and analysed the frequency of politeness strategy and the levels of politeness expressed in educators' and learners' utterances from all tutoring dialogues. Then, we answered RQ 1.2 by investigating the correlation between the levels of politeness expressed in educators' utterances and learners' problem-solving performance. Lastly, we answered RQ 1.3 by investigating the predictive power of using politeness in predicting learner performance. We employed a well-established machine learning method—Gradient Tree Boosting (GTB) [84]— to predict learner performance. GTB is an ensemble classification method based on decision trees and has been recently shown to successfully handle classification tasks with multiple outcome variables.

Contribution. In Chapter 2, we employed linguistic politeness theories by Brown and Levinson [81] to characterise politeness in tutoring dialogues and investigated the correlation between the politeness of educators' utterances and learners' problem-solving performance. We found that educators tended to be very polite at the start of tutoring dialogue sessions and became more direct in guiding learners as the sessions progressed. The learners with better performance in solving problems tended to be more polite at the beginning and the end of tutoring dialogue sessions than their counterparts who failed to solve problems. Additionally, inspired by the works of Morrison et al. [85] and Schegloff and Sacks [86], we investigated the correlation between polite expressions by educators and learners' problem-solving performances in instructional communication (e.g., provide corrective feedback or suggestions on learners' works) and non-instructional communication (e.g., greeting and self-introduction). We found that the correlation between polite expressions by educators and learners' problem-solving performances was not evident in non-instructional communication. We also evaluated the power of politeness in predicting learners' problem-solving performance by applying GTB. We found that politeness alone cannot adequately reveal learners' problem-solving performance, and thus other factors should also be taken into account (e.g., sentiment contained in utterances).

As there was no evident correlation between polite expressions by educators and learners' performance in non-instructional communication, we further investigated the use of politeness in instructional communication. Informed by prior research [44, 87], instructional communication can be categorised into many specific types of instructional strategies (e.g., corrective feedback and thought-provoking questions), which can be encoded into dialogue acts. Informed by previous studies [44, 87], dialogue acts are often used to understand the conversational structure within a dialogue and a dialogue act can reflect the meaning of an utterance from the dialogue. Many prior studies [44, 72, 87] used dialogue acts to identify effective communicative patterns from the tutoring dialogues but the understanding of delivering effective instructional strategies is still limited, especially in large-scale tutoring dialogue sessions. To this end, the current

[2]available at https://github.com/WolfNiu/polite-dialogue-generation

PhD study aimed to address Gap 1 by further investigating the role of dialogue acts in tutoring dialogue in **Chapter 3**. Specifically, we investigate the following **R**esearch **Q**uestions:

RQ 2 *What is the manifestation of dialogue acts in the tutoring dialogue provision process?*

- **RQ 2.1** What dialogue acts are commonly taken by educators and learners during tutoring dialogue sessions?
- **RQ 2.2** What patterns of actions, i.e., one or multiple consecutive actions, are associated with different levels of learners' problem-solving performance in tutoring dialogue sessions?
- **RQ 2.3** To what extent are the identified actions and action patterns predictive of the problem-solving performance of learners in tutoring dialogue sessions?

To answer RQ 2, we again used the *Yup* tutoring dialogue dataset and the annotation of learner prior progress from Chapter 2. We employed a well-known dialogue act scheme proposed by Vail and Boyer [69], which was proposed for annotating the dialogue acts in the online textual dialogue tutoring environment and contained sufficient dialogue acts of learners and educators. Inspired by the prior research in analysing the tutoring dialogue Rus et al. [44], we labelled a subset of the original dialogue dataset, which was used to train a classifier by applying supervised machine learning techniques. Then, we used the classifier to automatically infer the dialogue acts for the remaining dataset. To enable a rigorous reproduction, we have built and open-sourced the pre-trained dialogue act classifier which is available at `https://github.com/bertDA/BertDA`. Since the dialogue acts of the remaining tutoring dialogues were automatically identified by our dialogue act classifier, we then analysed the distribution of these dialogue acts in the whole dataset to answer RQ 2.1. To answer RQ 2.2, we employed the `TraMineR` package in R to identify the consecutive patterns of dialogue acts in our dataset. `TraMineR` is a popular tool used to analyse sequences of states or events in data, which has been successfully applied to analyse the consecutive patterns of dialogue acts in tutoring dialogues [45, 88]. To answer RQ 2.3, we evaluated the effectiveness of dialogue act relevant features (i.e., the dialogue acts and consecutive patterns of dialogue acts) in predicting the learner's problem-solving performance.

By analysing a large corpus consisting of both successful tutoring dialogues (i.e., learners solved the learning problems in the tutoring process) and unsuccessful dialogues (i.e., learners did not solve the learning problems), our study contributed to an in-depth understanding of educators' as well as learners' behaviour in human-human online tutoring dialogues and offered empirical evidence to support existing effective practices (e.g., providing timely feedback to learners) for the development of dialogue-based ITS.

Contribution. In Chapter 3, we identified a set of dialogue act patterns that were pertinent to educators and learners across dialogues of different traits. For example, learners without prior progress in solving problems, compared to those with prior progress, were likely to receive more thought-provoking questions from their educators. Then, overall, educators often took action to provide feedback and information to learners and to ask thought-provoking questions to guide learners to solve problems. Correspondingly, learners took more dialogue actions in answering questions or expressing agreement with educators and acceptance of the provided explanations or solutions. Additionally, we also investigated the predictive power of incorporating dialogue act relevant features (i.e., the dialogue acts and consecutive patterns of dialogue acts) in the prediction of learner problem-solving performance by using the GTB model. We demonstrated that the dialogue act relevant features extracted from learners' and educators' utterances could not adequately predict learner performance and should be considered together with other relevant factors (e.g., the informativeness of the utterances).

Despite achieving promising classification performance (F1 score of 0.742 and Cohen's κ score of 0.735) in identifying dialogue acts from educators and learners, our dialogue act classifier (as reported in Chapter 3) still has room for improvement. From the analysis of misclassified instances by our dialogue act classifier, we have identified two main limitations. First, inadequate discourse contextual information during the training process might lead to incorrect identification of dialogue acts. Second, the dataset for training the dialogue act classifier is not sufficient, especially for the high informative training instances (i.e., data samples that help the classifer learn the underlying patterns and improve the classification accuracy on the unseen data [89]). Thus, it is necessary to identify the high informative training instances from the dataset and invite the human coders to annotate these high informative instances which are further used for training the dialogue act classifier. In an effort to address Gap 2, we aimed to optimise the dialogue act classification performance and alleviate the human labelling budget to ensure less human labour annotation and more accurate classification performance. As detailed in Figure 1.2, which depicts the different stages of this PhD thesis implementation, Chapters 2 and 3 are focused on the investigation of politeness and dialogue acts in tutoring dialogues, while Chapter 4 is dedicated to enhancing the automated classification of dialogue act from online tutoring dialogues.

As discussed in the previous paragraph, discourse contextual information might be an important factor influencing classification performance. However, little research has examined the extent to which the incorporation of discourse contextual information has an effect on the performance of an automated dialogue act classifier. Additionally, given the limitations reported in the previous paragraph, we acknowledged that annotating more utterances can enable the classifier to capture comprehensive patterns to improve the dialogue act classification. The labour-intensive issue of annotating dialogue acts has been raised in many other studies [70, 76, 90]. To mitigate the labelling budget for the dialogue act classification task, the previous work by Nye et al.

[90] suggested employing statistical active learning to selectively sample the utterances from the dialogue dataset for human experts to annotate. However, the potential of statistical active learning for the educational dialogue act classification task remains unknown. Motivated by the need to facilitate the analysis of effective tutoring dialogues in previous studies [70, 76, 90], the current PhD study focused on investigating the enhancement of the dialogue act classification in Chapter 4. Specifically, we investigated the following **R**esearch **Q**uestions:

RQ 3.1 *To what extent does the discourse context relate to the performance of automatic classification of educational dialogue acts?*

RQ 3.2 *To what extent can widely-used active learning methods alleviate the labour-intensive issues in the automatic classification of educational dialogue acts?*

To answer RQ 3.1, we used the utterances from the *Yup* tutorial dialogue dataset that we annotated the dialogue acts in the study reported in Chapter 3. We explored contextual information and evaluated widely-used language models (i.e., BERT [91] and ELECTRA [92]) for classifying educational dialogue acts. In line with the prior studies [93–95], we referred to the sentence being classified as the current segment and preceding sentences of the current segment as preceding segments (i.e., discourse contextual information). In a previous study, Ribeiro et al. [93] examined the impact of preceding segments (i.e., incorporating the preceding segments from 0 to 5) on dialogue act classification for a dataset consisting of general human communication dialogues and found that the dialogue act classifier trained on input with the incorporation of preceding segments achieved higher classification accuracy than the classifier without the incorporation of preceding segments. However, since the dialogue act scheme built upon general human communication might not fit well to analyse the educational dialogues [96], the influence of the preceding segments (i.e., discourse context) on dialogue act classification in the educational domain still remains unknown. Inspired by the dialogue act classification with the incorporation of discourse context in general human communication [93], the current PhD study investigated the preceding segments from 0 to 5 on the tutorial dialogue dataset to observe the impact of contextual information on educational dialogue act classification.

To answer RQ 3.2, we aimed to investigate the extent to which statistical active learning methods can alleviate the labour-intensive issue. In the recent work by [97], they developed two statistical active learning methods, namely CoreMSE and CoreLOG, which achieved state-of-the-art performance on many textual data classification tasks. The current study included both CoreMSE and CoreLOG to evaluate their efficacy on the educational dialogue act classification task. Additionally, we also investigated the recent statistical active learning methods which were mostly related to the CoreMSE and CoreLOG methods as the baseline methods for comparison, including MOCU [98], WMOCU [98], BADGE [99], ALPS [100] and the random baseline.

Contribution. We found that the machine learning models trained on the input that included the discourse context achieved better performance than the models excluding the discourse context. Besides, the effectiveness of the contextual information decayed after the machine learning models achieved optimal performance in classifying the dialogue acts. Our findings provided a strong motivation for incorporating the discourse context and finding the optimal length of incorporated contextual utterances to train a dialogue act classifier. Furthermore, to combat the high demand for manually annotating dialogue acts, we proposed to exploit existing active learning methods (e.g., CoreLOG and CoreMSE [97]) to alleviate the demand for annotating dialogue acts. We found that compared with the random baseline, active learning methods (i.e., CoreLOG and CoreMSE [97]) can select informative samples from the training dataset to train ML models, which can alleviate the labour-intensive issues. Our findings indicated that both statistical active learning methods (CoreLOG and CoreMSE [97]) have the potential to support the training process of dialogue act classifiers. We expect that both statistical active learning methods can be applied to more educational classification tasks that are confronted with the issues of annotation cost (i.e., time and financial expense) for ML model training, which might help unlock the potential of ML models by improving the accuracy of the classifiers and minimising the cost of annotation.

In Chapter 2, we investigated the role of politeness in tutoring dialogues and found that the correlation between polite expressions by educators and learners' performances was not evident in non-instructional communication (e.g., greeting). Thus, it is worthwhile to investigate the use of politeness in instructional communication (e.g., providing corrective feedback and suggestions). In Chapter 3, we investigated the role of dialogue acts in tutoring dialogues and developed a dialogue act classifier to automatically identify the dialogue acts which reflected specific types of instructional and non-instructional communicative patterns from educators. In Chapter 5, we further investigated the interplay between the use of politeness and instructional dialogue acts. As noted in Gap 1, both benefits and hindrances of guiding learners politely coexist in the dialogue communication process. To maintain the effectiveness of tutoring dialogues, it is important to understand the extent to which educators express politeness in different types of instructional communication (e.g., providing hints and corrective feedback). A promising way to investigate how educators effectively communicate with learners is to analyse the instructional communication based on learner performance (e.g., whether learners successfully completed learning tasks) [44]. Thus, the appropriateness of expressing politely in instructional communication can be manifested by investigating the relationship between the use of politeness in instructional communication and learner performance. By doing so, we deem that the use of appropriate politeness could better facilitate the online one-on-one tutoring dialogues. However, to our knowledge, no previous study attempted to examine this relationship in tutoring dialogues. Thus, we proposed the fourth **R**esearch **Q**uestion as below:

RQ 4 *How is the politeness used in the instructional communication by educators in tutoring dialogue?*

To answer RQ 4, we also used the *Yup* tutoring dialogue dataset and the annotation of learner prior progress from Chapter 2. Building upon the efforts in Chapter 2 and Chapter 3, we first investigated the extent to which educators express politeness in different instructional communication in tutoring dialogues. Then, we investigated the capability of instructional politeness on the prediction of learner performance. The prediction of learner performance was implemented by adopting the GTB model which has demonstrated effectiveness in the prediction of learner problem-solving performance in Chapter 2 and Chapter 3.

Contribution. In the study reported in Chapter 5, we analysed the use of politeness in instructional communication from the dialogue corpus and demonstrated the politeness levels for multiple commonly used instructional communication (e.g., open question and positive feedback). Additionally, as informed by the findings in Chapter 2 (i.e., the politeness might be varied as the tutoring progressed), we also investigated the change of politeness levels for different types of instructional communication as the tutoring progressed but we did not observe a significant change of politeness in instructional communication. Furthermore, we examined the capability of using politeness levels in instructional communication to predict learner performance. We showed that by incorporating the features of politeness in instructional communication with the features proposed in [38] (e.g., the sentiment levels of the utterance, complexity, and informativeness), the GTB model achieved better model performance of predicting learner performance compared to the prediction performance of the model reported in Chapter 3. These findings indicate that the features of politeness in instructional communication can improve the prediction of learners' problem-solving performance. We expect that our results can provide guidelines for human educators to consider the extent to which politeness can be expressed in instructional communication and shed light on the design of dialogue-based ITS.

Given the promising analysis results of communicative patterns in tutoring dialogues, we also investigated the potential of communicative patterns in assessment feedback (i.e., a typical type of delayed feedback). As discussed in Gap 3, assessment written feedback is a widely-used form of feedback, especially for higher education [27]. However, by scrutinising the prior works of analysing communicative patterns in assessment written feedback [32–35], we argue that these works are insufficient to demonstrate the insights of effective communicative patterns from the assessment written feedback based on the design of dialogic feedback as suggested by [1, 22]. To support the practice of crafting feedback and the design of automated feedback systems, it is worthwhile to devote efforts to investigate more communicative patterns (e.g., how educators strengthen the relationships with learners via polite communication and how educators promote learner independence in the feedback process via the use of instructional information

in the written feedback) from the assessment written feedback. To this end, we propose the fifth **R**esearch **Q**uestion as below:

RQ 5 *What communicative patterns of assessment feedback are predictive of learners' task performance?*

To answer RQ5, we used the assessment feedback dataset collected from an introductory data science course taught in English at the postgraduate level. The dataset was provided by Monash University and the research on using this dataset was approved by the *Human Research Ethics Committee of Monash University* under Project ID 29874. The dataset contained the textual feedback comments offered by educators who provided feedback based on the learner's submission to the first assessment (namely, Assignment I). Assignment I required the learners to write a proposal report to introduce a data science problem to be solved and describe relevant application backgrounds and types of business models. After the learners had submitted Assignment I (i.e., proposal report), educators marked the learners' Assignment I and provided feedback to help the learners work on the second assessment (namely, Assignment II). In Assignment II, the learners could use the feedback from Assignment I to write a final report on their introduced data science topic. After learners submitted Assignment II (i.e., the final report), educators graded learners' work and provided corresponding feedback. To measure learners' task performance, inspired by the work Nicoll et al. [32], we measured learner grade changes by subtracting the learners' grades in Assignment I from their grades in Assignment II. The positive values by the subtraction indicated learners achieved performance increase on Assignment II, and we encoded the records of the positive values as the `Increase` group. Whereas, the other calculated results were encoded as the `Not Increase`.

We relied on a comprehensive learning-centred feedback framework [8] to extract the features of communicative patterns from assessment feedback by using the software, including Linguistic Inquiry and Word Count (LIWC) [41], Linguistic Politeness [101], and Coh-Metrix [42]. Then, we measured the correlation between the features of communicative patterns and the learner grade changes (i.e., `Increase` and `Not Increase`) by using statistical tests. Additionally, we employed widely-used machine models in previous feedback studies (e.g., Gradient Tree Boosting [35], Logistic Regression [32], and Random Forest [33, 35]) to predict learner performance. Though predicting learner performance has significant implications on the evaluation of feedback quality [32], we believe that it is equally important to understand how the features of communicative patterns contribute to the predicted outcome since the interpretability of the results can enhance human trust in the analysis results [102–104]. Thus, we decided to adopt the well-established interpretable framework **SHAP**[3] (**SH**apley **A**dditive ex**P**lanations) [104] to understand the contribution of features on model prediction performance.

[3] https://shap.readthedocs.io/

Contribution. We analysed communicative patterns of feedback on Assignment I and observed learner grade changes (i.e., `Increase` and `Not Increase`) on the subsequent assignment, which is connected to Assignment I. With the use of the learner-centred feedback framework [8], we examined the features of communicative patterns among different groups of learner grade changes (i.e., `Increase` and `Not Increase`). Our results revealed some insights about the use of communicative patterns in the assessment written feedback. For example, we observed that compared to the learners in the `Not Increase`, the learners in the `Increase` group received feedback with more use of detailed suggestions about future improvement on the subsequent assignment. Additionally, the learner's performance on Assignment I was considered the most important indicator for differentiating improvement on the subsequent assignment. Furthermore, we found that the **GTB** model achieved better performance in predicting learner grade changes compared to the baseline models. The **SHAP** framework demonstrated its potential to transparently reveal the feature importance in the machine learning model. For example, with the use of the **SHAP** framework, our study identified the positive emotional words in the feedback as one of the most significant features for model prediction performance and also demonstrated that the high frequency of positive emotional words in feedback negatively correlated with the learner grade increase on the subsequent assignment.

To summarise, we focused on the analysis of different communicative patterns on various types of feedback in Chapters 2–6, as detailed in Figure 1.2. Since the main focus of this PhD thesis was the dialogic feedback in the form of one-on-one online tutoring dialogue, we first investigated the communicative patterns (politeness and dialogue acts) in tutoring dialogues in Chapters 2–5. Then, we focused on the investigation of communicative patterns in assessment feedback in Chapter 6.

1.4 Thesis Structure

This thesis consists of seven chapters. All main chapters (Chapter 2–6) are based on full research papers published or submitted for publication in conferences or journals.

- **Chapter 1** includes research background, research questions, and contributions.
- **Chapter 2** is based on the paper submitted in the *British Journal of Educational Technology* (presently under review after minor revisions were requested).
- **Chapter 3** is based on the paper published in the *Future Generation Computer Systems* [38].
- **Chapter 4** is based on the paper submitted in the *IEEE Transactions on Learning Technologies*.

- **Chapter 5** is based on the paper published in the Proceedings of the *12th International Learning Analytics and Knowledge Conference* [37].
- **Chapter 6** is based on the paper accepted for publication in the Proceedings of the *13th International Learning Analytics and Knowledge Conference* [39].
- **Chapter 7** concludes and discusses the present and future work.

Chapter 2

Investigating the Role of Politeness in Tutoring Dialogues

2.1 Introduction

Researchers have been developing dialogue-based Intelligent Tutoring Systems (ITSs) to alleviate the workload of providing dialogue-based feedback and instruction by human educators [54–57]. Despite many successful implementations of dialogue-based ITSs for supporting learners' academic performance (e.g., AutoTutor assisted learners in learning the fundamentals of operating systems and many other learning subjects [49, 55–57], Rimac proactively supports learners to learn physics [58], and Beetle II boosted learner understanding of concepts in electronics and electricity [50]), the communicative patterns of these ITSs still need to be improved to become comparably effective as human educators [45, 105–109]. One of the major research gaps documented in prior research is the inability of dialogue-based ITSs to communicate politely with a learner (i.e., Gap 1 described in Chapter 1). Less polite feedback, that a system delivers to a learner (e.g., *"Not exactly! Try it again*"), often neglects learners' feelings. This, in turn, might impose negative feelings on the learners [66, 67], and further, impede learners' self-efficacy, motivation and acceptance of feedback [30]. Such expressions that can harm learners' self-efficacy are called direct expressions [81]. Human educators, in contrast, often tailor their feedback in a polite form (e.g., *"How about we check this error together?"*) and in this way build solidarity and rapport [66] that has been widely documented to benefit learners' academic performance [60–64, 68]. Therefore, we argue that the dialogue-based ITS should integrate linguistic politeness into their utterances to guide learners. By doing so, we deemed that the effectiveness of dialogue-based ITS can be further enhanced. However, limited prior studies

investigated politeness in online human-human tutoring dialogues and further discussed the necessity of equipping politeness in the dialogue-based ITSs. Thus, we aim to investigate the role of politeness in human-human tutoring dialogues (RQ1).

To address Gap 1 (described in Chapter 1), we adopt the learning analytic approaches to investigate the role of politeness by analysing a large-scale real-world dataset consisting of over 15K online human-human tutorial dialogues. By addressing this gap, we aimed to shed light on the design of dialogue-based ITS and the guidance of feedback practice. The research outputs that directly addressed this gap have been submitted/published in two academic publications. We include the second publication (a journal paper) in the next section since the journal paper is a substantial extension of the first conference paper:

- Lin, J., Lang, D., Xie, H., Gašević, D., & Chen, G. (2020, July). Investigating the Role of Politeness in Human-Human Online Tutoring. In *International Conference on Artificial Intelligence in Education* (pp. 174-179). Springer, Cham.
- Lin, J., Raković, M., Xie, H., Lang, D., Gašević, D., & Chen, G. (2023). On the Role of Politeness in Online HUman-Human Tutoring. *British Journal of Educational Technology* (Under Review, Minor Revision).

2.2 Publication: On the Role of Politeness in Online Human-Human Tutoring.

On the Role of Politeness in Online Human-Human Tutoring

Jionghao Lin[a], Mladen Raković[a], Haoran Xie[b], David Lang[c], Dragan Gašević[a] and Guanliang Chen[a]

[a]Faculty of Information Technology, Monash University, 20 Exhibition Walk, Clayton VIC 3800, Australia
[b]Department of Computing and Decision Sciences, Lingnan University, Hong Kong
[c]Stanford University, 450 Serra Mall, Stanford, CA 94305, United States

ARTICLE INFO

Keywords:
Intelligent Tutoring Systems
Learning Analytics
Politeness Strategies
Student Performance
Predictive Analysis

ABSTRACT

Researchers have demonstrated that dialogue-based Intelligent Tutoring Systems (ITS) can be effective in assisting students in learning. However, little research attempted to explore the necessity of equipping dialogue-based ITS with one of the most important capabilities of human tutors, i.e., maintaining politeness with students, which is essential to provide students with a pleasant learning experience. In this study, we examined the role of politeness by analysing a large-scale real-world dataset consisting of over 14K online human-human tutorial dialogues. Specifically, we employed linguistic theories of politeness to characterise the politeness levels of tutor-/student-generated utterances, investigated the correlation between the politeness levels of tutors' utterances and students' problem-solving performance, and quantified the power of politeness in predicting students' problem-solving performance by applying Gradient Tree Boosting. We showed that: i) in the effective tutorial sessions (i.e., those in which students successfully solved problems), tutors tended to be very polite at the start of a tutorial session and become more direct to guide students as the session progressed; ii) students with better performance in solving problems tended to be more polite at the beginning and the end of a tutorial session than their counterparts who failed to solve problems; iii) the correlation between polite expressions by tutors and students' performances was not evident on students' performance; and iv) politeness alone cannot adequately reveal students' problem-solving performance, and thus other factors (e.g., sentiment contained in utterances) should also be taken into account.

1. Introduction

Tutoring that involves human participants, i.e., human-human tutoring, has been widely recognised as an effective instructional method across different educational levels (Bloom, 1984; Slavin, 1987; Lin, Singh, Sha, Tan, Lang, Gašević and Chen, 2022b). Despite its effectiveness, human-human tutoring has often been deemed as a time-demanding and labour-intensive instructional approach. This is particularly true in high-enrolment courses where several tutors need to be hired and multiple tutoring sessions need to be arranged to provide sufficient support to each student. Inspired by the recent advancements in technology-enhanced learning environments, researchers have begun developing dialogue-based Intelligent Tutoring Systems (ITSs) (Alkhatlan and Kalita, 2019) to provide timely and adaptive support to students while taking into account their individual differences, e.g., personalities, emotional needs, and levels of engagement (Alkhatlan and Kalita, 2019; Hasan, Noor, Rahman and Rahman, 2020). Many dialogue-based ITSs developed to date have demonstrated empirical benefits in supporting hundreds of thousands of students to study a variety of subjects, e.g., AutoTutor for operating systems (Nye, Graesser and Hu, 2014) and Rimac for physics (Albacete, Jordan, Lusetich, Chounta, Katz and McLaren, 2018)

Despite their promises and comparable instructional performance to human-human tutoring, dialogue-based ITSs can be potentially improved to become more effective in guiding students (Almasri, Ahmed, Almasri, Abu Sultan, Mahmoud, Zaqout, Akkila and Abu-Naser, 2019; Maharjan and Rus, 2018; Maharjan, Rus and Gautam, 2018; Tang, Liang, Hare and Wang, 2020; Schez-Sobrino, Gmez-Portes, Vallejo, Glez-Morcillo and Redondo, 2020). One of the major challenges documented in prior research is the inability of dialogue-based ITSs to communicate politely with a student (Wang, Johnson, Mayer, Rizzo, Shaw and Collins, 2008; Lin, Lang, Xie, Gašević and Chen, 2020). In particular, less polite feedback delivered by ITSs (e.g., "*You made this same mistake again.*") may diminish students' motivation and, subsequently, affect their learning performance. Human tutors, in contrast, often tailor their feedback in a polite form (e.g., "*How about we check this error together?*") to build solidarity and rapport with students, which

guanliang.chen@monash.edu (G. Chen)
ORCID(s):

in turn may benefit their learning processes (Wang, Johnson, Rizzo, Shaw and Mayer, 2005; Wang and Johnson, 2008; Johnson and Wang, 2010; Wang et al., 2008; Gupta, Walker and Romano, 2007; Mikheeva, Schneider, Beege and Rey, 2019; McLaren, Lim, Yaron and Koedinger, 2007; McLaren, DeLeeuw and Mayer, 2011a) and enhance learning performance, especially of low performing students (McLaren, DeLeeuw and Mayer, 2011b; D'Mello, Olney, Williams and Hays, 2012). Therefore, we posit the expressions of politeness need to be integrated into utterances automatically generated by dialogue-based ITSs and thus resemble polite utterances from human tutors. In this way, the instructional effectiveness of an ITS can be further enhanced. However, little research investigated politeness in online human-human dialogue-based tutoring and further discussed the necessity of equipping politeness in the dialogue-based ITSs.

For the above reason, we attempted to examine (1) the role of politeness in online human-human tutoring dialogues and (2) the relationship between tutor politeness and student performance at the end of a tutoring session. Relying on the Brown and Levinson theoretical framework of politeness (Brown and Levinson, 1987), we focused on the politeness features composed by two groups of features, politeness strategy and politeness level. As the politeness is mutually influenced between the interlocutors (Group et al., 2011) in the communication process, we extracted and analysed politeness features from dialogue transcripts collected during tutorial sessions between students and tutors, and used these features as input to a Gradient Tree Boosting (GTB) (Chen and Guestrin, 2016) machine learning model to predict student performance. Formally, our study was guided by the following research questions:

- **RQ 1** What are the politeness features characterising dialogue between tutors and students in online human-human tutoring?
- **RQ 2** To what extent does the politeness level of tutors' utterances correlate with the students' performance?
- **RQ 3** To what extent do the politeness features predict the students' performance in online human-human tutoring?

2. Related Works

2.1. Politeness in Tutoring Communication

Polite communication helps to build solidarity and rapport, and induces effective interaction among interlocutors in both informal (e.g., everyday greetings, requests and expressions of gratitude) and formal settings (e.g., business letters (Brown and Levinson, 1987; Economidou-Kogetsidis, 2015) and instruction (Brummernhenrich and Jucks, 2013)). Brown and Levinson theorised that expressing politeness is associated with how a person would like to be perceived by other persons (Goffman, 1967; Brown and Levinson, 1987). In tutoring communication, the tutor's corrective feedback such as "*The information in this paragraph is not valid*" can threaten a student's confidence as it confronts the student with their own misunderstandings or mistakes (Brummernhenrich and Jucks, 2013; Jucks, Päuler and Brummernhenrich, 2016). Moreover, directive instructions (i.e., commands) such as "*Write this paragraph again*" can restrict the student's work autonomy (Brummernhenrich and Jucks, 2013; Jucks et al., 2016).

To alleviate these negative effects, tutors can articulate their utterances as *polite expression* to show their respect to students. Brown and Levinson suggested two politeness strategies to this end, *Positive* and *Negative Strategies*. *Positive Politeness Strategy* emphasises common ground and cooperation, e.g., "*Let's review this paragraph together to improve it further.*", while *Negative Politeness Strategy* aims at preserving student autonomy, e.g., "*Do you want to rewrite this paragraph?*". It should be noted that the politeness of tutors' expressions can also be affected by students' politeness in the tutoring communication as the politeness is mutually influenced between the interlocutors (Group et al., 2011) which indicates that human tutors who receive polite utterances from students may be more likely to formulate polite utterances (Brummernhenrich and Jucks, 2016).

2.2. Polite Guidance Can Boost Students' Performance

Polite expressions from tutors have been found to be more effective than direct expressions in promoting students' learning gains in different educational contexts (Wang et al., 2005; Wang and Johnson, 2008; Wang et al., 2008; Gupta et al., 2007; Mikheeva et al., 2019; McLaren et al., 2011b). For example, Wang et al. (2005) showed that students who received politely expressed utterances (e.g., "*We should set the planning methods now.*") from tutors in a web-based tutoring system outperformed their peers who were given direct expressions (e.g., "*Set your planning methods now.*") on a challenging problem-solving task. Later findings from the Wang et al. (2008) and Wang and Johnson (2008) studies further confirmed the positive effects of polite tutoring communication on students' performance in solving

industrial engineering problems and in learning a foreign language. McLaren et al. (2011b) examined the effects of polite tutoring in a web-based tutoring system and found that students with low prior knowledge who read tutorial instructions written in a polite way boosted their learning performance in a course. Further, Schneider, Nebel, Pradel and Rey (2015) demonstrated that students who read politely written task instructions outperformed their peers who were given directly expressed instructions. More recently, Mikheeva et al. (2019) found that feedback phrased politely improved student achievements in solving complex mathematical problems at the university level.

Despite growing evidence of positive effects of polite communication on learning performance, tutors often need to engage in additional work to properly phrase their utterances and make them polite (Person, Kreuz, Zwaan and Graesser, 1995). This may be a particularly challenging and time-consuming task in assignments that involve free responses to a prompt, e.g., proposing and justifying a solution to a real-world problem. The number of steps and solutions students propose while working on those problems can vary across individuals, and, to maintain productive communication, tutors need to attend to a variety of issues (e.g., selected approach, quality of evidence and clarity of description) and timely tailor polite utterances to help learners progress. Altogether, this adds to the already high workload of tutors. One way to alleviate this issue is to inform the dialogue-based ITSs about the linguistic properties of expressing politeness and the system will then use this information to generate utterances in a polite form (Sottilare, Graesser, Hu and Goldberg, 2014; Walker and Ogan, 2016). To understand how the construction of politeness plays a part in dialogue-based tutoring, a recent study (Lin et al., 2020) analysed the politeness levels of tutor-student utterances from the beginning and end of the tutoring dialogues. However, their analysis (Lin et al., 2020) did not incorporate the student ability (e.g., prior knowledge level) which is considered an important factor affecting the positive effect of politeness on students (Boyer, Phillips, Wallis, Vouk and Lester, 2008; McLaren et al., 2011b). As a remedy, a later study (Lin, Rakovic, Lang, Gasevic and Chen, 2022a) identified the students' working progress on a problem before the students joined the tutoring dialogues and further examined the politeness of the tutors' instructional strategies in the tutoring dialogues. Lin et al. (2022a) found that the politeness levels of some instructional strategies varied. One example is the instructional strategy of general positive feedback (e.g., "*Well done!*") was presented more politely compared to negative feedback (e.g., "*No, you are wrong.*"). Their results (Lin et al., 2022a) can inform the researchers of dialogue-based ITSs about the extent to which politeness should be expressed in instructional utterances. However, it still remains unclear how politeness has been adopted in online dialogue-based tutoring as well as the association between the polite expressions from tutors and students' problem-solving performance. With the current study, we aimed to contribute to the literature of politeness investigation with insights on how politeness was adopted from human-human online tutoring, and further shed light on the design of dialogue-based ITS.

3. Methods

3.1. Dataset

Table 1
The descriptive statistics of the dataset used in the study. **Without PP** and **With PP** stand for *Without Prior Progress* and *With Prior Progress* students, respectively. We employed Mann-Whitney tests to examine the differences of values in Rows 5-9 between any two of the session categories (Gap-clarified, Gap-explained, and Gap-bridged) in which students had the same status of prior progress. All differences were presented statistical significance ($p < 0.01$).

Metrics	All	Gap-clarified		Gap-explained		Gap-bridged	
		Without PP	With PP	Without PP	With PP	Without PP	With PP
1. # total sessions:	14,562	1,302	1,203	1,931	1,255	4,389	4,482
2. # total utterances:	1,216,784	30,128	31,014	113,099	78,575	488,119	475,849
3. # tutors:	116	96	92	99	98	106	110
4. # students:	5,165	962	763	1,419	908	2,168	1,800
5. Avg. Sess Dur (mins):	30.27 ± 30.66	9.75 ± 7.21	10.55 ± 7.64	22.88 ± 18.05	25.94 ± 19.03	38.60 ± 37.37	37.78 ± 32.17
6. Avg. # Uttr / Sess:	83.56 ± 81.05	23.14 ± 14.92	25.78 ± 16.68	58.57 ± 42.73	62.61 ± 43.79	111.21 ± 93.70	106.17 ± 87.62
7. Avg. # Words / Sess:	647.75 ± 596.12	198.13 ± 134.44	201.62 ± 131.81	489.56 ± 346.82	524.09 ± 351.18	845.28 ± 675.05	807.46 ± 649.08
8. Avg. % Uttr by tutors:	58.42 ± 7.86	56.46 ± 9.51	53.95 ± 9.49	60.25 ± 7.77	58.68 ± 7.82	59.75 ± 6.94	58.03 ± 7.07
9. Avg. % Words by tutors:	78.36 ± 9.10	80.54 ± 10.20	74.32 ± 11.80	82.21 ± 7.96	78.87 ± 8.59	79.30 ± 7.81	76.09 ± 8.69

This research was conducted with ethics approval from `Monash University`. We analysed an anonymised tutoring dialogue dataset provided by an online education company based in the US. The dataset contained data about 5,165 K-12 students and 116 qualified tutors. The students' demographic information (e.g., gender and age) was not

collected in the existing dataset. The students and tutors were working collaboratively in solving different problems, e.g., chemistry, mathematics, and physics. As per the tutoring service policy, tutors were required to guide students to solve problems themselves, rather than directly offering students a final solution. For this reason, the collected utterances in our dataset contained fine-grained details that reflect the scaffolding process. The dataset for our analysis contained 14,562 dialogue sessions and 92% of them were mathematics tutorial sessions. The details of the dataset are presented in Table 1. We observed that the majority of the dialogue sessions were categorised as Gap-bridged (8,871, 60.9%), indicating that more than a half of the students managed to solve their problems at the end of the session. Further, our dataset contained 21.9% of sessions as Gap-explained and 17.2% of sessions as Gap-clarified. Rows 5-7 of Table 1 indicate that the tutors and students invested more effort in the Gap-bridged than in the Gap-clarified and Gap-explained dialogues. One reason for this may be that the more efforts (as measured by the number of utterances, number of words or duration) the students and tutors made in a tutoring session the better the tutoring outcome was.

In the data preparation stage, we first manually annotated utterances reporting on students' prior progress. At the outset of a tutoring session, tutors commonly asked students about the progress they had made on the task. Yang and Li (2018) showed that a tutor should communicate with students by taking into account the students' prior progress and this positively correlated with the students' perceived self-efficacy and further predicted their problem-solving performance. As suggested by previous empirical studies (Wang and Johnson, 2008; Boyer et al., 2008), students with low progress toward the learning goal can benefit from polite utterances from tutors and increase their self-efficacy and subsequent achievements. For this reason, we captured prior progress in each session and included it as a control variable in our analysis. In our study, we annotated the student as *With prior progress* based on whether a student could demonstrate their effort in solving a problem. Specifically, we labelled each utterance reporting on students' prior progress as *With prior progress*, e.g., "*I have multiplied the two factors, but I'm a bit stuck after that*" or *Without prior progress*, e.g., "*I wasn't sure where to start.*". Three coders were involved in the utterance annotation process. Each utterance was annotated independently by two coders. The overall agreement between these two coders was substantial with a percentage agreement of 0.847 and Cohen's κ score of 0.735. The third coder was involved afterwards to resolve inconsistencies between the two coders. As the results of annotation, we observed that more than 47% of students made some progress in the task before attending the tutoring session (Table 1).

Next, we manually annotated each tutorial session relative to the problem-solving performance of a student at the end of the session. To this end, we developed the following three subcategories: *Gap-clarified*, *Gap-explained* and *Gap-bridged*.

- **Gap-clarified**, a tutor identified the problem but was uncertain whether the student gained any progress at the end of the session;
- **Gap-explained**, a tutor identified the problem and helped the student make progress throughout the session, however, the student did not reach a full solution at the end of the session; and
- **Gap-bridged**, a tutor identified the problem and helped the student reach the solution.

It should be noted that each dialogue session was first annotated by an educational expert employed by the online education company that collected the dialogue dataset. To examine the reliability of the annotations, we recruited one coder who randomly sampled 500 tutorial dialogues and annotated them independently. We observed a percentage agreement score of 0.884 and Cohen's κ score of 0.787. We attached sample dialogues for each of the subcategories in the digital appendix[1].

3.2. Inferring Politeness from Tutorial Utterances

To answer RQ1, we extracted and analysed the politeness features in tutors' utterances. In our analysis, we considered politeness features from two perspectives, politeness strategy and politeness level, informed by Brown and Levinson's (1987) Politeness Theory. To extract politeness strategy and politeness level for tutors' utterances, we utilised computational tools developed in previous research (Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec and Potts, 2013; Niu and Bansal, 2018).

We adopted the politeness strategy classifier (PSC) (Danescu-Niculescu-Mizil et al., 2013) to extract politeness strategies in this study. The PSC is a domain-independent classifier based on the Politeness Theory (Brown and Levinson, 1987). It utilises politeness markers in written text, e.g., "*please*", to detect politeness strategies across 21 different

[1] http://bitly.ws/uRpZ

types (Table 2). We show the names of 21 strategies in PSC, strategy interpretation, and example utterances in the Table 2. It should be noted that strategies marked with ⋆ are more likely to incur a sense of direct expression, while the other strategies are more likely to incur a sense of polite expression (Danescu-Niculescu-Mizil et al., 2013). The PSC in the present study has been implemented using the `ConvoKit` toolkit[2].

To extract politeness levels from tutors' utterances, we adopted politeness level identifier (PLI), a state-of-the-art classification tool proposed by Niu and Bansal (2018)[3] which was built upon the Politeness Theory (Brown and Levinson, 1987). This tool can be used to determine whether a piece of text is (e.g., utterance, sentence, and paragraph) polite or direct. Specifically, the tool outputs the politeness level as a score ranging between 0 and 1, representing from very direct to very polite Niu and Bansal (2018). Since PLI was originally developed using textual data different from the data in this study, we validated the tool on the utterances in our dataset. Specifically, we randomly selected 500 utterances to be independently annotated as either polite or direct by two human coders. The percentage agreement between the coders was 0.724 and the Cohen's *kappa* score was 0.531, which was sufficient to proceed with the analysis (Neuendorf, 2017). The third human coder was involved in resolving the misaligned cases. We compared the 500 utterances annotated in this way to the results produced by the PLI tool and obtained Cohen's *kappa* score of 0.639, which was deemed a sufficient indicator that the tool can be validly applied in the context of our study.

Table 2
The 21 politeness strategies according to **Danescu-Niculescu-Mizil et al. (2013)**.

Strategy	Interpretation	Examples in our dataset
Gratitude	Strategy expresses gratitude	*Thank you for using TUTORINGSERVICE.*
Deference	Strategy expresses praise or approval	*Awesome!*
Greeting	Strategy expresses greeting	*Hey!*
HASPOSITIVE	Positive sentiment lexicons from Liu, Hu and Cheng (2005)	*Great. You did an amazing job.*
HASNEGATIVE ⋆	Negative sentiment lexicons from Liu et al. (2005)	*You are wrong*
Apologizing	Strategy expresses apology	*I am sorry if I confused you.*
Please	The word *Please* in a sentence but not at start	*OK, hold on please*
Please_start ⋆	The word *Please* at the start of a sentence	*Please tell me what it is and what its domain is.*
Indirect_(btw)	Strategy avoids using words expressing imperative	*By the way, here goes an example.*
Direct_question ⋆	Questions start with "What", "Why", "Who", and "How"	*What is the subscript on it?*
Direct_start ⋆	Sentences start with words expressing imperative	*Check this definition*
SUBJUNCTIVE	Counterfactual modal words	*Could you elaborate more on this factor?*
INDICATIVE	Indicative modal words	*Can you tell me what is the value of X?*
1st_person_start	First-person singular form at the start of a sentence	*I will try how I'm doing*
1st_person_pl	First-person plural form	*We can factor this trinomial.*
1st_person	First-person singular form in a sentence but not at start	*Yes, I will start working on it*
2nd_person	Second-person form in a sentence but not at the start	*As you can see, it would be outside the triangle.*
2nd_person_start ⋆	Second-person form at the start of a sentence	*You do not have to calculate in decimals*
Hedges	Hedging words after a noun or pronoun	*It seems a little confusing for me*
Factuality ⋆	Words indicating being actual	*No, I mean that the easy one that I really know*
HASHEDGE	Strategy expresses hedging in a sentence	*Probably being careful*

3.3. Effectiveness of Polite Tutoring

To answer RQ2, we investigated how the politeness expressed by tutors during the session predicts students' performance at the end of the session. Expanding upon prior studies that typically analysed instructional expressions at course-grained levels, i.e., direct vs polite, (Wang and Johnson, 2008; Wang et al., 2008; Gupta et al., 2007; Mikheeva et al., 2019; McLaren et al., 2011b, 2007, 2011a), we followed a fine-grained analytical approach by Niu and Bansal (2018) to categorise tutors' utterances into five politeness groups *Direct*, *Weak Direct*, *Neutral*, *Weak Polite*, and *Polite*. Our study investigated the average PLI politeness level score across all the utterances a tutor created in a session. The average scores from all tutorial sessions were sorted in ascending order and then equally divided into five categorical politeness groups.

[2] `https://convokit.cornell.edu/`
[3] https://github.com/WolfNiu/polite-dialogue-generation

3.4. Prediction on Student Performance

3.4.1. Prediction Model

To answer RQ3, we trained a Gradient Tree Boosting (GTB) (Chen and Guestrin, 2016) machine learning classifier. The classifier predicted student performance at the end of tutoring sessions as a categorical outcome (i.e., Gap-clarified, Gap-explained, and Gap-bridged) relying upon an extensive set of 145 politeness features as input (as detailed below). GTB is an ensemble classification method based on decision trees and has been recently shown to successfully handle classification tasks with multiple outcome variables, similar to our previous studies (i.e., (Lin et al., 2022b; Chen, Lang, Ferreira and Gasevic, 2019). We selected the Random Forest model as the baseline model. The Random Forest model was also built upon the ensemble classification methods, which had been demonstrated to generally outperform 179 commonly used machine learning models (e.g., Decision Trees, Support Vector Machines, and Naive Bayes) on 121 different classification tasks (Fernández-Delgado, Cernadas, Barro and Amorim, 2014).

3.4.2. Feature Engineering

The politeness displayed by interlocutors involved in a conversation is often mutually affected, which means that a student's politeness is likely to be affected by a tutor's politeness (Group et al., 2011; Brummernhenrich and Jucks, 2013). Therefore, we also include students' politeness in our feature engineering step. Then, to understand the extent to which politeness predicts students' performance, we decided to capture the frequency and percentage of using the politeness strategies in the dialogue, which were considered commonly used features in conducting educational research (Korb, 2013) and we named these features as holistic features. For each specific politeness strategy (e.g., *Please_start*), it is worthwhile to investigate how each politeness strategy can affect the prediction, so we also recorded as features the frequency and percentage of each politeness strategy in the dialogue. As tutors and students exhibited statistically different politeness levels (Lin et al., 2020), we opted to calculate the average politeness levels of the utterances in a dialogue. Therefore, our study categorised the features into three groups (i.e., **Holistic features**, **Strategy-level features**, and **Politeness-level features**). To train the GTB classifier, we engineered a total of 145 politeness features based on three feature groups from the utterances by both tutors and students. We calculated each feature by considering utterances generated by (i) tutors only, (ii) students only, and (iii) general (both tutors and students). The politeness features were broadly categorised into the following groups:

- **Holistic features**, capturing general characteristics of politeness in a dialogue. Using `PSC`, we extracted: (i) *# Politeness Strategies*, the number of politeness strategies per utterance, across 21 proposed types [4] and (ii) *% Politeness Strategies*, the fraction of different types of polite or direct strategies contained in utterances.
- **Strategy-level features**, which measured the usage of a specific strategy in dialogue by a tutor, by a student, and by both of them: (i) *Sum. strategy usage*, the total occurrence of a strategy in dialogue; and (ii) *Avg. strategy usage*, the total occurrence of a strategy divided by the total number of utterances in dialogue.
- **Politeness-level features**, which calculated average politeness levels (identified by `PLI`) of the utterances only by a tutor, only by a student, and by both of them.

Overall, the feature set included: (i) 16 holistic features; (ii) 126 strategy-level features; iii) 3 politeness-level features. We denoted these features as *Politeness features*. We note that a student's prior progress on a problem might be another useful predictor to enhance the model's performance. As we aimed to build a model for real-time prediction (i.e., the model's input features should be automatically collected from the real-time data), we did not add students' prior progress into the engineered feature set. Subsequently, we plan to build a machine learning model to automatically identify the students' prior progress in a session and incorporate it as an additional feature for students' performance prediction.

The current study mainly focused on the politeness features extracted from tutor-student utterances. However, it should be noted that students' problem-solving performance might not only be affected by the polite expression but also by other features including student sentiment (Jiménez, Juárez-Ramírez, Castillo and Armenta, 2018), student effort (Goodman, Jaffer, Keresztesi, Mamdani, Mokgatle, Musariri, Pires and Schlechter, 2011), and task difficulty levels (Blanchard, Baker, Ocumpaugh and Brawner, 2014). Therefore, we decided to incorporate those relevant features in our final model. As suggested by (Lin et al., 2022b), we engineered additional 543 non-politeness features to enhance our predictor set. Due to the word limit, the description of these features was detailed in the electronic appendix that can be accessed by following this link `http://bitly.ws/uRpZ`.

[4] Note that one utterance can contain multiple types of politeness strategies.

3.4.3. Ablation Analysis

For RQ3, we conducted an ablation analysis to gain a better understanding of the predictive power of politeness features on model performance. An ablation analysis is a widely-used method to measure the contribution made by different features on the model performance (Lin, Pan, Lee and Oviatt, 2019). The contribution of a specific feature was assessed by measuring the difference between the model performance trained on the feature set with that specific feature included vs excluded. In other words, the ablation analysis assessed the feature importance by removing a specific feature of the model, and then measuring how removing that feature affected the model prediction performance. To conduct the ablation analysis, we performed the following steps: 1) assume we have a dataset containing m features; 2) use all of the m features to train the machine learning model and calculate the model performance (e.g., as classification accuracy and F1-score) on the testing set; 3) remove each of the m features separately from the training data and use the $m-1$ features to train the model again and calculate the model performance on the testing set; and 4) compare the feature importance between features in the full model (i.e., from the model using m features) and the features in the model using $m-1$ features.

3.5. Study Setup

Model Training for Predicting Student Performance We randomly split the annotated dataset into *training*, *validation*, and *testing*, per the ratio of 80%:10%:10% (Goodfellow, Bengio and Courville, 2016), respectively. To evaluate the predictive performance of GTB, we used Random Forest (RF) model as a baseline model. The GTB and RF models were trained using the Python package `scikit-learn`[5]. The parameters in both models (e.g., number of trees and max depth) were optimised by using grid search on the validation dataset. The models' performance was finally evaluated on the testing dataset.

Evaluation Metrics To evaluate the models' performance, we utilised four representative metrics, i.e., classification accuracy, F1 score, Area Under the Curve (AUC), and Cohen's kappa coefficient (Cohen's κ).

Table 3
The top 10 most frequent politeness strategies in our dataset. **Without PP** and **With PP** stand for *Without Prior Progress* and *With Prior Progress* students, respectively. Strategies marked with ⋆ are likely to incur a sense of direct expression, while the others are polite. We employed Mann-Whitney tests to examine the difference of politeness levels between **Without PP** and **With PP** for each session category (significant results were marked in bold font, $p < 0.001$) and the difference among any two of the session categories (i.e., Gap-clarified, Gap-explained, and Gap-bridged) where students had the same status of prior progress (significant results were marked with the symbol, i.e., †, ♣, and ◇[6]in a row, $p < 0.001$).

Strategy	Role	All	Gap-clarified		Gap-explained		Gap-bridged	
			Without PP	With PP	Without PP	With PP	Without PP	With PP
HASPOSITIVE	T	23.23%	◇♣**22.33%**	◇♣**25.91%**	†◇**20.84%**	†◇**22.25%**	†♣**23.03%**	†♣**24.29%**
2nd_person	T	20.72%	◇♣**28.38%**	◇♣**25.92%**	†◇21.17%	†◇21.83%	†♣**18.21%**	†♣**19.06%**
1st_person_pl	T	12.68%	◇♣**11.41%**	◇♣**7.04%**	†◇**15.86%**	†◇**13.80%**	†♣**13.79%**	†♣**11.77%**
HASNEGATIVE *	T	10.73%	◇♣**18.38%**	◇♣**13.55%**	†◇**11.98%**	†◇**10.86%**	†♣**9.23%**	†♣**8.65%**
Direct_start *	T	6.84%	◇♣4.99%	◇♣4.53%	◇7.35%	◇7.40%	♣**7.52%**	♣**6.96%**
Direct_question *	T	5.76%	**6.68%**	◇♣**5.43%**	6.00%	◇6.02%	**5.72%**	♣**5.45%**
HASHEDGE	T	5.50%	**6.20%**	◇♣**5.15%**	†**5.95%**	†◇**6.56%**	†5.13%	†♣5.26%
1st_person	S	5.33%	**5.33%**	◇♣**7.47%**	**4.41%**	◇**6.01%**	**4.45%**	♣**5.82%**
HASPOSITIVE	S	5.24%	♣**4.28%**	◇♣**7.93%**	†**3.50%**	†◇**4.59%**	†♣**4.74%**	†♣**6.21%**
1st_person_start	T	5.01%	◇♣**9.57%**	◇♣**8.65%**	†◇5.55%	†◇5.61%	†♣**3.42%**	†♣**3.86%**

4. Results

4.1. Results on RQ1

The top 10 frequently used politeness strategies were given in Table 3. We calculated the percentage of the politeness strategies for each tutorial session and then averaged the percentage over the entire dataset, per session category

[5]https://scikit-learn.org/

[6]We used three symbols (e.g., †, ♣, and ◇) to mark the statistical differences. † marked the significant difference between Gap-explained and Gap-bridged, ♣ marked the significant difference between Gap-clarified and Gap-bridged, ◇ marked the significant difference between Gap-clarified and Gap-explained

(i.e., Gap-clarified, Gap-explained, and Gap-bridged). As one utterance could contain multiple strategies, the gross fraction of utterances associated with these strategies was larger than 100%. It should be noted that 3 out of 10 strategies in Table 3 were direct: *HasNegative*, *Direct_Start*, and *Direct_Question*. The remaining 7 strategies we identified were polite.

Considering students' prior progress, we reported several findings in Table 3. Firstly, the fraction of the HASPOSITIVE strategies tutors delivered to *With Prior Progress* students were generally higher than the fraction of the same strategies delivered to *Without Prior Progress* students across the three session categories. In contrast, the fraction of the HASNEGATIVE strategies tutors delivered to *With Prior Progress* students were generally lower than the same set of strategies delivered to the *Without Prior Progress* students across the three session categories. Secondly, the fraction of the strategy *1st_person_pl* by tutors delivered to the *With Prior Progress* students was generally lower than the strategy delivered to the *Without Prior Progress* students, across the three session categories. Thirdly, we noted that the proportion of the *HasNegative* strategy delivered to the students with and without prior progress in the Gap-clarified sessions was higher than that of the other two session types.

Table 4
The average politeness score of tutors and students across dialogues in various session categories. **Without PP** and **With PP** stand for *Without Prior Progress* and *With Prior Progress*, respectively. We employed Mann-Whitney tests to examine the difference of politeness levels between **Without PP** and **With PP** for each session category (significant results were marked in bold font, $p < 0.001$) and the difference among any two of the session categories (i.e., Gap-clarified, Gap-explained, and Gap-bridged) where students had the same status of prior progress (significant results were marked with the symbol, i.e., †, ♣, and ◇ in a row, $p < 0.001$).

	All	Gap-clarified		Gap-explained		Gap-bridged	
		Without PP	With PP	Without PP	With PP	Without PP	With PP
1. Avg. Politeness Score (Tutor & Student)	0.62 ± 0.06	◇♣**0.64 ± 0.12**	◇♣**0.67 ± 0.11**	†♣0.60 ± 0.09	♣0.60 ± 0.09	†◇**0.59 ± 0.08**	◇**0.61 ± 0.08**
2. Avg. Politeness Score (Tutor)	0.66 ± 0.08	◇♣**0.71 ± 0.10**	◇♣**0.72 ± 0.09**	†♣0.65 ± 0.08	♣0.65 ± 0.08	†◇0.64 ± 0.07	◇0.65 ± 0.07
3. Avg. Politeness Score (Student)	0.56 ± 0.07	◇♣**0.58 ± 0.10**	◇♣**0.62 ± 0.11**	◇0.54 ± 0.07	†◇ 0.55 ± 0.07	♣ **0.55 ± 0.06**	†♣**0.56 ± 0.07**

The results shown in Table 3 indicate that polite strategies were more frequently used than the direct ones in human-human dialogue-based tutoring, which can be further confirmed by the results shown in Table 4. Table 4 presents the politeness levels of the tutors and students from all tutorial dialogues. The politeness levels in Table 4 were calculated by computing the mean of the average politeness levels across all dialogue sessions. When considering the utterances made by the tutors and students from all dialogues (Row 1, Table 4), we found that the politeness level was 0.62, which implied that the tutoring dialogue was relatively polite. When comparing the politeness levels from tutors among three various dialogue sessions, we observed that the Gap-clarified sessions (i.e., students achieved the least progress) were the most polite ones (0.72 in the group *With Prior Progress* and 0.71 in *Without Prior Progress*).

To better understand the use of politeness in different tutoring sessions, it is worth further scrutinising the role of politeness expressed by tutors and students at the beginning and end of the dialogues, as these are the two stages where tutors and students of different dialogue categories likely differed from each other regarding the politeness levels. Figure 1 presents the changes of politeness levels at the beginning and the end of sessions by tutors and students. We calculated the politeness levels by considering the first and last 15 utterances from tutors and students, and, in each plot, we distinguished between the students with and without prior progress. Figure 1 (a) shows that all types of dialogue categories from ***tutors*** had similar politeness levels in the first 15 utterances. As the dialogue progressed, the level of politeness was decreasing. By analysing the frequent strategies adopted by tutors between 8th and 15th utterances, we found that tutors in the Gap-clarified group used notably less direct strategies of *Direct Start* and *Direct Question* than the tutors in the Gap-explained and Gap-bridged groups. When analysing the end of the dialogues (Figure 1 (b)), we found that tutors and students were gradually increasing politeness in their communication. It is worth noting that the politeness level of Gap-bridged dialogues from ***students*** in both *With Prior Progress* and *Without Prior Progress* increased between -7th (i.e., the last seventh utterances) and -1st utterances, which distinguished the difference between various dialogue categories. By analysing the frequent strategies adopted by tutors and students between -7th and -1st utterances, we found that the tutors in the Gap-clarified dialogue group used more the *HASNEGATIVE* strategy than their colleagues in the Gap-explained and Gap-bridged groups. This is probably because, in the Gap-clarified dialogues, the tutors showed more negative comments on students' works (e.g., *"Why are you unsure it?"*, *"No, it's wrong."*), and thus a higher usage of *HASNEGATIVE*. In contrast, the tutors in the Gap-bridged dialogues used more the *HASPOSITIVE* strategy than the tutors in the other two categories of dialogues; whereas, the students in the Gap-

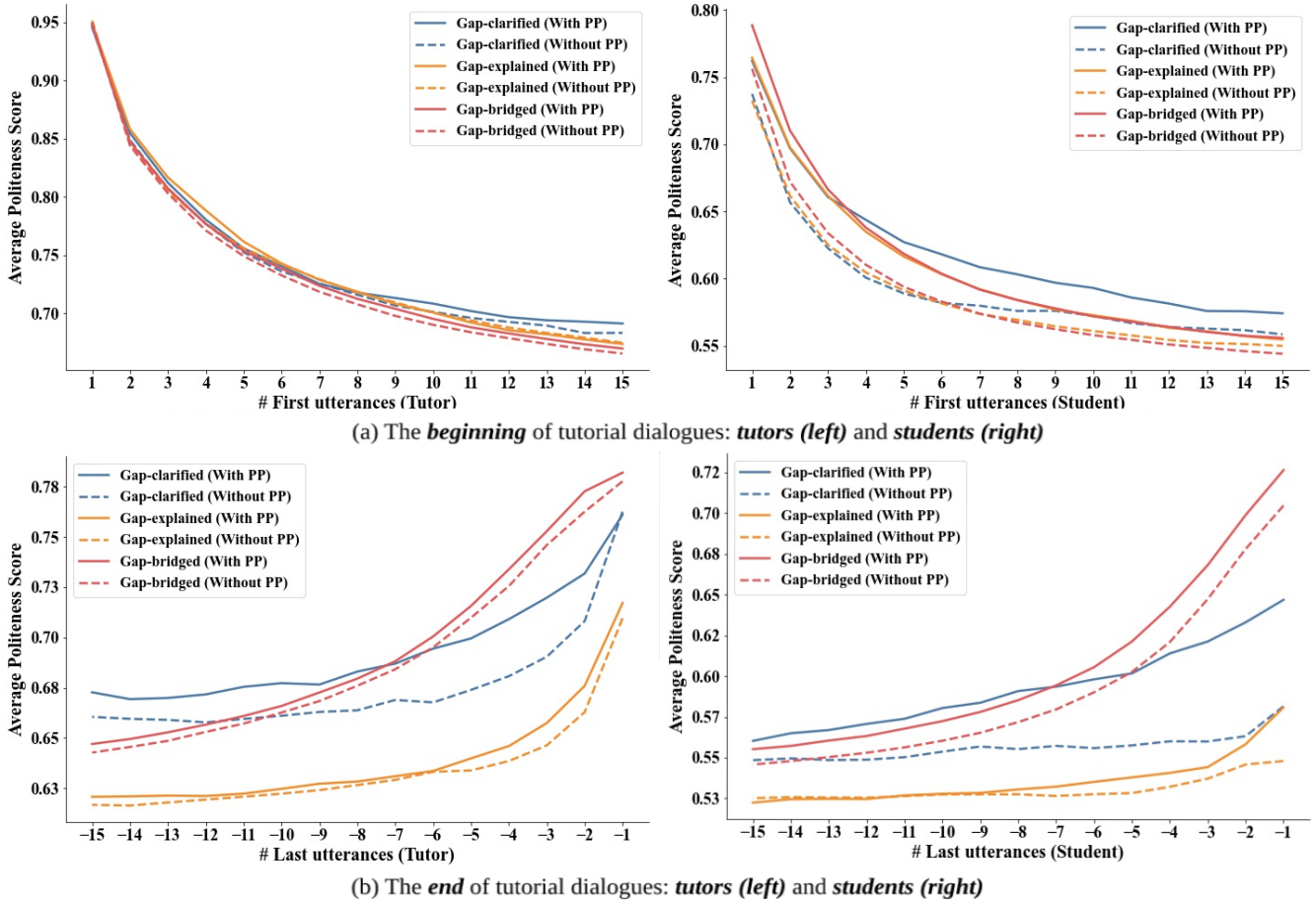


Figure 1: The average politeness scores of tutors and students at the start and end of tutoring dialogue. **With PP** and **Without PP** denote *With Prior Progress* and *Without Prior Progress*, respectively.

bridged dialogues used more polite strategies of *Gratitude* and *HASPOSITIVE* than in the other two category dialogues.

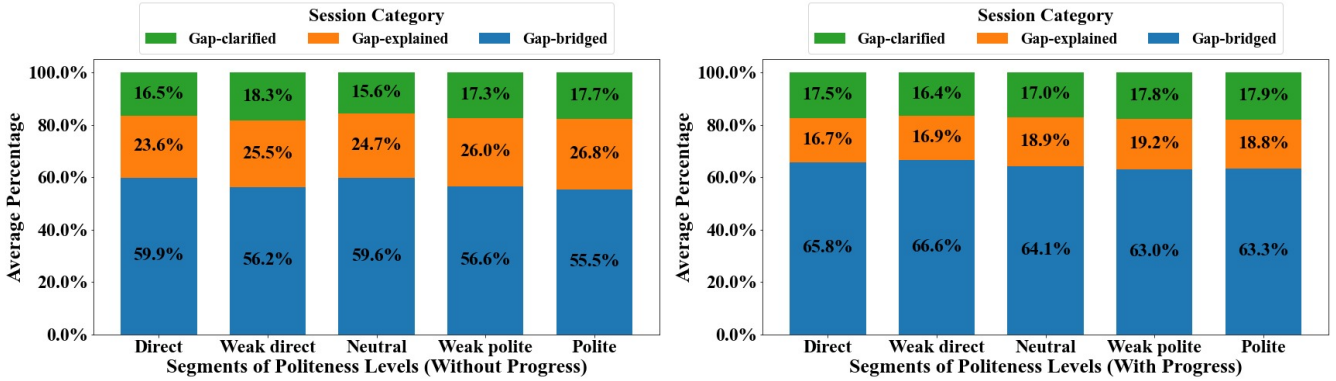


Figure 2: The distribution of dialogue categories in different politeness groups by considering **first 5 utterances** from tutors. Notably, the students *Without Progress* are on the **left** and *With Progress* on the **right**.

4.2. Results on RQ2

To answer RQ2, we investigated the correlation between the politeness of tutors' expressions and students' performance under the condition of students' prior progress. We categorised the average politeness levels from the tutors' utterances into five politeness groups (i.e., *Direct*, *Weak direct*, *Neutral*, *Weak polite*, and *Polite*, described in Sec. 3.3). Then, as expressing politely at the start of the tutoring is important for tutors to build rapport with students (Bleistein

and Lewis, 2014), we aimed to answer RQ2 by exploring the politeness of tutors' utterances from the start of the tutoring. As a tutoring process can be divided into two essential modes: *Non-instructional mode* and *Instructional mode* (Morrison, Nye, Rus, Snyder, Boller and Miller, 2015), our study mainly focused on analysing the correlation between tutors' politeness and students' performance based on these two modes.

In the *Non-instructional mode*, tutors and students had the non-instructional communication to build rapport (e.g., "*How are you?*"). Building upon the work by (Morrison et al., 2015; Schegloff and Sacks, 1973), tutors commonly use several utterances at the beginning of the tutorial session to finish the non-instructional communication (e.g., Greeting and Self-introduction). To gain an understanding of politeness in non-instructional communication, we investigated the first 5 utterances from tutors and the results were shown in Figure 2. When checking the tutoring session where students had no prior progress (i.e., *Without Prior Progress*), the fraction of Gap-bridged sessions in the *Direct* group were slightly higher than that in the *Weak polite* and *Polite* groups; whereas, regarding the tutoring session where students had a certain amount of prior progress (i.e., *With Prior Progress*), the result showed that the fraction of Gap-bridged sessions in the *Direct*, *Weak direct* and *Neutral* groups were slightly higher than that in the *Polite* and *Weak polite* groups. The results indicated that the polite expression by tutors might not have a direct impact on students' final problem-solving performance.

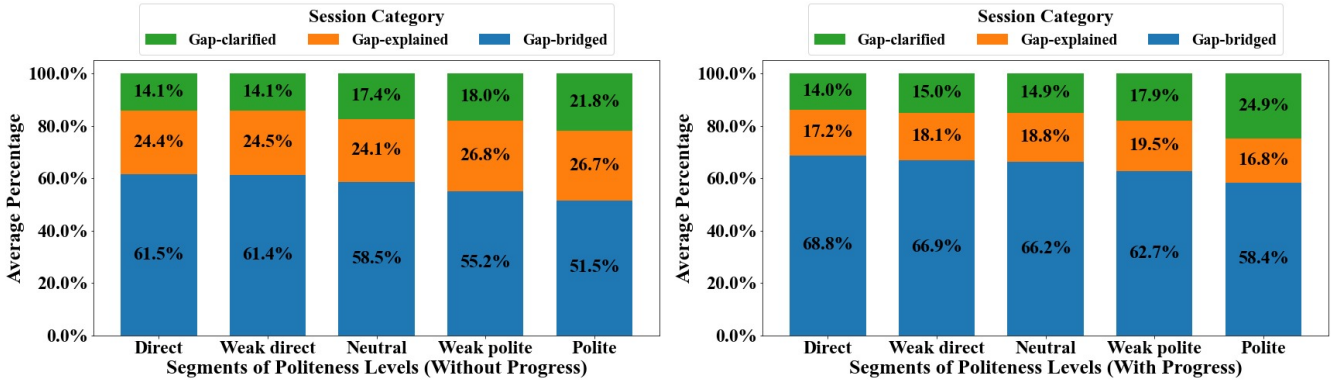


Figure 3: The distribution of dialogue categories in different politeness groups by considering the **first 10 utterances** by tutors. The students *Without Progress* are on the **left** and *With Progress* on the **right**.

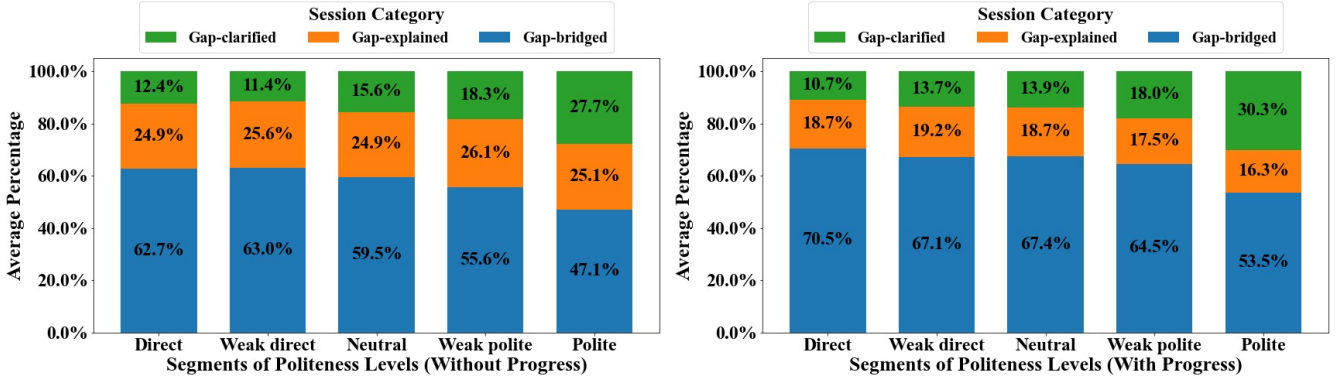


Figure 4: The distribution of dialogue categories in different politeness groups by considering **first 15 utterances** from tutors. The students *Without Progress* are on the **left** and *With Progress* on the **right**.

When diving into the *Instructional mode*, tutors commonly start with some evaluation activities about the students' understanding of the problem and the difficulty of the problem (Morrison et al., 2015), which can help tutors understand students' prior progress and provide personalised instructional methods. Additionally, tutors complete these activities by several utterances followed by non-instructional communication (Morrison et al., 2015). To ensure our analysis covered most of the tutors' utterances related to the evaluation activities, we investigated the politeness used in the evaluation process by the first 10 utterances from tutors (shown in Figure 3). When scrutinising the tutoring session where students had no prior progress (i,e., *Without Prior Progress*), the fraction of the Gap-bridged sessions in the

Direct group was much higher than that in the *Polite* group; the same finding was also observed for the tutoring session where students had a certain amount of prior progress (i.e., *With Prior Progress*). This means that when diving into the instructional activities, the tutor would start using more direct expressions to guide students who had or had no prior progress.

After the early evaluation in the *Instructional mode*, tutors spend the most time on the scaffolding process (Morrison et al., 2015), which is considered one of the most effective methods in supporting students (VanLehn, 2011). To ensure that our analysis covered most of the tutors' utterances related to scaffolding activities and some short dialogue sessions (e.g., Gap-clarified) where the tutors delivered few utterances, we investigated the politeness used in the scaffolding process in the first 15 utterances by tutors (shown in Figure 4). The result demonstrated that the fraction of the Gap-bridged sessions was higher in the *Direct* group than the *Polite* group in both students had no prior progress *Without Prior Progress* and had prior progress *With Prior Progress*. This result indicates that the tutors tended to communicate in a direct manner during the scaffolding process.

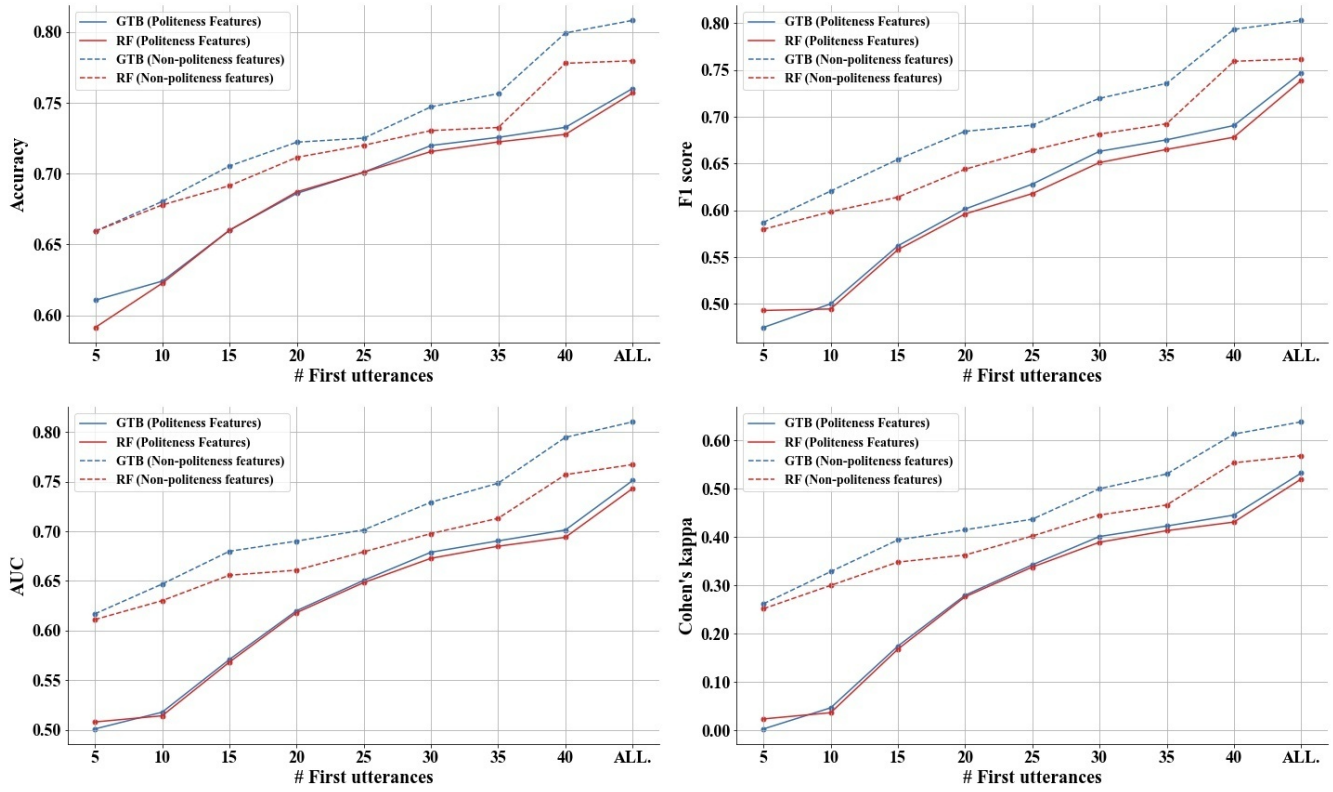


Figure 5: The performance of GTB model (lines coloured in blue) and RF model (lines coloured in red) in predicting student problem-solving performance.

4.3. Results on RQ3

To answer RQ3, we investigated the capability of politeness features (i.e., politeness strategies and politeness levels) in predicting students' performance. We used the first N utterances (where N ϵ [5, 10, 15, 25, 30, 35, 40, *All*.] utterances) to extract the politeness features introduced in Sec. 3.4. Then, we used those features as input for training the GTB and RF models and predicting students' performance. The prediction performance of GTB and RF models was measured using four metrics (i.e., Classification accuracy, F1 score, AUC, and Cohen's *kappa*) as shown in Figure 5. By scrutinising the solid line in Figure 5, we found that the GTB outperformed RF model across all the prediction performance metrics and the prediction performance of GTB presented a generally increasing trend. When including all the utterances (i.e., N = ALL), we found that the GTB model reached the performance of 0.760 (Accuracy), 0.747 (F1 score), 0.751 (AUC score), and 0.533 (Cohen's κ). However, in practice, the earlier a student's performance can be predicted, the more tailored support a tutor can provide to help students (Moreno-Marcos, Pong, Munoz-Merino and Kloos, 2020; Gutjahr, Menon and Nedungadi, 2017). In Figure 5 (the solid blue line), we observed that the best

trade-off between early tutoring intervention and prediction accuracy was likely to be after 10 utterances where the Accuracy, F1 score, AUC and Cohen's *kappa* were 0.624, 0.500, 0.047, and 0.518, respectively. Then, to examine the power of politeness features for predicting students' performance, we further used the non-politeness features proposed in (Lin et al., 2022b) to train both models in comparison. The results of using non-politeness features (the dash line as shown in Figure 5) indicated that the GTB model achieved better prediction performance by non-politeness features than that by politeness features when using all possible values of N utterances from tutorial dialogues. This indicates that, despite being useful to identify different tutorial session categories, politeness features were insufficient to predict students' performance in online dialogue-based tutoring.

Table 5
The ablation test results of the GTB model when only considering the first 10 utterances in a dialogue.

Feature name	Accuracy	F1 score	Cohen's *kappa*	AUC
1. Politeness Features	0.624	0.500	0.047	0.518
2. w/o Holistic Features	0.623 (-0.23%)	0.497 (-0.55%)	0.042 (-10.51%)	0.516 (-0.31%)
3. w/o Strategy-level Features	**0.620 (-0.63%)**	**0.491 (-1.81%)**	**0.031 (-33.78%)**	**0.512 (-1.11%)**
4. w/o Politeness-level features	0.625 (+0.11%)	0.504 (+0.84%)	0.054 (+15.05%)	0.520 (+0.52%)

To deepen our understanding of the predictive power of politeness features, we implemented the ablation test (described in Sec. 3.4.3) to measure the respective contributions made by the specific politeness features. In Table 5, we presented the feature importance made by each group of features described in Sec. 3.4. Here, Row 1 in Table 5 presents the performance of the GTB model taking all politeness features as the input, and Rows 2-4 present the performance of the GTB model without taking a specific type of features as input. The rates within brackets were calculated by comparing the model performance without taking a specific type of features as input to the model performance trained on all politeness features (i.e., Row 1). The results with the maximum performance decrease are in bold. As we observed that using first 10 utterances for model prediction might be the most effective trade-off point between early intervention and prediction accuracy, we decided to present the results of the ablation test by using the first 10 utterances from the tutoring dialogues (shown in Table 5). Though the group of **Strategy-level features** contributed most to the prediction performance at the first 10 utterances, the use of politeness might not be a strong indicator to differentiate the student performance at the early stage the tutoring dialogues.

5. Discussion and Conclusion

The development of dialogue-based ITSs with the consideration of linguistic politeness is important in motivating students. In this paper, we analysed the characteristics of the politeness features (i.e., politeness strategies and levels) of utterances from tutors and students in online human-human tutoring. The results further motivated us to quantify the correlation between the tutors' politeness and students' performance. Our study displayed some insights to support developing existing dialogue-based ITS and also contributed to the tutoring practice with the following main findings:

- In the most effective sessions (i.e., students successfully solved the problem), the tutors tended to use more the polite strategy (`HASPOSITIVE`) to guide the students with prior progress than those without; on the other hand, the tutors tended to use the less polite strategy (`1st_person_pl`) and the direct strategies (`Direct_start` and `Direct_question`) to guide the students with prior progress than those without.
- Compared to the most ineffective tutoring sessions (i.e., where students had no progress on the problem after tutoring sessions), the tutors in the most effective tutoring sessions tended to use less the direct strategy (`HasNegative`) but more the direct strategy (`Direct start`) to guide the students with prior progress.
- In the most effective tutoring sessions, the tutors were likely to express politely at the beginning and gradually used more direct expressions to guide the students as a tutorial session progressed. The students in the most effective sessions were slightly more polite at the beginning and end of tutorial sessions compared with the students in the most ineffective tutoring sessions.
- By investigating the politeness in the instructional and non-instructional communication, our study found that in the non-instructional communication, there was no evident correlation between the politeness of the tutors'

expressions and the students' performance. In comparison, in instructional communication, the students would benefit more from the tutors' direct expressions than polite expressions.

- Our results demonstrated the politeness features (i.e., politeness strategies and levels) were insufficient to predict the students' performance and should be incorporated with other non-politeness features (e.g., the sentiment of the utterances).

5.1. Implications

Our study advanced the existing research on politeness in an educational context by examining the large scale tutor-student dialogue dataset in the online learning environment. Firstly, we found that human tutors used more polite strategies than direct ones in tutoring. In particular, the most frequently used polite strategies from tutors were `HASPOSITIVE`, `2nd_person` and `1st_person_pl`. The strategy `HASPOSITIVE` was used to characterise the utterances containing the positive sentiment lexicons such as praise and encouragement. Our study revealed that in the most effective tutorial sessions (i.e., Gap-bridged), the tutors tended to employ more the `HASPOSITIVE` strategy for the students with prior progress than without prior progress. It was not a surprising result as tutors commonly gave praise to the students who made certain efforts on their problem (Maclellan, 2005), which could maintain or enhance students' positive face for getting approval. Thus, using more of the `HASPOSITIVE` strategy to express politeness for students with prior progress is recommended for tutoring practice and design of dialogue-based ITS, which might encourage students to take a further effort.

Secondly, we found that the tutors in the effective tutoring sessions (i.e., Gap-bridged) had lower usage of the polite strategy `2nd_person` than the tutors in the ineffective tutoring sessions (i.e., Gap-clarified and Gap-explained). Additionally, in the effective tutoring sessions, the students without prior progress received less the strategy `2nd_person` from tutors compared to the students with prior progress. The polite strategy `2nd_person` was related to the use of the word "*you*" but not placed at the first word of the utterance (Danescu-Niculescu-Mizil et al., 2013). This polite strategy can be used to rephrase a direct expression (e.g., "*No, it is wrong*" and "*You are wrong*") to the polite form (e.g, "*It seems that you made a mistake there*") (Brown and Levinson, 1987). An explanation for more usage of this polite strategy in ineffective sessions could be that the misuse of polite strategy might make the instruction unclear (Jucks et al., 2016), which might further hinder students' understanding. For example, by using the strategy `2nd_person`, the hints with the direct sense (e.g., "*You should calculate the value of X first*") could be rephrased in a polite form (e.g., "*Should you calculate the value of X first?*"). In the example, the direct hints presented more explicit instruction than the polite ones. Therefore, the polite strategy `2nd_person` could help tutors to mitigate the sense of direct but this strategy should be carefully used in delivering instructional hints, especially for the students without prior progress.

Regarding the use of the `1st_person_pl` strategy, we found that in the effective tutoring sessions (i.e., Gap-bridged), the tutors tended to employ more a polite strategy `1st_person_pl` for the students without prior progress than for their counterparts with prior progress. The polite strategy `1st_person_pl` expressed the sense of working together towards the joint goal (Wang et al., 2008), which might be helpful to improve students' learning experience in the tutoring process. As there was no direct evidence demonstrating the helpfulness of using `1st_person_pl`, we suggest that a randomised control trial should be conducted in the future to examine the effectiveness of using this strategy on students' learning experience.

When scrutinising the changes of politeness levels as the tutorial progressed (e.g., shown in Figure 1), we found that the effective tutoring sessions (i.e., Gap-bridged) gradually used more direct expressions, which is in line with the results in Lin et al. (2020). Specifically, we observed that tutors in effective sessions used more the `Direct_start` strategy than that was the case in one type of ineffective sessions (i.e., Gap-clarified). The reason for this result could be that the tutors in the effective sessions used more direct and explicit hints than they did in the ineffective sessions. The direct hints expressed the sense of requesting the students to do something, which can be categorised into a direct strategy `Direct_start`. We also observed that the tutors used more the `Direct_start` strategy with the student without prior progress than they did with those with prior progress. Therefore, these results might indicate that to better assist the students without prior progress, tutors might consider using more direct expressions to explicitly guide students.

Inspired by the works (Morrison et al., 2015; Schegloff and Sacks, 1973), our study investigated the relationship between the politeness of tutors' expression and students' performance by differentiating the tutoring dialogues into instructional and non-instructional communication. In the non-instructional communication (e.g., greeting), we did not find the evident correlation between the politeness from tutors' expressions and students' performance. This was not a surprising result as the tutors in our dataset were trained on how to communicate properly with students and these tutors

might express at a similar politeness level in the non-instructional communication. When scrutinising the politeness of tutors' expression in the instructional communication, we found that the percentage of effective sessions was higher in the groups guided by direct tutors than the polite ones. This result indicated that the students' performance could benefit from tutors' direct expression in the instructional communication. Our results on tutors' politeness from non-instructional and instructional communication might not support the finding from some existing works (Wang et al., 2005; Wang and Johnson, 2008; Wang et al., 2008; Gupta et al., 2007; Mikheeva et al., 2019; McLaren et al., 2011b), i.e., tutors' polite expressions positively correlate with student performance. An explanation to the difference between our result and existing works might be that the direct effect of politeness was weak on students' performance in the real-world learning environment compared to the lab study (McLaren et al., 2007, 2011a).

Finally, our prediction results showed that the politeness features alone might be insufficient to predict the student performance compared with the non-politeness features. However, we suggest that the politeness features should be incorporated in the future work of student performance prediction and the development of dialogue-based ITS. Our study demonstrated that the use of some politeness strategies, as discussed above, was different regarding the students' prior progress and also students' performance. By incorporating the politeness features in the dialogue-based ITS, we expect that more personalised interventions (e.g., use the strategy `1st_person_pl` when students did not make any progress before joining the tutoring) can be employed to better support students.

5.2. Limitations and Future Work

Though our study produced some interesting findings, several limitations in this study should be noted and highlight a need for future research revolving around this topic. *Firstly*, our study did not find an evident positive correlation between politeness and students' performance, but it remains unknown whether politeness has a positive effect on students' performance in real-world human-human dialogue-based online tutoring. We suggested that a randomised experiment is required to examine the effect of politeness in human-human dialogue-based online tutoring so as to further guide the development of dialogue-based ITSs. *Secondly*, our analysis was conducted on a dataset that was collected in the USA. Future research needs to be conducted with students from other cultural backgrounds to examine the generalisability of our results. Based on the existing literature, a positive effect of politeness was mostly found with the students in the US (Wang et al., 2005, 2008; McLaren et al., 2011b) and Germany (Schneider et al., 2015; Mikheeva et al., 2019). However, Brom, Hannemann, Starkova, Bromová and Děchtěrenko (2017) found that students from Czech preferred the tutors guiding them with direct expressions in instruction. Therefore, the effectiveness of politeness in supporting students might vary across cultures. *Thirdly*, the consistency between human annotation of politeness levels in our study and the PLI results was not very high. It should be noted that the PLI method was originally trained on the human-annotated dataset provided by Danescu-Niculescu-Mizil et al. (2013). In their work, Danescu-Niculescu-Mizil et al. (2013) claimed that the boundary between somewhat polite and impolite was blurry in the annotation process. The reason could be the perception of politeness is subjective which might be varied from the individual (e.g., cultural background, gender, and personality) (Brown and Levinson, 1987). Thus, the agreement of the sanity check on the PLI tool was not such promising. *Fourthly*, the current study did not analyse tutoring sessions across different subjects (i.e., mathematics, physics, and chemistry). As 92% of the tutoring sessions in our dataset were tutoring mathematics, it is worthwhile to collect more tutoring sessions in chemistry, physics, and subjects to estimate the extent to which the findings of the current study can be generalised. Additionally, it should be noted that math tutoring might be a unique domain as tutors often use strictly objective instruction, which might be a factor influencing the use of direct or polite expressions. *Lastly*, though the current study demonstrated the effectiveness of the GTB model in predicting students' problem-solving performance, future research should investigate the efficacy of other predictive models (e.g., deep neural networks).

Furthermore, an extension of future work would be to use a qualitative method (e.g., interview) to investigate the perspectives of politeness in the online tutoring dialogues from tutors and students. Inspired by the results related to RQ2, we could design a mixed-method study where tutors' perspectives of expressing and students' perspectives of receiving different levels of politeness (e.g., polite, neutral and direct) will be collected in different tutoring modes (e.g., instructional and non-instructional communication) by interview. Finally, we will collect the students' problem-solving performance from the online dialogue tutoring sessions and perspectives on the different levels of politeness from tutors and students, which will be further analysed to compare the results with the quantitative analysis presented in the current study.

6. Conflict of Interest

There is no potential conflict of interest in this study.

7. Ethics Statement

The ethics approval was obtained prior to the analysis of the archived data.

8. Data Availability Statement

Due to privacy issues, the dialogue data cannot be made publicly available.

References

Albacete, P., Jordan, P., Lusetich, D., Chounta, I.A., Katz, S., McLaren, B.M., 2018. Providing proactive scaffolding during tutorial dialogue using guidance from student model predictions, in: International Conference on Artificial Intelligence in Education, Springer. pp. 20–25.

Alkhatlan, A., Kalita, J., 2019. Intelligent tutoring systems: A comprehensive historical survey with recent developments. International Journal of Computer Applications 181, 1–20. doi:10.5120/ijca2019918451.

Almasri, A., Ahmed, A., Almasri, N., Abu Sultan, Y.S., Mahmoud, A.Y., Zaqout, I.S., Akkila, A.N., Abu-Naser, S.S., 2019. Intelligent tutoring systems survey for the period 2000-2018 .

Blanchard, N., Baker, R., Ocumpaugh, J., Brawner, K., 2014. I feel your pain : A selective review of affect-sensitive instructional strategies.

Bleistein, T., Lewis, M., 2014. One-on-one language teaching and learning: Theory and practice. Springer.

Bloom, B.S., 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational researcher 13, 4–16.

Boyer, K.E., Phillips, R., Wallis, M., Vouk, M., Lester, J., 2008. Balancing cognitive and motivational scaffolding in tutorial dialogue, in: International conference on intelligent tutoring systems, Springer. pp. 239–249.

Brom, C., Hannemann, T., Starkova, T., Bromová, E., Děchtěrenko, F., 2017. The role of cultural background in the personalization principle: Five experiments with czech learners. Computers & Education 112, 37–68.

Brown, P., Levinson, S., 1987. Politeness: some universals in language usage, Cambridge University Press, Cambridge, UK.

Brummernhenrich, B., Jucks, R., 2013. Managing face threats and instructions in online tutoring. Journal of Educational Psychology 105, 341.

Brummernhenrich, B., Jucks, R., 2016. "he shouldn't have put it that way!" how face threats and mitigation strategies affect person perception in online tutoring. Communication Education 65, 290–306.

Chen, G., Lang, D., Ferreira, R., Gasevic, D., 2019. Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues, in: EDM.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., Potts, C., 2013. A computational approach to politeness with application to social factors, in: Proceedings of the 51st ACL, Association for Computational Linguistics, Sofia, Bulgaria. pp. 250–259.

D'Mello, S., Olney, A., Williams, C., Hays, P., 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. International Journal of human-computer studies 70, 377–398.

Economidou-Kogetsidis, M., 2015. Teaching email politeness in the efl/esl classroom. Elt Journal 69, 415–424.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research 15, 3133–3181.

Goffman, E., 1967. On face-work. Interaction ritual , 5–45.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Goodman, S., Jaffer, T., Keresztesi, M., Mamdani, F., Mokgatle, D., Musariri, M., Pires, J., Schlechter, A., 2011. An investigation of the relationship between students' motivation and academic performance as mediated by effort. South African Journal of Psychology 41, 373–385.

Group, L.P.R., et al., 2011. Discursive approaches to politeness. volume 8. Walter de Gruyter.

Gupta, S., Walker, M.A., Romano, D.M., 2007. How rude are you?: Evaluating politeness and affect in interaction, in: International Conference on Affective Computing and Intelligent Interaction, Springer. pp. 203–217.

Gutjahr, G., Menon, K., Nedungadi, P., 2017. Using an intelligent tutoring system to predict mathematics and english assessments, in: 2017 5th IEEE International Conference on MOOCs, Innovation and Technology in Education (MITE), IEEE. pp. 135–140.

Hasan, M.A., Noor, N.F.M., Rahman, S.S.A., Rahman, M.M., 2020. The transition from intelligent to affective tutoring system: A review and open issues. IEEE Access .

Jiménez, S., Juárez-Ramírez, R., Castillo, V.H., Armenta, J.J.T., 2018. Affective Feedback in Intelligent Tutoring Systems: A Practical Approach. Springer.

Johnson, W.L., Wang, N., 2010. The role of politeness in interactive educational software for language tutoring. Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology. Auerbach Publications , 91–113.

Jucks, R., Päuler, L., Brummernhenrich, B., 2016. "i need to be explicit: You're wrong": Impact of face threats on social evaluations in online instructional communication. Interacting with Computers 28, 73–84.

Korb, K., 2013. Conducting educational research: calculating descriptive statistics.

Lin, J., Lang, D., Xie, H., Gašević, D., Chen, G., 2020. Investigating the role of politeness in human-human online tutoring, in: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (Eds.), Artificial Intelligence in Education, Springer International Publishing, Cham. pp. 174–179.

Lin, J., Pan, S., Lee, C.S., Oviatt, S., 2019. An explainable deep fusion network for affect recognition using physiological signals, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA. p. 2069–2072. URL: https://doi.org/10.1145/3357384.3358160, doi:10.1145/3357384.3358160.

Lin, J., Rakovic, M., Lang, D., Gasevic, D., Chen, G., 2022a. Exploring the politeness of instructional strategies from human-human online tutoring dialogues, in: LAK22: 12th International Learning Analytics and Knowledge Conference, Association for Computing Machinery, New York, NY, USA. p. 282–293. URL: https://doi.org/10.1145/3506860.3506904, doi:10.1145/3506860.3506904.

Lin, J., Singh, S., Sha, L., Tan, W., Lang, D., Gašević, D., Chen, G., 2022b. Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues. Future Generation Computer Systems 127, 194–207.

Liu, B., Hu, M., Cheng, J., 2005. Opinion observer: analyzing and comparing opinions on the web, in: Proceedings of the 14th international conference on World Wide Web, pp. 342–351.

Maclellan, E., 2005. Academic achievement: The role of praise in motivating students. Active learning in higher education 6, 194–206.

Maharjan, N., Rus, V., 2018. A tutorial markov analysis of effective human tutorial sessions, in: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pp. 30–34.

Maharjan, N., Rus, V., Gautam, D., 2018. Discovering effective tutorial strategies in human tutorial sessions, in: The Thirty-First International Flairs Conference.

McLaren, B.M., DeLeeuw, K.E., Mayer, R.E., 2011a. Polite web-based intelligent tutors: Can they improve learning in classrooms? Computers & Education 56, 574–584.

McLaren, B.M., DeLeeuw, K.E., Mayer, R.E., 2011b. A politeness effect in learning with web-based intelligent tutors. International Journal of Human-Computer Studies 69, 70–79.

McLaren, B.M., Lim, S.J., Yaron, D., Koedinger, K.R., 2007. Can a polite intelligent tutoring system lead to improved learning outside of the lab? Frontiers in Artificial Intelligence and Applications 158, 433.

Mikheeva, M., Schneider, S., Beege, M., Rey, G.D., 2019. Boundary conditions of the politeness effect in online mathematical learning. Computers in Human Behavior 92, 419–427.

Moreno-Marcos, P.M., Pong, T.C., Munoz-Merino, P.J., Kloos, C.D., 2020. Analysis of the factors influencing learners' performance prediction with learning analytics. IEEE Access 8, 5264–5282.

Morrison, D.M., Nye, B., Rus, V., Snyder, S., Boller, J., Miller, K., 2015. Tutorial dialogue modes in a large corpus of online tutoring transcripts, in: International Conference on Artificial Intelligence in Education, Springer. pp. 722–725.

Neuendorf, K.A., 2017. The Content Analysis Guidebook. SAGE Publications.

Niu, T., Bansal, M., 2018. Polite dialogue generation without parallel data. Transactions of the Association for Computational Linguistics 6, 373–389.

Nye, B.D., Graesser, A.C., Hu, X., 2014. Autotutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education 24, 427–469.

Person, N.K., Kreuz, R.J., Zwaan, R.A., Graesser, A.C., 1995. Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. Cognition and instruction 13, 161–188.

Schegloff, E.A., Sacks, H., 1973. Opening up closings .

Schez-Sobrino, S., Gmez-Portes, C., Vallejo, D., Glez-Morcillo, C., Redondo, M.A., 2020. An intelligent tutoring system to facilitate the learning of programming through the usage of dynamic graphic visualizations. Applied Sciences 10, 1518.

Schneider, S., Nebel, S., Pradel, S., Rey, G.D., 2015. Mind your ps and qs! how polite instructions affect learning with multimedia. Computers in Human Behavior 51, 546–555.

Slavin, R.E., 1987. Making chapter 1 make a difference. The Phi Delta Kappan 69, 110–119.

Sottilare, R.A., Graesser, A., Hu, X., Goldberg, B., 2014. Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management. volume 2. US Army Research Laboratory.

Tang, Y., Liang, J., Hare, R., Wang, F.Y., 2020. A personalized learning system for parallel intelligent education. IEEE Transactions on Computational Social Systems 7, 352–361.

VanLehn, K., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational psychologist 46, 197–221.

Walker, E., Ogan, A., 2016. We're in this together: Intentional design of social relationships with aied systems. International Journal of Artificial Intelligence in Education 26, 713–729.

Wang, N., Johnson, W.L., 2008. The politeness effect in an intelligent foreign language tutoring system, in: International Conference on Intelligent Tutoring Systems, Springer. pp. 270–280.

Wang, N., Johnson, W.L., Mayer, R.E., Rizzo, P., Shaw, E., Collins, H., 2008. The politeness effect: Pedagogical agents and learning outcomes. International journal of human-computer studies 66, 98–112.

Wang, N., Johnson, W.L., Rizzo, P., Shaw, E., Mayer, R.E., 2005. Experimental evaluation of polite interaction tactics for pedagogical agents, in: Proceedings of the 10th international conference on Intelligent user interfaces, pp. 12–19.

Yang, F., Li, F.W., 2018. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers & Education 123, 97–108.

2.3 Chapter Summary

To address Gap 1 and RQ 1, in Chapter 2, we employed the politeness tools (politeness strategy tools and politeness scoring tool) to extract the politeness strategies and levels from the tutoring dialogues. We also investigated the predictive power of politeness on the prediction of learner problem-solving performance. In Chapter 2, we presented the results of the study aimed to understand the role of politeness in dialogue-based feedback. Specifically, the main contributions of the study in Chapter 2 are summarised below:

Contribution 1: Track the change of politeness in the provisions of tutoring dialogues.

- In the effective tutoring dialogue sessions where learners managed to solve problems, the educators were likely to use polite expressions at the beginning and gradually used more direct expressions (e.g., "*Solve the value of x first*) to guide the learners as the tutoring dialogue progressed.
- Compared to the most ineffective tutoring sessions where learners had no progress on the problem-solving after tutoring sessions, the educators in the most effective tutoring session tended to use more the directive statements (a type of direct expression, "*Remember the denominator cannot be zero*") to guide the learners with prior progress.
- The learners in the effective sessions were slightly more polite at the beginning and the end of tutoring dialogues compared with the learners in the most ineffective tutoring sessions.

Contribution 2: The capability of politeness features in predicting learner problem-solving performance.

- By investigating the politeness in instructional communication (e.g., providing hints and corrective feedback) and non-instructional communication (e.g., greeting and off-topic statements), our study found that in the non-instructional communication, there was no evident correlation between the politeness of the educators' expression and the learner's performance. In comparison, in instructional communication, the learners would benefit more from the educators' direct expressions than polite expressions.
- Our results demonstrated the politeness features (i.e., politeness strategies and levels) were insufficient to reveal the learners' problem-solving performance and should incorporate with other features (e.g., the sentiment of the utterances).

Our results to RQ1 were derived from real-world online tutoring, which can provide guidance for novice educators to consider the use of politeness in tutoring (**Contribution 1**) and the design of dialogue-based ITS regarding the use of politeness in tutoring dialogue (**Contribution 1** and **Contribution 2**).

Chapter 3

Mining Effective Instructional Strategies from Tutoring Dialogues

3.1 Introduction

In Chapter 2, we investigated politeness in instructional communication (e.g., providing hints and corrective feedback) and non-instructional communication (e.g., greeting and off-topic statements), we found that in non-instructional communication, there was no evident correlation between polite expressions by educators and learner problem-solving performance. In comparison, in instructional communication, the learners would benefit more from the educators' direct expressions than polite expressions. To gain a better understanding of how educators express politeness in instructional communication, we argue that it is necessary to identify the communicative patterns related to instructional communication from the tutoring dialogues. By doing so, a widely used method is to use dialogue acts, which can characterise educators' and learners' intention behind their utterances in tutoring dialogues [44, 70]. Several prior studies [44, 70, 87, 110] have used dialogue acts to identify effective communicative patterns from the tutoring dialogues. However, the understanding of delivering effective tutoring dialogues is still limited (i.e., Gap 1 described in Chapter 1). Most prior studies [44, 45, 70, 87, 88] typically focused on the analysis of tutoring dialogue sessions where learners successfully solved the problems and the identification of effective instructional strategies that educators should take. We argue that educators should also learn from unsuccessful tutorial sessions (i.e., learners failed to solve the problems) and understand the factors that are responsible for causing those failures. Thus, it is worthwhile to devote more efforts to analysing the role of dialogue acts in tutoring dialogues (RQ 2). As shown in prior research [44, 87], instructional communication can be categorised into many specific types of strategies (e.g., corrective feedback and thought-provoking questions), which can be encoded into dialogue acts. Before exploring how politeness should be

expressed in instructional communication, we need to understand how educators communicate with learners in tutoring dialogues.

Driven by this, in Chapter 3, we adopted a widely-used educational dialogue act scheme [87] to characterise the conversational intentions behind utterances made by an educator/learner in tutoring dialogue. We aimed to identify communicative patterns from educators and learners in the form of dialogue acts in tutoring dialogues and evaluate the predictive power of dialogue acts in predicting learner problem-solving performance. Notably, the identified dialogue acts can be further used to facilitate the analysis of politeness in instructional communication (in Chapter 5). The results of this research have been published in the journal of *Future Generation Computer Systems*. The complete version of the manuscript is included in the following section.

- Lin, J., Singh, S., Sha, L., Tan, W., Lang, D., Gašević, D., & Chen, G. (2022). Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems*, 127, 194-207.

3.2 Publication: Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues.

Future Generation Computer Systems 127 (2022) 194–207

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues

Jionghao Lin [a,b], Shaveen Singh [a,b], Lele Sha [a,b], Wei Tan [b], David Lang [c], Dragan Gašević [a,b,d,e], Guanliang Chen [a,b,*]

[a] Centre for Learning Analytics, Monash University, Australia
[b] Faculty of Information Technology, Monash University, Australia
[c] Graduate School of Education, Stanford University, United States
[d] School of Informatics, University of Edinburgh, United Kingdom
[e] Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia

ARTICLE INFO

Article history:
Received 7 February 2021
Received in revised form 20 August 2021
Accepted 3 September 2021
Available online 9 September 2021

Keywords:
Intelligent Tutoring Systems
Educational dialogue analysis
Tutoring strategies
Dialogue acts
Student performance
Learning analytics

ABSTRACT

To construct dialogue-based Intelligent Tutoring Systems (ITS) with sufficient pedagogical expertise, a trendy research method is to mine large-scale data collected by existing dialogue-based ITS or generated between human tutors and students to discover effective tutoring strategies. However, most of the existing research has mainly focused on the analysis of successful tutorial dialogue. We argue that, to better inform the design of dialogue-based ITS, it is also important to analyse unsuccessful tutorial dialogues and gain a better understanding of the reasons behind those failures. Therefore, our study aimed to identify effective tutoring strategies by mining a large-scale dataset of both successful and unsuccessful human–human online tutorial dialogues, and further used these tutoring strategies for predicting students' problem-solving performance. Specifically, the study adopted a widely-used educational dialogue act scheme to describe the action behind utterances made by a tutor/student in the broader context of a tutorial dialogue (e.g., asking/answering a question, providing hints). Frequent dialogue acts were identified and analysed by taking into account the prior progress that a student had made before the start of a tutorial session and the problem-solving performance the student achieved after the end of the session. Besides, we performed a sequence analysis on the inferred actions to identify prominent patterns that were closely related to students' problem-solving performance. These prominent patterns could shed light on the frequent strategies used by tutors. Lastly, we measured the power of these tutorial actions in predicting students' problem-solving performance by applying a well-established machine learning method, Gradient Tree Boosting (GTB). Through extensive analysis and evaluations, we identified a set of different action patterns that were pertinent to tutors and students across dialogues of different traits, e.g., students without prior progress in solving problems, compared to those with prior progress, were likely to receive more thought-provoking questions from their tutors. More importantly, we demonstrated that the actions taken by students and tutors during a tutorial process could not adequately predict student performance and should be considered together with other relevant factors (e.g., the informativeness of the utterances).

1. Introduction

Dialogue-based Intelligent Tutoring Systems (ITS), similar to conventional ITS like SQL-Tutor [1], Algebra Tutor PAT [2], and eTeacher [3], aim at helping students construct knowledge and skills of different subjects by providing them with immediate and personalized instructions or feedback. Compared to conventional ITS, dialogue-based ITS deliver instructions or feedback by having natural and meaningful conversations with students [4], and are expected to act as competent as human tutors to engage students and provoke more in-depth thinking and learning. Given the promising potentials, both academic researchers and industrial practitioners have put great efforts in building various dialogue-based ITS, among which CIRCSIM-Tutor [5], AutoTutor [6], BEETLE II [7], and Why2 [8] are notable representatives. Noticeably, these systems have been deployed for use in practice and have assisted millions of students with their learning.

Despite being popular, most of the existing dialogue-based ITS are plagued by their inability in delivering personalized learning experiences to students [9]. The current dialogue-based ITS, as yet, fail to achieve their full potential and are unable to act as

* Corresponding author.
E-mail address: guanliang.chen@monash.edu (G. Chen).

https://doi.org/10.1016/j.future.2021.09.001
0167-739X/

competently as human tutors [10]. One main reason is that these dialogue-based ITS, more often than not, lack sufficient pedagogical expertise as human tutors in guiding students [11,12]. That is, these dialogue-based ITS typically have little knowledge about the tutoring strategies that can be of use to facilitate the tutoring process [13]. For instance, questioning a student's progress of learning problems is a common strategy used to help tutors detect knowledge gaps of the student at the beginning of a tutorial session [14]. Then, in follow-up teaching activities, tutors can better direct their efforts, e.g., introducing relevant learning contents and designing appropriate teaching activities to enable students to develop mastery of those concepts. It should also be noted that successful applications of such a tutoring strategy often depends on (i) a tutor's experience (and domain/contextual knowledge) in applying the strategy (e.g., when to ask questions and what type of question should be asked) and (ii) information about students (e.g., mastery level and learning progress) [15–18].

Numerous studies have been conducted to investigate how dialogue-based ITS can be equipped with relevant pedagogical expertise to apply appropriate tutoring strategies [4,11,17,19–22]. In this strand of research, a recent trend is to mine large-scale data collected by existing dialogue-based ITS or generated between human tutors and students to discover effective tutoring strategies [11,12,19,23]. However, existing data-intensive studies typically focused on the analysis of successful tutorial sessions (i.e., those in which students successfully solved problems or achieved meaningful learning) and the identification of effective tutoring strategies that tutors should take. We argue that, to provide students with necessary help, tutors should also learn from unsuccessful tutorial sessions and gain a better understanding of the factors contributing to such failures. Therefore, unsuccessful tutorial sessions should also be analysed to better guide the design and development of future dialogue-based ITS.

This study aimed to identify the frequent tutoring strategies used by tutors in not only successful but also unsuccessful tutorial sessions by mining a large-scale human–human tutorial dialogue dataset. The study also aimed to examine the extent to which these identified tutoring strategies are predictive of students' problem-solving performance. Here, we described a tutoring strategy as the actions taken by a tutor in the tutorial process (e.g., asking thought-provoking questions and providing hints). Formally, our work was guided by three **R**esearch **Q**uestions:

- **RQ1** What actions are commonly taken by tutors and students during tutorial sessions?
- **RQ2** What patterns of actions, i.e., one or multiple consecutive actions, are associated with different levels of students' performance in solving problems in tutorial sessions?
- **RQ3** To what extent are the identified actions and action patterns predictive the problem-solving performance of students in tutorial sessions?

To answer the above questions, we first employed a widely-used dialogue act (DA) scheme (proposed by [24]) to characterize tutors' (as well as students') actions behind their utterances in a tutorial dialogue. Then, the derived actions were analysed by applying a sequence analysis to shed light on the frequent tutoring strategies employed by tutors, which were further used as input for a well-established machine learning method—Gradient Tree Boosting (GTB) [25]—to measure the contribution made by these strategies in predicting students' problem-solving performance. To our knowledge, our study is the first to take students' prior progress into account to reveal effective tutorial strategies in human–human online tutoring. By analysing a large corpus consisting of both successful and unsuccessful tutorial dialogues, our study contributed with an in-depth understanding of tutors' as well as students' behaviour in human–human online tutoring and offered empirical evidence to support existing good practices (e.g., providing timely feedback to students) for the development of dialogue-based ITS.

2. Related work

Tutoring strategies refer to principles and approaches employed by instructors to better assist students to learn in various educational settings [15,26], e.g., raising a question to trigger in-depth thinking and acknowledging students' achievement to motivate them to continue to learn. Effective tutoring strategies play essential roles in helping instructors better direct their teaching efforts and enabling students to construct meaningful knowledge, and thus have been investigated for years [26].

Given the increasingly important role of dialogue-based ITS, there have been growing debates and research endeavours on investigating to what extend and how dialogue-based ITS should be constructed to make use of effective tutorial strategies to students. As summarized in [15], there are three typical methods employed to develop dialogue-based ITS with tutoring strategies:

- *Observing from human expert instructors*. Researchers crafted a set of tutoring rules by observing from human tutors' effective tutoring practices, which were further incorporated into the development of ITS. Take *AutoTutor* [27] as examples, which were developed by applying rule-based algorithms to simulate human tutors to employ tutoring strategies such as asking questions and providing short feedback to students. The strategies applied in AutoTutor were demonstrated effective in assisting undergraduate students to develop the mastery in introductory computer literacy courses [27].
- *Deriving from learning theories*. The design and development of an ITS were guided by well-established learning theories. For instance, by building upon the adaptive control of thought theory [28], which describes the characteristics of human cognition in the process of memory, [29] developed tutoring strategies such as decomposing the problem into a set of sub-problems and providing timely feedback on errors. These strategies were adopted by *cognitive tutors* [29] and demonstrated benefits to promote high-school students' success rate in the studies of algebra, geometry, and computer programming.
- *Observing students*. Researchers developed tutoring strategies by observing how students of different characteristics (e.g., gender, age, and mastery level in the learning subject) responded to different teaching practices. For example, Arroyo et al. [30] analysed how elementary-school students of different gender and different levels of cognitive development reacted to various types of hints (e.g., hints provided in a form of numeric symbolic or concrete visual shape) when students attempted to solve mathematical problems.

It is worth noting that all of the methods described rely heavily on the collection and analysis of the data generated between human tutors and students or between ITS and students, especially the methods that involve *observing from human expert instructors* and *observing students* [11,12,23,24,31]. As the first step to reveal effective tutoring strategies, existing studies often developed a coding scheme to characterize the actions hidden behind the utterances made by tutors and students in a tutorial dialogue (i.e., DAs) [32–34]. One of the pioneering studies in this strand of research [35] proposed to use a five-step framework to describe and characterize the process of one-on-one tutoring including the following steps: (i) a tutor starts by asking a

question; (ii) a student attempts to answer the question; (iii) the tutor provides feedback on the quality of the answer; (iv) the tutor and the student collaborate to improve the quality of the answer; and (v) the tutor assesses the student's understanding of the answer. By analysing two samples of tutorial dialogues, this framework was demonstrated as effective in identifying frequent dialogue patterns that characterized the collaborative nature of the one-on-one tutorial process. The effectiveness of this five-step framework was also validated in several studies [6,36–38]. As another example, Hennessy et al. [34] developed a coding scheme to take interlocutors' sociocultural backgrounds into account (e.g., the relationship between the interlocutors), which consists of 33 dialogue act codes including `express or invite ideas` and `nake reasoning explicit`. Other representative studies on developing DA schemes include [24,39–41]. In particular, the scheme presented in [24] was developed by building upon several prior schemes, which consisted of two levels of DA tags, with the first level describing the general flow of a tutorial process (e.g., a tutor raised a question) and the second level capturing more fine-grained information of a specific action taken by a tutor/student (e.g., what type of question was asked by the tutor). This DA scheme has been widely adopted in recent studies [42–46] and demonstrated effective in revealing varying dialogue patterns.

With the DA tags determined for the utterances, effective tutoring strategies were investigated by analysing the relationship between students' learning performance and dialogue actions taken by tutors or students [11,24,47–52]. For example, Boyer et al. [47] demonstrated that the actions performed by tutors played a significant role in delivering different kinds of learning outcomes, e.g., actions offering encouragement to students were useful in bolstering self-efficacy, while actions providing positive cognitive feedback were helpful in boosting learning gains. In a similar view, Vail and Boyer [24] applied correlation analysis to reveal actions and action bigrams (i.e., two consecutive actions) that were indicative of student learning, e.g., the action bigram consisting of positive feedback given by tutors after a confirmation question from the student was positively correlated with the learning gain of students. In particular, there are three studies most relevant to our work [11,12,19]. All of these three studies made use of the DA scheme developed by Morrison et al. [39] to capture the interaction between human tutors and students recorded in a large-scale dataset consisting of over 19K tutorial dialogues. To locate the effective tutoring strategies adopt by tutors, Maharjan et al. [11,12] applied sequence analysis to characterize the significant action patterns displayed in successful tutorial dialogues. These patterns indicated that tutors tended to be more expressive and encouraging at the beginning of successful tutorial dialogues and used more scaffolding strategies (e.g., providing a series of hints to students) during the tutorial process.

Compared to the works described above, our work distinguished itself from several perspectives. Firstly, the studies presented in [11,12,19] relied on a large amount efforts of experienced coders to label the dialogues, which was rather costly and time-consuming. Our research, instead, applied the state-of-the-art pre-trained language model BERT with a small sample of labelled data as input to automatically infer the actions taken by tutors and students and tutors in all tutorial dialogues. Secondly, most of the existing studies have mainly focused on analysing successful tutorial dialogues, i.e., those in which students achieved meaningful learning. In contrast, we distinguished dialogues in a more fine-grained manner by considering both students' prior progress in solving a task before the start of a tutorial session as well as the students' ultimate problem-solving performance after the end of the session, and revealed action patterns that were specific to different categories of dialogues. Thirdly, to our knowledge, our work is the first study to apply a state-of-the-art machine learning method–GTB–to systematically quantify the power of tutorial actions and action patterns in predicting students' problem-solving performance in human–human online tutoring.

3. Methods

3.1. Dataset

With the ethics approval from Monash University for secondary data use (Project ID 26156), we used a deidentified tutorial dataset that was prepared by an educational technology company. The educational technology company provides a mobile phone application for tutors and students to work together to solve problems covering subjects like mathematics, chemistry, and physics. With the mobile application, a student could take a picture of an unsolved problem and initialize a request for help. Then, the application connected the student with an experienced tutor who guided the student to solve the problem by leveraging texts and images to communicate. According to the policy of the educational technology company, tutors should give their best to guide students to solve problems by themselves and are disallowed to directly share answers with the students. That means, the dialogues contained in this dataset detailed processes of how tutors and students collaborated to solve various problems.

Recall that this study aimed to reveal frequent tutoring strategies occurring in not only successful but also unsuccessful tutorial dialogues. Notice that students with different learning progress often requires different support to complete a learning task. A tutor is often suggested to select appropriate tutoring strategies by taking into account the prior progress the student has achieved before entering a tutorial session [53,54]. Therefore, we manually labelled the dataset in two steps. That is, for each dialogue, we first determined the level of prior progress that a student made in solving a problem before talking to a tutor, and then further characterized the level of problem-solving performance the student achieved after the end of a tutorial session, as described below.

Step 1: Prior Progress. At the beginning of a tutorial session, a tutor often raised questions to a student to learn about whether the student had made certain progress in solving a problem. For instance, a tutor might ask *"Can you tell me what you have tried so far?"*, and a student might answer *"I haven't tried anything"* or *"Yes"* and then described the progress she had made. By manually scrutinizing the first few utterances of a tutorial dialogue, we were able to determine a student's level of prior progress and, correspondingly, labelled the dialogue as either *With-Prior-Progress* or *Without-Prior-Progress*. In total, we recruited three human coders to label the whole dataset. Each dialogue was labelled by two coders and the disagreements between the two coders were resolved by the third coder. The overall agreement percentage score was 0.847 and the Cohen's κ score was 0.735, which demonstrated a substantial level of agreement.

Step 2: Problem-solving Performance. Different from previous studies [11,12,19,24], we distinguished the performance level of a student in a more fine-grained manner and labelled each dialogue in our dataset to one of the following categories:

- **Gap-clarified**: a tutor was able to identify the problem but uncertain whether the student had made any progress;
- **Gap-explained**: a tutor was able to identify the problem and help the student make certain progress, but the student had not identified a correct or full solution;

Table 1
The descriptive statistics of the dataset used in the study. With PP and Without PP denote *With Prior Progress* and *Without Prior Progress*, respectively. Mann–Whitney tests were applied to examine the difference (Rows 5–9) between any two of the Gap-clarified, Gap-explained, and Gap-bridged categories in which students had the same level of prior progress. All differences were significant ($p < 0.01$).

Metric	All	Gap-clarified		Gap-explained		Gap-bridged	
		With PP	Without PP	With PP	Without PP	With PP	Without PP
1. # total sessions:	14,562	1,203	1,302	1,255	1,931	4,482	4,389
2. # total uttrances:	1,216,784	31,014	30,128	78,575	113,099	475,849	488,119
3. # tutors:	116	92	96	98	99	110	106
4. # students:	5,165	763	962	908	1,419	1,800	2,168
5. Avg. Sess Dur (mins):	30.27 ± 30.66	10.55 ± 7.64	9.75 ± 7.21	25.94 ± 19.03	22.88 ± 18.05	37.78 ± 32.17	38.60 ± 37.37
6. Avg. # Uttr/Sess:	83.56 ± 81.05	25.78 ± 16.68	23.14 ± 14.92	62.61 ± 43.79	58.57 ± 42.73	106.17 ± 87.62	111.21 ± 93.70
7. Avg. # Words/Sess:	647.75 ± 596.12	201.62 ± 131.81	198.13 ± 134.44	524.09 ± 351.18	489.56 ± 346.82	807.46 ± 649.08	845.28 ± 675.05
8. Avg. % Uttr by tutors:	58.42 ± 7.86	53.95 ± 9.49	56.46 ± 9.51	58.68 ± 7.82	60.25 ± 7.77	58.03 ± 7.07	59.75 ± 6.94
9. Avg. % Words by tutors:	78.36 ± 9.10	74.32 ± 11.80	80.54 ± 10.20	78.87 ± 8.59	82.21 ± 7.96	76.09 ± 8.69	79.30 ± 7.81

- **Gap-bridged**: a tutor was able to identify the problem and guide the student to successfully solve the problem or a similar problem.

In this step, each tutorial dialogue was labelled by an independent educational expert employed by the educational technology company that collected the dataset. To validate the reliability of these expert-crafted labels, we randomly selected 500 tutorial dialogues and labelled them independently by using the same coding rules. Our labels reached a percentage agreement score of 0.884 and Cohen's κ score of 0.787 with those expert-crafted labels. We provided sample dialogues for each of these categories in an electronic appendix, which is accessible via https://github.com/bertDA/BertDA/blob/main/DA_Appendix.pdf

It is worth noting that the three dialogue categories were on an ordinal scale relative to the level of students' problem-solving performance, i.e., there was an increasing amount of problem-solving progress obtained by students from the Gap-clarified dialogues to the Gap-bridged dialogues. The dataset originally consisted of 18,203 dialogues. Since tutors were unlikely to deliver meaningful tutoring in short dialogues, we removed dialogues that (i) contained less than 10 utterances; (ii) lasted less than 1 min; (iii) were difficult to be determined whether a student had made any progress before the start of a tutorial session or during the tutorial session, e.g., those in which a student quit a session because no tutor was assigned to help the student or a student did not reply to a tutor at all in the whole tutoring process. After removal, there were 14,562 dialogues left, among which about 92% were related to math tutoring. The descriptive statistics of the dataset are given in Table 1. Most of the dialogues were of the category Gap-bridged (8,871, 60.9%), followed by Gap-explained (3,186, 21.9%) and then Gap-clarified (2,505, 17.2%). This suggests, in our case, more than a half of the students successfully solved problems. Besides, over 47% of the students had made certain progress before joining a tutorial session, and compared to their counterparts without any prior progress, these students were more likely to identify correct solutions (i.e., Gap-bridged dialogues), which was in line with our expectations. Also, it would be intuitive to assume that, the better problem-solving performance achieved by a student, the more efforts the student as well as the tutor had invested in a tutorial session. To corroborate this assumption, we depicted the characteristics of the dialogues in rows 5–7 of Table 1. As we can observe, there was a steady increase from Gap-clarified and Gap-explained to Gap-bridged in terms of the session duration and the number of utterances and words contained in a dialogue.

3.2. Dialogue act scheme and dialogue act labelling

In line with previous studies [24,39], we also characterized the underlying actions taken by tutors and students in a tutorial session by using the two-level DA scheme presented in [24], whose effectiveness in depicting online one-on-one tutorial process has been validated in several studies [43,55,56]. The structure of the DA scheme is detailed in Table 2. Specifically, there are 12 first-level DA tags in the scheme, which can be used to portray the general tutor–student interaction, e.g., tutors raised thought-provoking questions to students (i.e., the tag `Question`) and students answered the questions raised by the tutors (i.e., the tag `Answer`). To capture more fine-grained information from tutor–student interaction, the 12 first-level labels are further expanded to 31 second-level DA tags. For instance, the tag `Question` is extended to distinguish between types of questions raised by tutors, including the questions requiring students to recall specific learning concepts (`Factual Question`), the questions prompting students' critical thinking (`Probing Question`), and the questions encouraging students to reason and reflect (`Open Question`). Noticeably, some DA tags are pertinent to only tutors or students while the others are pertinent to both. For example, the tag `Request Feedback` can only be used to describe the utterances generated by students to seek feedback, while the tags `Positive Feedback` and `Negative Feedback` can only be used to describe the utterances generated by tutors to provide feedback to students. As for tags like `Acknowledge` (expressing agreement with or acknowledgement of their interlocutors) and `Correction` (correcting the typo errors made in their previous utterance) can be used to describe utterances made by both tutors and students. Recall that our dataset was obtained from a mobile phone application used for online one-on-one tutoring, which allows tutors and students to use not only texts but also images to communicate. Based on our observations on the dataset, tutors typically used images to provide hints to inspire students and students often used images to seek feedback from tutors for the partial or full solution developed by them. None of the existing tags in the DA scheme can be used to depict these actions. Therefore, to better capture the tutor–student interaction observed in our dataset, we added two new second-level tags to the DA scheme, i.e., `Hint by Image` within the first-level tag `Hint` and `Request Feedback by Image` within the first-level tag `Request Feedback`.

Considering the number of tags contained in the adopted DA scheme and the number of tutorial dialogues contained in our dataset, it would be a very time-consuming and costly process if we purely relied on human coders to identify the DA tags for the whole dataset. Inspired by Rus et al. [19,57], we labelled a subset of the original dataset, which was used to train a classifier by applying machine learning techniques and then further used the classifier to automatically infer the DA tags for the remaining data. Specifically, we recruited two educational experts who have been involved in teaching for years as coders to label 50 randomly-selected tutorial dialogues in our dataset, which contained a total of 3,629 utterances. It should be noted that an utterance often contained multiple sentences and different

Table 2
The description of the DA scheme developed in [24]. The DA tags marked with ♣ are added by us to better depict the tutorial process in our dataset. The column Role indicates whether a DA is only specific to tutors (T), students (S), or specific to both. (* `Operational Question` tag was Originally denoted as `Question` in the DA scheme in [24]. We described it as `Operational Question` to better illustrate the difference between this tag and other tags.)

First-level DA Tag	Second-level DA Tag	Role	Examples in our dataset
Hint	Information	T	*"It can be any one of the cards in the deck."*
	Hint by Image ♣		*[Image]*
	Observation	T&S	*"We have 80."*
Directive	Directive	T	*"Check this definition."*
Acknowledge	Acknowledge	T&S	*"Alright!"*
Request Confirmation	Evaluation Question	T	*"Does that make sense?"*
Request Feedback	Request Feedback by Image ♣	S	*[Image]*
	Confirmation Question		*"Would the answer be 30?"*
Positive Feedback	General Positive Feedback	T	*"Correct!"*
	Elaborated Positive Feedback		*"Your formula for period is correct!"*
Negative Feedback	Negative Feedback	T	*"No, it is incorrect."*
Lukewarm Feedback	Lukewarm Feedback	T	*"Almost correct, but the sign is missing."*
Correction	Correction	T&S	*"*We will"*
Question	Direction Question	S	*"How do I do that?"*
	Information Question		*"What are the units for W?"*
	Probing Question		*"How many options can it be?"*
	Open Question	T	*"What do you think we could try next?"*
	Factual Question		*"What is the value of x?"*
	Operational Question *		*"Any questions on this?"*
	Ready Question		*"Are you ready to begin?"*
	Extra Domain Question	T&S	*"How are you doing today?"*
Answer	Yes-No Answer	T&S	*"Yes, that would be very helpful."*
	WH Answer	T&S	*"It is 6."*
	Ready Answer	S	*"Yes, I'm ready."*
	Extra Domain Answer	T&S	*"I'm good."*
Statement	Explanation	T&S	*"The straight line is the line on the bottom."*
	Greeting	T&S	*"Hello!"*
	Extra Domain Other	T&S	*"Welcome to use this app!"*
	Reassurance	T	*"No problem, I will help you."*
	Understanding	S	*"Ok, got it."*
	Not Understanding		*"I don't know why."*

sentences could indicate different actions (i.e., sentences could be assigned with different DA tags), the labelling was performed on a sentence level. Also, to enable enough fine-grained information to be captured, we asked the coders to identify not only the first-level but also the second-level tags for each sentence. Before starting the labelling, the two coders were required to develop a clear understanding of each tag contained in the DA scheme and correspondingly crafted a set of labelling rules (e.g., sentences containing keywords like "*hello*" and "*welcome*" should be assigned with the tag `Greeting`). Then, the two coders started to annotate five tutorial dialogues together, through which the labelling rules were revised and expanded to facilitate the subsequent labelling. Then, each of the remaining 45 tutorial dialogues was labelled by the two coders independently and their overall agreement score was 0.77 (measured by Cohen's κ), which indicates a substantial agreement between the two coders and the derived labels were reliable. The cases with disagreements were resolved by inviting a third educational expert to discuss together with the original two coders. As we aimed to reveal the frequent actions and action patterns used in tutorial dialogues, only the sentences labelled with first-level tags which occurred in more than 5% of total sentences were further labelled with second-level DA tags.

3.3. Inferring educational dialogue acts

With the labelled data derived in Section 3.2, we aimed to construct a classifier to automatically infer the DA tags for the remaining dialogues. Driven by the great success achieved by pre-trained language models in deriving accurate representations of textual data [58], which can be further utilized to facilitate downstream prediction tasks (e.g., DA identification in our case), we also used BERT [59,60] in our study. Notably, BERT has been demonstrated as effective in various settings, even with a limited amount of labelled data [59]. Here, we concatenated a single classification layer as the task model on top of BERT's output for the [CLS] and [SEP], which are the special tokens used in BERT embedding to encode the information of the sentence segmentation from the whole input data. As indicated in [61], the assignment of a DA tag to a sentence often depends on the context in which the sentence was uttered. Therefore, in order to make use of the context related to a sentence for DA prediction, we concatenated the following information for each labelled sentence as input to train the classifier:

- The text of a sentence;
- The person who uttered the sentence (i.e., tutor or student), which enabled BERT to relate the linguistic difference of tutor/student-generated utterances to different tutor/student-specific DA tags;
- The order of a sentence in a tutorial dialogue, which enabled BERT to account for the occurrence likelihood of different DA tags throughout a tutorial process;
- The session ID, which enabled BERT to capture the overall context in which a sentence was uttered; and
- The text of the sentence preceding the current sentence, which enabled BERT to pay specific attention to the local context surrounding a sentence.

With the classifier built, the DA tags of the remaining dialogues were automatically inferred. Then, the distribution of

these DA tags in the whole dataset as well as in each category of dialogues were further analysed to answer RQ1.

3.4. Mining frequent action patterns

To answer RQ2, we employed the `TraMineR` package in R to identify the discriminant action patterns in our dataset. `TraMineR` is a popular tool used to mine, describe, and visualize discriminant sequences or discrete sequences of states or events in data. Though primarily developed to analyse biographical longitudinal data, `TraMineR` has been successfully applied to other kinds of categorical sequence data, including sequences of actions in tutorial dialogues [11,12]. Specifically, we used `TraMineR` as follows: (i) we first extracted the frequent action patterns by counting their occurrence frequency in all dialogues with the aid of the `seqefsub()` function of `TraMineR`; then (ii) these frequent action patterns were used as input to the `seqecmpgroup()` function of `TraMineR`, which applied the Pearson Independence Chi-squared test with Bonferroni correction to retrieve action patterns that can be used to discriminate dialogues with different levels of student performance. To depict how discriminative an action pattern is, we further computed the value of *Pearson Residual*, which is a statistic used to compare the dispersion of the observed action pattern with the expected occurrence and indicate the degree of its departure to the expected occurrence. A positive Pearson Residual indicates that the actual occurrence of an action pattern is higher than its expected occurrence, while a negative Pearson Residual indicates a lower actual occurrence than the expected occurrence. Here, we selected a p-value threshold of 0.01 so as to reveal the most likely distinctive action patterns for different categories of dialogues. It is worth noting that an action pattern is not necessarily a contiguous sequence of DAs observed in the data. Instead, the order of the observed DAs is preserved. For instance, *(Tutor, Information)-(Tutor, General Positive Feedback)* may be formed from the contiguous sequence of *(Tutor, Info)-(Student, Confirmation Question)-(Tutor, General Positive Feedback)*.

3.5. Predicting student performance

Prediction Model. For RQ3, we aimed to evaluate the effectiveness of the observed actions or action patterns in predicting the problem-solving performance of students, i.e., predicting which label (among Gap-clarified, Gap-explained, and Gap-bridged) should be assigned to a tutorial dialogue. Essentially, this can be treated as a multi-class classification problem. Examples of typical techniques used for multi-class classification problems are Naive Bayes, decision trees, and support vector machines, while recent studies suggested that techniques like Gradient Tree Boosting [62] can also be of use. Our recent study [63] utilized this technique for predicting students' satisfaction with a tutoring service by leveraging a set of different features derived from the dialogue discourse. GTB is designed based on the rationale of ensemble learning [64], which states that multiple predictors aiming to predict the same target variable are more likely to deliver better performance than any single predictor alone. In fact, GTB is highly similar to random forests, both of which construct multiple decision trees as the predictors, and the final prediction is generated by combining the predictions of all constructed predictors with techniques like weighted average and majority vote. It is worth noting that each decision tree is constructed with a random sub-sample of the data. By doing this, each decision tree is slightly different from the others and more importantly, these decision trees together can adequately capture the characteristics of the data and thus deliver better prediction performance. The main difference between GTB and random forests lies in that, the predictors in random forests are constructed independently, which means, there can be multiple predictors producing the same type of prediction error. On the contrary, the GTB predictors are built in a sequential manner, in which the errors produced by previous predictors can be corrected by the latter predictors, and thus GTB takes less time to reach close to actual labels. In particular, previous research has demonstrated the effectiveness of GTB in dealing with various types of feature data for a wide range of machine learning problems. Therefore, in line with [63], whose prediction task is highly similar to ours (i.e., classifying educational dialogues), we also adopted GTB as the predictive modelling method.

Feature Engineering. To measure the extent to which the observed actions and action patterns in a tutorial dialogue can indicate the problem-solving achievement accomplished by students, we engineered the following three groups of features:

- **# Individual DA**: the number of a specific DA made by a tutor/student in a dialogue;
- **% Individual DA**: the fraction of a specific DA made by a tutor/student in a dialogue (divided by the total number of the identified DA);
- **# Significant action patterns**: the number of significant action patterns appeared in a dialogue (as discovered by applying the method described in Section 3.4).

In total, we designed 218 features based on the DA produced by both tutors and students, including (i) 18 first-level DA (here 6 out of the 12 original first-level DA are shared between tutors and students); (ii) 41 second-level DA (here 10 out of the 31 original second-level DA are shared between tutors and students); (iii) the percentage of 59 first-level and second-level DAs; and (iv) 100 discriminative action patterns found by `TraMineR`. We denoted these features as *DA features*. We acknowledged that a student's prior progress in solving a problem might be beneficial in boosting the prediction performance. However, the acquisition of such information relied on the manual analysis of the first few utterances in a tutorial dialogue in our current study. As we aimed to develop a prediction model that can be deployed for real-time use in practice, i.e. the input features to the model should be directly and automatically engineered from the observed data, we did not incorporate this into the feature set. In the future, we plan to develop methods to automatically determine the prior progress of a student as a tutorial session proceeds, and further take this information into account for predicting problem-solving performance.

Though we mainly focused on DAs produced by tutors and students in this study, as student performance may not be solely determined by the DAs, it would be necessary to include other relevant features to quantify the role of DA features in predicting student performance. For instance, the informativeness and complexity of utterances expressed by tutors may be greatly related to student performance [63]. Therefore, in addition to the *DA features* described above, we further used [63] as a reference and engineered another 325 features from the dataset and used them for predicting the labels of dialogues. These features include:

- **Efforts**, i.e., the efforts that a tutor/student invested in a tutorial session, which were measured by calculating the duration of the tutorial session, the number of utterances uttered by the tutor/student, and the number of words contained in those utterances;
- **Informativeness**, i.e., the informativeness of the utterances uttered by a tutor/student, which was measured by calculating the number (or fraction) of unique words and concepts contained in those utterances;

Table 3
Top 10 most frequent DAs identified in our dataset (sorted according to the fraction of utterances associated with a specific DA in the whole dataset in a descending order, i.e., the column **All**). The top 3 largest fraction numbers in each column are in bold. **T** denotes tutors and **S** denotes students. **With PP** and Without PP denote *With Prior Progress* and *Without Prior Progress*, respectively. Mann–Whitney tests were applied to examine the difference between any two of the Gap-clarified, Gap-explained, and Gap-bridged categories in which students had the same level of prior progress. Except for the results marked with the same symbol in a row (e.g., ◊, †), the others were all significant ($p < 0.01$).

	Dialogue Act	Role	All	Gap-clarified		Gap-explained		Gap-bridged	
				With PP	Without PP	With PP	Without PP	With PP	Without PP
1.	General Positive Feedback	T	**10.16%**	**9.18%**	6.71%	**8.16%**	**7.73%**	**11.41%**	**11.16%**
2.	Information	T	**8.62%**	6.57%	8.03%	**9.05%**	**10.35%**	**7.77%**	**8.96%**
3.	Probing Question	T	**8.10%**	6.81%	7.23%	◊ 8.33%	**8.46%**	◊ 7.87%	**8.50%**
4.	Yes-No Answer	S	7.19%	8.22%	**9.49%**	6.66%	◊ 7.59%	6.49%	◊ 7.02%
5.	WH Answer	S	6.41%	6.64%	† 6.97%	◊ 6.78%	† 6.39%	◊ 6.15%	6.44%
6.	Acknowledge	S	5.67%	7.37%	7.12%	5.35%	◊ 5.93%	5.32%	◊ 5.34%
7.	Request Feedback by Image	S	5.10%	**8.60%**	6.34%	◊ 5.13%	3.84%	◊ 5.45%	3.95%
8.	Extra Domain Other	T	4.93%	**9.43%**	**11.13%**	5.34%	5.70%	3.46%	3.30%
9.	Confirmation Question	S	4.89%	◊ 5.39%	5.19%	4.65%	4.59%	◊ 4.92%	4.93%
10.	Operational Question	T	4.73%	7.93%	**8.53%**	4.12%	◊ 4.45%	3.89%	◊ 3.98%

- **Complexity**, i.e., the complexity of the utterances uttered by a tutor/student, which was measured by applying Flesch readability score [65];
- **Responsiveness**, i.e., the average amount of time that a student needed to wait before receiving a reply from a tutor after the student sent an utterance;
- **# Questions**, i.e., the number of questions asked by a tutor/student in a tutorial session;
- **Entrainment**, which calculates a score to describe the degree to which tutors' utterances and students' utterances were aligned with each other in a tutorial session;
- **Sentiment**, i.e., the overall sentiment polarity scores of utterances sent by a tutor/student in a tutorial dialogue;
- **Experience**, i.e., the number of tutorial sessions that a tutor/student had prior to the current one;
- **N-grams**. The top 100 most frequent unigrams, bigrams, and trigrams contained in the utterances of a dialogue.

3.6. Study setup

Model Training for DA Classification. We implemented the classifier by using a BERT pre-trained language model [60]. For classifying DA, the number of neurons contained in the classification layer coupled with BERT was set to 768 and *softmax* was selected as the activation function. The labelled sentences were randomly split to *training*, *validation*, and *testing* datasets in the ratio of 80%:10%:10%. We set the maximum sequence length to 512 and fine-tuned on a batch size of 32 for 6 epoches. AdamW optimizer was used with the learning rate of 2e-5 to optimize the training of the classifier. All experiments were implemented on Titan GTX 2080ti and 2.50 GHz Intel Xeon E5-2678 v3 CPU processor.

Model Training for Student Performance Prediction. For predicting student performance, we randomly assigned the dialogues to the *training*, *validation*, and *testing* datasets in the ratio of 80%:10%:10%. For comparison, we selected random forests as the baseline method to demonstrate the effectiveness of GTB. Both random forests and GTB were implemented with the aid of the `scikit-learn`[1] library in Python and their parameters were fine-tuned by applying grid search on the validation data, and then we evaluated the performance of the two methods on the testing data.

Evaluation Metrics. For both of the two classification tasks above, we adopted three representative metrics for measuring the competency of the classification models, i.e., Area Under the Curve (AUC), F1 score, and Cohen's κ coefficient (Cohen's κ). We also present the result of classification accuracy as a reference.

[1] https://scikit-learn.org/.

4. Results

With the method described in 3.3, we built a DA classifier which successfully assigned correct labels for 75% of the sentences in the labelled dataset. More specifically, the classifier achieved a performance of 0.742 and 0.828 in terms of F1-score and AUC, respectively. In particular, the classifier achieved a Cohen's κ of 0.735, which demonstrated a sufficient performance level, especially given the large number of DA contained in our dataset (i.e., 31 second-level tags). This pre-trained DA classifier is available at https://github.com/bertDA/BertDA.

In the following, we detail the results obtained in response to the three RQs raised in Section 1.

4.1. Results on RQ1

The top 10 most frequent second-level DAs are shown in Table 3. These DAs, in total, accounted for 65.80% of the sentences in the dataset. We can observe that `General Positive Feedback`, `Information`, and `Probing Question` were ranked 1st, 2nd, and 3rd in the table, respectively. These actions were often taken by tutors to give necessary hints (`Information`), to raise thought-provoking questions (`Probing Question`), or to offer positive feedback (`General Positive Feedback`) to acknowledge students' achievement. The high occurrence frequency of such tutor-specific DAs suggests that, in online one-on-one tutoring, tutors tended to take the lead role in this collaborative problem-solving process. On the other hand, the most frequent actions by students were `Yes-No Answers` (i.e., a tag used to annotate students' responses, which typically start with a "*yes*" or "*no*", to simple questions), `WH Answers` (i.e., a tag used to annotate students' responses to complex questions with starting words including "*what*", "*why*", and "*how*") and `Acknowledge` (i.e., a tag used to annotate students' statements made to express acknowledgement or agreement to the explanations provided by tutors), which ranked 4th, 5th, and 6th in the table, respectively. Again, this is not a surprising result given the large number of the probing questions raised by tutors.

By differentiating the levels of students' prior progress, we can observe several findings in Table 3. Firstly, the fraction of the DA tag `General Positive Feedback` given to *With-Prior-Progress* students was generally higher than that to *Without-Prior-Progress* students in all three session categories. This is not a surprising result, as indicated in Table 1, a student with prior progress was more likely to successfully solve a problem, and thus received more positive feedback from tutors. Also, in the dialogues where students successfully solved problems (i.e., Gap-bridged), tutors had the highest usage of `General Positive`

Table 4
The top 10 most frequent action patterns from each category of dialogues. The patterns are sorted according to their occurrence frequency in an descending manner. The action patterns that occurred in only one performance category of dialogues (i.e., Gap-clarified, Gap-explained, and Gap-bridged) are in bold, and the action patterns that occurred in only one prior-progress category of dialogues (e.g., With or Without Prior Progress) are marked with ♠. T denotes tutors and S denotes students. Here are the abbreviation of the DA tags: Oprt-Ques (`Operational Question`), Y-N-Ansr (`Yes-No Answer`), Req-Fdbk-Img (`Request Feedback by Image`), G-Pos-Fdbk (`General Positive Feedback`), and Prob-Ques (`Probing Question`).

	Gap-clarified	Gap-explained	Gap-bridged
With Prior Progress	(S, Req-Fdbk-Img)-(T, Greeting) **(S, Req-Fdbk-Img)-(T, Oprt-Ques)** **(T, Greeting)-(T, Oprt-Ques)** **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Oprt-Ques)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ **(T, Greeting)-(T, Greeting)** ♠ **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Greeting)** ♠ (S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)	(S, Req-Fdbk-Img)-(T, Greeting) (S, Req-Fdbk-Img)-(T, Prob-Ques) **(T, Greeting)-(T, Prob-Ques)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) ♠ **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Prob-Ques)** ♠ (S, Req-Fdbk-Img)-(T, G-Pos-Fdbk) (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ (T, Greeting)-(T, G-Pos-Fdbk) **(S, Req-Fdbk-Img)-(T, Information)**	(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk) (S, Req-Fdbk-Img)-(T, Greeting) (T, Greeting)-(T, G-Pos-Fdbk) (S, Req-Fdbk-Img)-(S, Y-N-Ansr) ♠ **(S, Req-Fdbk-Img)-(T, Greeting)-(T, G-Pos-Fdbk)** **(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** **(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** (T, Greeting)-(S, Y-N-Ansr) ♠ **(T, Greeting)-(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** (S, Req-Fdbk-Img)-(T, Prob-Ques)
Without Prior Progress	(S, Req-Fdbk-Img)-(T, Greeting) (S, Req-Fdbk-Img)-(T, Oprt-Ques) (T, Greeting)-(T, Oprt-Ques) **(S, Req-Fdbk-Img)-(T, Greeting)-(T, Oprt-Ques)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ **(S, Req-Fdbk-Img)-(T, Oprt-Ques)-(S, Y-N-Ansr)** ♠ **(T, Oprt-Ques)-(S, Y-N-Ansr)** ♠ **(T, Greeting)-(T, Oprt-Ques)-(S, Y-N-Ansr)**	(S, Req-Fdbk-Img)-(T, Greeting) (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (T, Greeting)-(S, Y-N-Ansr) **(S, Req-Fdbk-Img)-(T, Information)** (S, Req-Fdbk-Img)-(T, Prob-Ques) ♠ **(S, Greeting)-(T, Information)** ♠ (S, Req-Fdbk-Img)-(T, Oprt-Ques) **(T, Greeting)-(T, Prob-Ques)** (S, Req-Fdbk-Img)-(T, Greeting)-(S, Y-N-Ansr) ♠ (S, Greeting)-(T, Oprt-Ques)	**(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)** (S, Req-Fdbk-Img)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Greeting) **(T, Greeting)-(T, G-Pos-Fdbk)** (T, Greeting)-(S, Y-N-Ansr) (S, Req-Fdbk-Img)-(T, Prob-Ques) ♠ (S, Req-Fdbk-Img)-(T, Oprt-Ques) **(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** **(S, Req-Fdbk-Img)-(T, G-Pos-Fdbk)-(T, G-Pos-Fdbk)** ♠ **(S, Y-N-Ansr)-(T, G-Pos-Fdbk)**

`Feedback`. This may indicate that positive feedback provided by tutors to students may be treated as a strong discriminator in revealing the problem-solving performance of students. Secondly, *Without-Prior-Progress* students, compared to *With-Prior-Progress* ones, received more `Information` and `Probing Question` from tutors in all three session categories. This indicates the extra scaffolding provided by tutors to assist students without making much progress in solving problems before entering a tutorial session. An interesting observation is that, compared to Gap-bridged students, Gap-explained students tended to receive more `Information` hints. Thirdly, it is noted that *Without-Prior-Progress* students in the Gap-clarified categories had high usage of `Yes-No-Answer` (9.49%). This suggests that tutors might have allocated extra efforts to engage and guide these students by asking simple questions. Among these questions, some utterances can be tagged as `Operational Questions`, e.g., *"Do you have any progress on it?"*. In addition, tutors in Gap-clarified dialogues used `Extra Domain Other` more frequently than their counterparts in the other two session categories.

4.2. Results on RQ2

To answer RQ2, we first extracted the frequent action patterns from all of the dialogues and counted their occurrence frequency in each category of dialogues (shown in Table 4). Here, we only considered patterns that appeared at least in 1% sentences in our dataset. In total, we identified 100 frequent action patterns. As action patterns consisting of only one DA tag, e.g., `Probing Question` and `Yes-No Answer`, were rather common in all categories of dialogues, we only present patterns that consist of at least two DA tags in Table 4. It is worth noting that all categories of dialogues have action patterns that are only specific to themselves, respectively. For instance, in Gap-clarified dialogues, a student's request for feedback (`Request Feedback by Image`) was often followed by responses from tutors who did not directly address the problem to be solved, such as `Operational Question` ("*Are you following me?*"). This, again, signified the extra efforts invested by a tutor to build the common problem-solving ground with a student. While scrutinizing the frequent patterns of Gap-bridged dialogues, we can easily observe that the same action taken by a student (i.e., `Request Feedback by Image`) was often followed by tutors' `General Positive Feedback`. As for the actions followed behind `Request Feedback by Image` in the Gap-explained dialogues, we can observe a high occurrence of `Information` (e.g., *"You should add the value of x"*) and `Probing Question` (e.g., *"How many elements do you get?"*), but not `General Positive Feedback`. These results, together, suggest that a student's problem-solving performance can largely be revealed by the varying usage of DA tags like `Operational Question`, `Information`, `Probing Question`, and `General Positive Feedback`. When comparing *With-Prior-Progress* with *Without-Prior-Progress* dialogues, we can see that students with prior progress received more `General Positive Feedback` from tutors while students without prior progress often faced more `Operational Questions` from tutors. This further corroborates our findings presented in Table 3.

With the aid of `TraMineR`, we identified a total of 100 discriminant action patterns that could be used to distinguish among the three categories of dialogues. The top 10 most discriminant patterns from the group *With Prior Progress* and *Without Prior Progress* are presented in Table 5. Interestingly, all of the top 10 patterns in both groups contained at least one of the following three DA tags pertinent to tutors, i.e., `General Positive Feedback`, `Probing Question`, and `Information`. This corroborates the findings presented in Table 4. This suggests that, when applying machine learning to predict the likelihood of a student successfully solving a task in a real-time manner, the thought-provoking questions asked by tutors and the hints and feedback provided by tutors can potentially be regarded as strong discriminators to distinguish different categories of tutorial sessions. Interestingly, `General Positive Feedback` is of a higher occurrence in the action patterns of *Without-Prior-Progress* dialogues that those of *With-Prior-Progress* dialogues in our case.

4.3. Results on RQ3

For RQ3, we aimed to investigate the extent to which DAs can be used to reveal students' problem-solving performance. It should be noted that, in the real-world tutorial scenarios, the earlier an unsuccessful session can be identified, the more interventions a tutor can take to help a student. Therefore, we were particularly interested in investigating whether the observed actions and action patterns displayed a varying predictive power as a tutorial session progressed. To this end, we selected the first N utterance in a tutorial dialogue to extract the DA features described in Section 3.5 as input for both GTB and random forests, and the results are depicted in Fig. 1 (the dash lines). We used $N \epsilon$ [5, 10, 15, 25, 30, 35, 40] as well as all of the available utterances as input for student performance prediction (denoted as ALL in Fig. 1). Based on Fig. 1, we can conclude that GTB was generally more effective than random forests in predicting student performance in terms of all the evaluation metrics, though the performance of these two models was relatively limited. For instance, the performance discrepancy between GTB and random

Table 5
The top 10 discriminant actions or action patterns in the group of *With Prior Progress* and *Without Prior Progress*. Action patterns that only occurred in either *With Prior Progress* or *Without Prior Progress* are in bold. Here, T denotes tutors and S denotes students. The value of Pearson Residual is used to compare the dispersion of the observed action pattern with the expected occurrence. The action patterns are sorted according to their Pearson Chi-square statistics in each group in a descending order, which indicate the extent to which an action pattern can be used to discriminate the three different categories of dialogues.

		Action Patterns	Pearson residual		
			Gap-clarified	Gap-explained	Gap-bridged
With prior progress	1	(T, Information)-(T, General Positive Feedback)-(T, General Positive Feedback)	-22.87	-3.34	13.62
	2	(T, Probing Question)-(T, General Positive Feedback)-(T, General Positive Feedback)	-19.78	-2.60	11.62
	3	(S, Request Feedback by Image)-(T, Information)-(T, General Positive Feedback)	-20.25	-0.17	10.59
	4	(T, Information)-(T, General Positive Feedback)	-20.25	-0.17	10.59
	5	**(T, General Positive Feedback)-(T, Information)-(T, General Positive Feedback)**	-22.48	-0.96	12.16
	6	(T, General Positive Feedback)-(T, Probing Question)-(T, General Positive Feedback)	-19.67	-1.46	10.97
	7	**(T, Information)-(T, Probing Question)-(T, General Positive Feedback)**	-22.80	-0.99	12.34
	8	**(T, Greeting)-(T, Information)-(T, General Positive Feedback)**	-20.21	-0.06	10.51
	9	**(T, Probing Question)-(T, Probing Question)-(T, General Positive Feedback)**	-21.09	-1.02	11.47
	10	**(T, Information)-(T, Information)-(T, General Positive Feedback)**	-23.41	0.01	12.13
Without prior progress	1	**(T, General Positive Feedback)-(T, General Positive Feedback)-(T, General Positive Feedback)**	-24.35	-4.64	16.33
	2	(T, Probing Question)-(T, General Positive Feedback)-(T, General Positive Feedback)	-24.06	-3.97	15.73
	3	(T, Information)-(T, General Positive Feedback)-(T, General Positive Feedback)	-24.72	-3.97	16.09
	4	**(T, General Positive Feedback)-(T, General Positive Feedback)**	-20.99	-2.02	12.76
	5	**(S, Request Feedback by Image)-(T, General Positive Feedback)-(T, General Positive Feedback)**	-20.98	-2.02	12.76
	6	(T, General Positive Feedback)-(T, Probing Question)-(T, General Positive Feedback)	-23.83	-2.60	14.70
	7	(T, Information)-(T, General Positive Feedback)	-22.13	0.08	11.99
	8	(S, Request Feedback by Image)-(T, Information)-(T, General Positive Feedback)	-22.12	0.09	11.99
	9	**(S, Yes-No-Answer)-(T, General Positive Feedback)-(T, General Positive Feedback)**	-22.07	-2.61	13.75
	10	**(T, Probing Question)-(T, General Positive Feedback)**	-20.69	-0.13	11.35

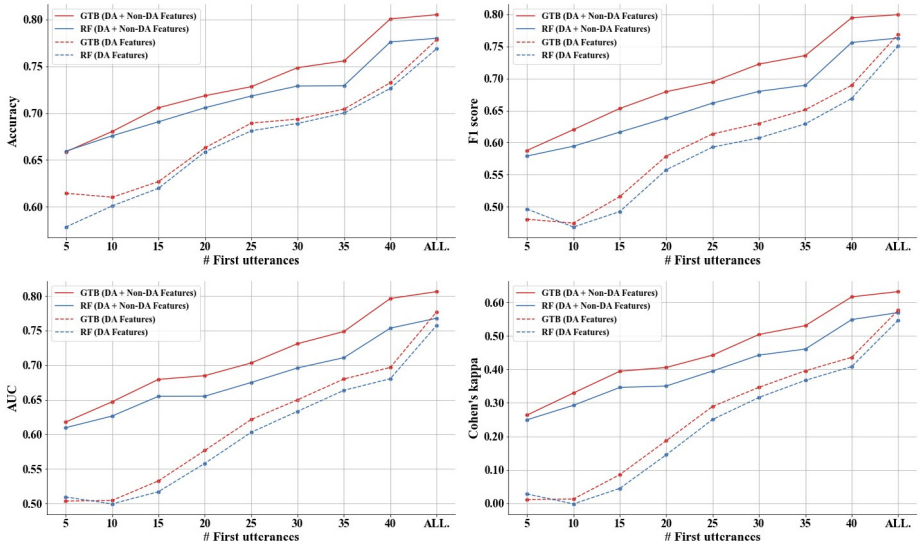


Fig. 1. The performance of GTB and random forests in predicting student performance in solving problems.

forests measured by the F1 score remained rather stable regardless of the number of input utterances. We can make similar observations when scrutinizing other evaluation metrics. When taking all of the available utterances as input, GTB achieved the performance of 0.779, 0.769, 0.577, and 0.777 as measured by accuracy, F1 score, AUC, and Cohen's κ, respectively. These results imply there is still space to further boost the prediction performance. Therefore, we further incorporated the non-DA features together with the DA features as input to the models for student performance prediction (the solid lines in Fig. 1). Unsurprisingly, the results indicate that both GTB and random forests achieved better prediction performance when taking both DA and non-DA features into account. If we take GTB ($N = 10$) as an example, we can see that the Accuracy was boosted from 0.610 to 0.680 and the F1 score was boosted from 0.475 to 0.620. The results of GTB for $N = ALL$ showed that the model achieved Cohen's κ score of 0.632, which indicates a substantial prediction performance. This suggests that, though being useful in characterizing different categories of tutorial sessions, tutors' tutoring actions and action patterns were insufficient in revealing students' problem-solving performance in online one-on-one tutoring. To gain a better understanding of the distinctive predictive power of different types of features, we further conducted an ablation test. That is, the contribution made by a feature is calculated as the difference between the prediction performance of a model when including the feature and that when excluding the feature [66]. Due to the limited space, we presented the results of the ablation test in the electronic appendix.[2] We found that the N-grams features (e.g., terms and phrases like "*great*" and "*good job*" in the positive feedback provided by tutors) were of particular importance in predicting students' problem-solving performance.

5. Discussion and conclusion

The construction of dialogue-based ITS with adequate pedagogic expertise is a longstanding task in the pathway towards

[2] Accessible via https://github.com/bertDA/BertDA/blob/main/DA_Appendix.pdf.

delivering on-time, personalized, and meaningful learning experiences to students. Though quite some studies have been carried out, these studies often ignored the analysis of unsuccessful tutorial sessions and seldom paid attention to the reasons behind these unsuccessful tutorial sessions. This motivated us to analyse a large-scale dialogue corpus (over 14 K), which consisted of both successful and unsuccessful online human–human tutorial sessions, to identify frequent tutoring strategies adopted by tutors and further use these tutoring strategies for predicting students' problem-solving performance. Through extensive analysis and evaluations, our study provided empirical evidence to support existing good practices for developing dialogue-based ITS and contributed to the research of educational dialogue analysis with the following main findings:

- Overall, tutors often took actions to provide feedback and information to students and to raise questions to guide students to solve problems. Correspondingly, students took more actions in answering questions or expressing agreement with tutors and acceptance of the provided explanations or solutions.
- In tutorial sessions where students delivered correct or partially correct solutions to the problems, tutors tended to ask more thought-provoking questions, offer more information hints, and pose less irrelevant statements or questions to students compared to tutorial sessions with lower problem-solving performance.
- If a student had made certain progress in solving a problem before entering a tutorial session, the student was likely to receive fewer hints and thought-provoking questions from a tutor, but still had a higher chance to successfully solve the task and received more positive feedback from the tutor.
- Positive feedback expressed by tutors can be used as a strong discriminator to differentiate dialogues of different student performance in solving problems.
- DA and DA patterns alone were insufficient to reveal the problem-solving performance of students and should be utilized together with other relevant factors (e.g., the informativeness, complexity, and the sentimental polarities of the utterance).
- We have released a DA classifier, which was constructed by applying state-of-the-art pre-trained language model BERT, to better support researchers for relevant research studies, which is accessible via https://github.com/bertDA/BertDA.

5.1. Implications

Firstly, utterances with tags such as `Information`, `Probing Question`, `Operational Questions`, and `General Positive Feedback` can be used to characterize tutorial sessions with different level of student performance. As reported in Table 3, the top three most frequently used tutorial dialogue actions taken by tutors in our dataset were `General Positive Feedback`, `Information`, and `Probing Question`. These tutorial actions, especially `General Positive Feedback` such as "*Correct!*" and "*Great!*", were more frequently observed in successful tutorial sessions (i.e., Gap-bridged), which is also evident in Tables 4 and 5. In fact, this is in line with the findings presented by previous studies [11,24,67], which suggested that the positive feedback provided by tutors played a vital role in verifying the correctness of the students' work, increasing a students' level of self-efficacy in accomplishing a learning task and encouraging the student to proceed with the remaining activities [67]. This implies that tutors providing online one-on-one tutoring service may consider, whenever it is appropriate, providing timely and adequate positive feedback to acknowledge students' achievements and further motivate them to deliver correct solutions. Though it might be possible that Gap-bridged students generally had a higher level of prior knowledge than their counterparts of the other two categories, which enabled them to be more likely to solve problems and thus received more positive feedback from tutors [68]. Future research should investigate the impact of students' prior knowledge on solving learning problems and what strategies should be utilized to better scaffold students with a low level of prior knowledge in our future research.

Secondly, the use of strategies like `Probing Question` and `Information` should be dependent on a student's level of prior progress. For instance, as shown in Table 3, *With-Prior-Progress* students generally received less `Information` hints than their *Without-Prior-Progress* counterparts. This observation is in line with the findings presented in [67], i.e., tutors should provide more scaffolding to students with little progress being made. Besides, we observed that, compared to Gap-explained dialogues, there was a lower usage of `Information` in Gap-bridged dialogues. Previous studies (e.g., [67,69]) showed that, to effectively engage a student in a learning task, tutors should avoid scaffolding the student with excessive information hints. Given that it might be rather challenging for tutors to determine a suitable amount of information hints to be provided to students in practice, we plan to develop automatic methods to measure the levels of both confusion and engagement of students and help tutors (or dialogue-based ITS) better direct their teaching efforts.

Thirdly, the failure of a tutorial session might not be entirely attributed to the extra use of DA tags that were not directly related to solving a learning problem. Table 3 shows that there was a higher occurrence of `Extra Domain Other` and `Operational Question` in Gap-clarified dialogues than the other two categories. To investigate the underlying reasons causing the use of such extra problem-solving-irrelevant utterances, we manually checked 200 randomly selected Gap-clarified dialogues which contained utterances tagged as `Extra Domain Other` and `Operational Question`, and found several issues. A common one is that tutors did not give enough time to a student to think and work on a problem and frequently asked operational questions like "*Are you working on the problem?*", which caused extra pressure to the student and further impeded her from solving the problem. Another common issue is that there were communication issues between tutors and students (e.g., "*I am sorry. I don't understand what you are trying to say*" and "*Let's clarify the information a bit*"). These communication issues, oftentimes, made a student quit a tutorial session before being able to receive any meaningful guidance. In addition, we observed there was a small portion of dialogues (about 12%) in which a student played against the rules (e.g., a student uttered "*Just give me the answer!*" and then a tutor replied with "*I know you might be frustrated, but handling out easy answers goes against our pledge and hurts you in the long run*") or a tutor did not provide timely responses (e.g., "*Sorry for the late reply*"), which also tended to make the student terminate the tutorial session before receiving help. These findings indicate that, when providing online tutoring service or designing dialogue-based ITS, appropriate methods should be developed to (i) remind tutors to allow students to have enough time to work on problems and provide timely feedback to them, (ii) facilitate the communication process between tutors and students (e.g., using a digital whiteboard), and (iii) provide students with a clear guideline (e.g., it is disallowed to directly share answers with students) to avoid potential misunderstandings held by certain students about the terms of service of the tutoring service and system.

Fourthly, GTB can be used to detect potentially successful tutorial sessions in real-world tutoring practices. As depicted in Fig. 1, GTB was capable of accurately classifying about 68% tutorial dialogues by only taking the first 10 utterances as input.

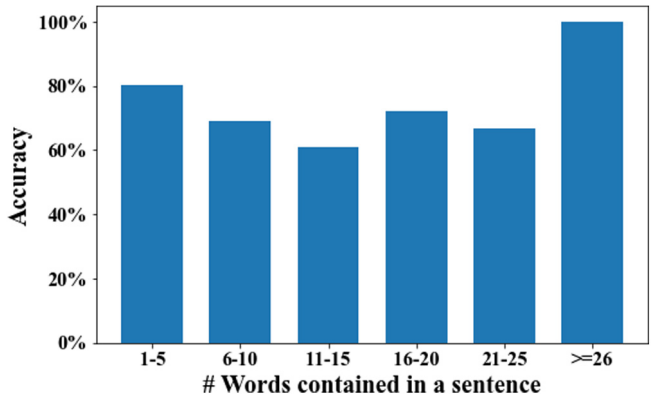


Fig. 2. The classification accuracy for sentences of different length.

With more utterances taken into account, the prediction accuracy kept increasing. This demonstrated that GTB is a reliable machine learning method that can be of practical use to locate potentially unsuccessful dialogues from effective ones. With these potentially unsuccessful dialogues detected in the early stage of a tutorial session, we expect that more interventions (e.g., changing a tutoring strategy) can be taken to better engage and assist a student.

5.2. Limitations

Misclassified Dialogue Acts. Recall that the DA classifier we presented in Section 3.3 was only able to deliver correct labels for 75% sentences in the labelled dataset, which suggests there is still certain improvement space. To gain a more in-depth understanding of the performance of the DA classifier, we calculated the prediction accuracy for each DA tag and found that some tags were more likely to be correctly identified by the classifier than others. For instance, the classifier was capable of distinguishing tags with characteristic keywords that rarely occurred in other tags, e.g., `Greeting` with keywords like "*hello*" and "*welcome*" and `General Positive Feedback` with keywords like "*great*", "*correct*", and "*awesome*". On the other hand, we noticed that tags with relatively fewer training samples were more likely to be misclassified. For instance, the DA `Elaborated Positive Feedback`, which was related to only 2.08% sentences in our labelled dataset, was often misclassified as `General Positive Feedback`. Though `Elaborated Positive Feedback` can also be characterized by keywords like "*great*" and "*correct*", sentences of this DA tag often contained more fine-grained information (e.g., "*Your formula for period is correct!*"). Our DA classifier was unable to distinguish this minor difference when there was a lack of enough training samples of `Elaborated Positive Feedback`. This suggests, for the future improvement of DA identification, it is worthwhile to label more tutorial sentences, especially those with tags of a low occurrence frequency, to enable the classifier to capture the fine-grained difference between various DA tags.

Next, as an initial step to investigate whether the amount of information contained in a sentence would likely impact the prediction performance of the classifier, we treated the number of words contained in a sentence as a proxy of the amount of information conveyed in the sentence, and plot the classification accuracy for sentences of different length in Fig. 2. Interestingly, we observed that, in our case, the classifier tended to deliver better performance when a sentence is particularly short ($[1, 5]$) or particularly long ($>= 26$) than the other sentences. However, considering that, as demonstrated before, the classifier tended to deliver varying performance when dealing with different DA tags, we could not conclude that the particularly short or long sentences were favoured by the classifier. In the future, it would be worthwhile to label more data for each DA tag, especially DA tags associated with sentences of varying lengths, and further investigate the impact of sentence length on the prediction performance of the classifier.

Table 6
Examples of the predicted DA tags delivered by the BERT-based classifier. Here are the abbreviation of the DA tags: Oprt-Ques (`Operational Question`), Extr-Dom-Othr (`Extra Domain Other Statement`), Y-N-Ansr (`Yes-No Answer`), and Ack (`Acknowledge`).

Role	Sentence	Actual	Prediction
Tutor	Anything else I can help you with?	Oprt-Ques	Oprt-Ques
Student	Awesome!	Extr-Dom-Othr	Y-N-Ansr
Student	That should be all thank you	Y-N-Ansr	Ack

Lastly, we observed that, for certain sentences, the wrong identification of DA tags could be explained by the lack of enough contextual information. As described in Section 3.3, when predicting the DA tag for a sentence, the text of the preceding sentence was incorporated as part of the input to capture relevant contextual information. However, as shown in Table 6, the contextual information related to a sentence might span more than one preceding sentence. That is, the student's response "*That should be all thank you*" was uttered to answer the question raised by the tutor ("*Anything else I can help you with?*"), but this response was preceded by another response uttered by the student (i.e., "*Awesome!*"). Therefore, the question raised by the tutor was not taken into account and the DA tag of "*That should be all thank you*" was misclassified as `Acknowledge`. This suggests that, to further improve the prediction performance of the DA classifier, it might be worthwhile to take additional contextual information (i.e., more than one proceeding sentence) into account.

Students' Performance Analysis *Firstly*, the categorical labels of the tutorial dialogues, i.e., Gap-clarified, Gap-explained, and Gap-bridged, were derived by one educational expert. Though a sanity check, which involved a second educational expert to use the same coding rules to label 500 dialogues randomly selected from the whole dataset, was conducted and a percentage agreement score of 0.884 was reached, future research efforts should be allocated to explore other methods to enhance the validity of the labelling results (e.g., employing crowd-sourcing workers to label the whole dataset). *Secondly*, as indicated before, students in certain dialogues played against the rules by asking tutors to directly share answers with them or students quit tutorial sessions because of not receiving timely responses from tutors. After manually scrutinizing 200 randomly-selected Gap-clarified dialogues, we found 12% were of this kind. We leave the automatic identification and exclusion of such dialogues for more fine-grained analysis for the future research. *Thirdly*, it has been widely recognized that the prior knowledge level of a student can significantly impact her performance in a learning task [70]. For instance, students with high prior knowledge, compared to those with low prior knowledge, are able to solve a learning problem with less information hints [71]. However, the information about students' prior knowledge level was not available in our dataset. As a remedy, we took into the level of prior progress achieved by a student before entering a tutorial session and demonstrated similar findings, i.e., students with prior progress were able to solve problems with less information hints compared to their counterparts without any prior progress. In the future, it would be worthwhile to further distinguish students' prior progress in a more fine-grained level (e.g., without progress, with small progress, and with much progress) and analyse its impact on

students' problem-solving behaviours and performance. *Fourthly*, we did not separate the analysis of dialogues of different subjects (i.e., math, chemistry and physics) in this study. Given that about 92% of the dialogues in our current dataset related to math tutoring, we plan to collect more dialogues of physics and chemistry to enrich the dataset, separate the dialogues of different subject areas for analysis, and further provide more in-depth insights to support the tutoring practices in different subject areas in our future work. *Fifthly*, though our work successfully revealed tutoring strategies that were frequently observed in both successful and unsuccessful tutorial dialogues, it still remains largely unknown when these tutoring strategies should be or should not be used. To further guide the development of future dialogue-based ITS, future research should focus on the content analysis of the utterances related to each DA tag and investigate the relationship between these DA tags and other relevant tags, e.g., whether students' statements specifying their confusion (i.e., the DA tag `Not Understanding`) always triggers an action from tutors to provide hints (i.e., the DA tag `Information`) and whether the continued use of such hint-providing actions likely promotes better student performance. For this purpose, causal models [72, 73] can be explored in future research. *Lastly*, though GTB was demonstrated to be effective in predicting students' problem-solving performance, there is still space to further improve the prediction performance. Given the wide success achieved by deep neural networks in tackling various types of tasks, especially those in the field of natural language processing, future research should investigate methods based on deep neural networks to deliver more accurate student performance predictions.

CRediT authorship contribution statement

Jionghao Lin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Shaveen Singh:** Conceptualization, Methodology, Validation, Formal analysis, Data curation. **Lele Sha:** Software, Validation, Data curation. **Wei Tan:** Software, Validation. **David Lang:** Resources. **Dragan Gašević:** Conceptualization, Validation, Supervision, Writing – original, Writing – review & editing, Project administration, Funding acquisition. **Guanliang Chen:** Conceptualization, Methodology, Validation, Supervision, Writing – original, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Mitrovic, Learning SQL with a computerized tutor, in: Proceedings of the Twenty-Ninth SIGCSE Technical Symposium on Computer Science Education, SIGCSE '98, ACM, 1998, pp. 307–311.

[2] S. Ritter, J. Anderson, M. Cytrynowicz, O. Medvedeva, Authoring content in the PAT algebra tutor, J. Interact. Media Educ. 1998 (2) (1998).

[3] S. Schiaffino, P. Garcia, A. Amandi, eTeacher: Providing personalized assistance to e-learning students, Comput. Educ. 51 (4) (2008) 1744–1754.

[4] V. Rus, S. D'Mello, X. Hu, A. Graesser, Recent advances in conversational intelligent tutoring systems, AI Mag. 34 (3) (2013) 42–54.

[5] M.W. Evens, J.A. Michael, One-on-One Tutoring by Humans and Computers, 2006.

[6] A.C. Graesser, P. Chipman, B.C. Haynes, A. Olney, AutoTutor: an intelligent tutoring system with mixed-initiative dialogue, IEEE Trans. Educ. 48 (2005) 612–618.

[7] M. Dzikovska, N. Steinhauser, E. Farrow, J. Moore, G. Campbell, BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics, IJAIED 24 (3) (2014) 284–332.

[8] K. VanLehn, A.C. Graesser, G.T. Jackson, P.W. Jordan, A. Olney, C.P. Rosé, When are tutorial dialogues more effective than reading? Cogn. Sci. 31 (1) (2007) 3–62.

[9] A. Almasri, A. Ahmed, N. Al-Masri, Y.S.A. Sultan, A.Y. Mahmoud, I. Zaqout, A.N. Akkila, S.S. Abu-Naser, Intelligent tutoring systems survey for the period 2000- 2018, Int. J. Acad. Eng. Res. 5 (3) (2019) 21–37.

[10] A. Alkhatlan, J. Kalita, Intelligent tutoring systems: A comprehensive historical survey with recent developments, Int. J. Comput. Appl. 181 (43) (2019) 1–20.

[11] N. Maharjan, V. Rus, D. Gautam, Discovering effective tutorial strategies in human tutorial sessions, in: The Thirty-First International Flairs Conference, 2018.

[12] N. Maharjan, V. Rus, A tutorial Markov analysis of effective human tutorial sessions, in: Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 30–34.

[13] A.C. Graesser, K. VanLehn, C.P. Rosé, P.W. Jordan, D. Harter, Intelligent tutoring systems with conversational dialogue, AI Mag. 22 (4) (2001) 39–51.

[14] K. Cotton, Classroom questioning, School Improv. Res. Ser. 5 (1988) 1–22.

[15] B. Du Boulay, R. Luckin, Modelling human teaching tactics and strategies for tutoring systems: 14 Years on, IJAIED 26 (1) (2016) 393–404.

[16] V.J. Shute, SMART: Student modeling approach for responsive tutoring, User Model. User-Adapt. Inter. 5 (1) (1995) 1–44.

[17] S. Katz, P.L. Albacete, A tutoring system that simulates the highly interactive nature of human tutoring., J. Educ. Psychol. 105 (4) (2013) 1126–1141.

[18] J. Paladines, J. Ramirez, A systematic literature review of intelligent tutoring systems with dialogue in natural language, IEEE Access 8 (2020) 164246–164267, http://dx.doi.org/10.1109/ACCESS.2020.3021383.

[19] V. Rus, N. Maharjan, L.J. Tamang, M. Yudelson, S. Berman, S.E. Fancsali, S. Ritter, An analysis of human tutors' actions in tutorial dialogues, in: The Thirtieth International Flairs Conference, 2017.

[20] K. VanLEHN, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educ. Psychol. 46 (4) (2011) 197–221.

[21] S. Sundararajan, S. Nitta, Designing engaging intelligent tutoring systems in an age of cognitive computing, IBM J. Res. Dev. 59 (6) (2015) 10:1–10:9.

[22] V. Rus, R. Banjade, N. Niraula, E. Gire, D. Franceschetti, A study on two hint-level policies in conversational intelligent tutoring systems, in: Innovations in Smart Learning, Springer, 2017, pp. 175–184.

[23] K.E. Boyer, R. Phillips, A. Ingram, E.Y. Ha, M. Wallis, M. Vouk, J. Lester, Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach, Int. J. Artif. Intell. Educ. 21 (1–2) (2011) 65–81.

[24] A.K. Vail, K.E. Boyer, Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes, in: ITS, 2014.

[25] M. Kumar, M. Kumar, et al., XGBoost: 2D-object recognition using shape descriptors and extreme gradient boosting classifier, in: Computational Methods and Data Engineering, Springer, 2021, pp. 207–222.

[26] A.C. Ornstein, T.J. Lasley, Strategies for Effective Teaching, Harper & Row New York, 1990.

[27] A.C. Graesser, N. Person, D. Harter, T.R. Group, et al., Teaching tactics in AutoTutor, Modell. Hum. Teach. Tact. Strateg. IJAIED 11 (2000) 1020–1029.

[28] J.R. Anderson, The Architecture of Cognition, Vol. 5, Psychology Press, 1996.

[29] J.R. Anderson, A.T. Corbett, K.R. Koedinger, R. Pelletier, Cognitive tutors: Lessons learned, J. Learn. Sci. 4 (2) (1995) 167–207.

[30] I. Arroyo, J.E. Beck, B.P. Woolf, C.R. Beal, K. Schultz, Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism, in: International Conference on Intelligent Tutoring Systems, Springer, 2000, pp. 574–583.

[31] J.G. Cromley, What do reading tutors do? A naturalistic study of more and less experienced tutors in reading, Discourse Process. 40 (2) (2005) 83–113.

[32] J.C. Marineau, P.M. Wiemer-Hastings, D. Harter, B.A. Olde, P. Chipman, A. Karnavat, V. Pomeroy, S. Rajan, A. Graesser, Classification of Speech Acts in Tutorial Dialog, 2000.

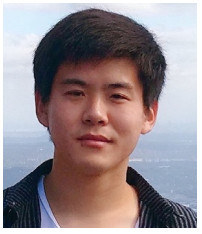
[33] R. Pilkington, Analysing Educational Discourse: The DISCOUNT Scheme, University of Leeds, Computer Based Learning Unit, 1999.

[34] S. Hennessy, S. Rojas-Drummond, R. Higham, A.M. Márquez, F. Maine, R.M. Ríos, R. García-Carrión, O. Torreblanca, M.J. Barrera, Developing a coding scheme for analysing classroom dialogue across educational contexts, Learn. Cult. Soc. Interact. 9 (2016) 16–44.

[35] A.C. Graesser, N.K. Person, J.P. Magliano, Collaborative dialogue patterns in naturalistic one-to-one tutoring, Applied Cognitive Psychology 9 (6) (1995) 495–522.

[36] K. Forbes-Riley, D. Litman, Investigating human tutor responses to student uncertainty for adaptive system development, in: Conference on ACII, Springer, 2007, pp. 678–689.

[37] K.E. Boyer, R. Phillips, A. Ingram, E.Y. Ha, M. Wallis, M. Vouk, J. Lester, Characterizing the effectiveness of tutorial dialogue with hidden markov models, in: ITS, Springer, 2010, pp. 55–64.
[38] A.C. Graesser, S. Lu, G.T. Jackson, H.H. Mitchell, M. Ventura, A. Olney, M.M. Louwerse, AutoTutor: A tutor with dialogue in natural language, Behav. Res. Methods Instrum. Comput. 36 (2) (2004) 180–192.
[39] D. Morrison, B. Nye, B. Samei, V.V. Datla, C. Kelly, V. Rus, Building an intelligent pal from the tutor. com session database phase 1: Data mining, in: Educational Data Mining 2014, Citeseer, 2014.
[40] S. D'Mello, A. Olney, N. Person, Mining collaborative patterns in tutorial dialogues, J. Educ. Data Min. 2 (1) (2010) 1–37.
[41] M.T. Chi, S.A. Siler, H. Jeong, T. Yamauchi, R.G. Hausmann, Learning from human tutoring, Cogn. Sci. 25 (4) (2001) 471–533.
[42] F.J. Rodríguez, K.M. Price, K.E. Boyer, Exploring the pair programming process: Characteristics of effective collaboration, in: Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, 2017, pp. 507–512.
[43] A.K. Vail, J.F. Grafsgaard, K.E. Boyer, E.N. Wiebe, J.C. Lester, Predicting learning from student affective response to tutor questions, in: ITS, Springer, 2016, pp. 154–164.
[44] A. Ezen-Can, K.E. Boyer, A tutorial dialogue system for real-time evaluation of unsupervised dialogue act classifiers: Exploring system outcomes, in: AIED, Springer, 2015, pp. 105–114.
[45] J.F. Grafsgaard, et al., Multimodal affect modeling in task-oriented tutorial dialogue., 2014.
[46] P. Robe, S.K. Kuttal, Y. Zhang, R. Bellamy, Can machine learning facilitate remote pair programming? Challenges, insights & implications, in: 2020 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC, IEEE, 2020, pp. 1–11.
[47] K.E. Boyer, R. Phillips, M.D. Wallis, M.A. Vouk, J.C. Lester, Learner characteristics and feedback in tutorial dialogue, in: Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, ACL, 2008, pp. 53–61.
[48] M.G. Core, J.D. Moore, C. Zinn, The role of initiative in tutorial dialogue, in: EACL, 2003.
[49] K. Forbes-Riley, D. Litman, A. Huettner, A. Ward, Dialogue-learning correlations in spoken dialogue tutoring, in: Proceedings of the 2005 Conference on AIED: Supporting Learning Through Intelligent and Socially Informed Technology, 2005, pp. 225–232.
[50] S. Katz, G. O'Donnell, H. Kay, An approach to analyzing the role and structure of reflective dialogue, IJAIED 11 (2000) 320–343, Part I of the Special Issue on Analysing Educational Dialogue Interaction.
[51] D.E. Meltzer, Relation Between Students ' Problem-Solving Performance and Representational Format, 2005.
[52] H. Muhonen, E. Pakarinen, A.-M. Poikkeus, M.-K. Lerkkanen, H. Rasku-Puttonen, Quality of educational dialogue and association with students' academic performance, Learn. Instr. 55 (2018) 67–79.
[53] F. Yang, F.W. Li, Study on student performance estimation, student progress analysis, and student potential prediction based on data mining, Comput. Educ. 123 (2018) 97–108.
[54] F.J. Dochy, G. Moerkerke, R. Martens, Integrating assessment, learning and instruction: Assessment of domain-specific and domaintranscending prior knowledge and progress, Studi. Educ. Eval. 22 (4) (1996) 309–339.
[55] A. Ezen-Can, J.F. Grafsgaard, J.C. Lester, K.E. Boyer, Classifying student dialogue acts with multimodal learning analytics, in: LAK, 2015, pp. 280–289.
[56] J.F. Grafsgaard, J.B. Wiggins, A.K. Vail, K.E. Boyer, E.N. Wiebe, J.C. Lester, The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring, in: ICMI, 2014, pp. 42–49.
[57] V. Rus, N. Maharjan, R. Banjade, Dialogue act classification in human-to-human tutorial dialogues, in: Innovations in Smart Learning, Springer, 2017, pp. 185–188.
[58] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in bertology: What we know about how bert works, Trans. Assoc. Comput. Linguist. 8 (2021) 842–866.
[59] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[60] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on EMNLP and the 9th International IJCNLP, 2019, pp. 3606–3611.
[61] H. Khanpour, N. Guntakandla, R. Nielsen, Dialogue act classification in domain-independent conversations using a deep recurrent neural network, in: Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2012–2021.
[62] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[63] G. Chen, D. Lang, R. Ferreira, D. Gasevic, Predictors of student satisfaction: A large-scale study of human-human online tutorial dialogues, in: EDM, 2019.
[64] T.G. Dietterich, et al., Ensemble learning, in: The Handbook of Brain Theory and Neural Networks, Vol. 2, MIT press Cambridge, MA, 2002, pp. 110–125.
[65] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL-Int. J. Appl. Linguist. 165 (2) (2014) 97–135.
[66] J. Lin, S. Pan, C.S. Lee, S. Oviatt, An explainable deep fusion network for affect recognition using physiological signals, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, ACM, 2019, pp. 2069–2072.
[67] V.J. Shute, Focus on formative feedback, Rev. Educ. Res. 78 (1) (2008) 153–189.
[68] E.R. Fyfe, B. Rittle-Johnson, M.S. DeCaro, The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters, J. Educ. Psychol. 104 (4) (2012) 1094.
[69] J. Van de Pol, M. Volman, J. Beishuizen, Scaffolding in teacher–student interaction: A decade of research, Educ. Psychol. Rev. 22 (3) (2010) 271–296.
[70] T. Hailikari, N. Katajavuori, S. Lindblom-Ylanne, The relevance of prior knowledge in learning and instructional design, Am. J. Pharm. Educ. 72 (5) (2008).
[71] S. Narciss, S. Sosnovsky, L. Schnaubert, E. Andrès, A. Eichelmann, G. Goguadze, E. Melis, Exploring feedback and student characteristics relevant for personalizing feedback strategies, Comput. Educ. 71 (2014) 56–76.
[72] S. Athey, G.W. Imbens, Machine learning methods for estimating heterogeneous causal effects, Stat 1050 (5) (2015) 1–26.
[73] Y. Luo, J. Peng, J. Ma, When causal inference meets deep learning, Nat. Mach. Intell. 2 (8) (2020) 426–427.



Jionghao Lin is a Ph.D. student in the Centre for Learning Analytics at Monash University, Melbourne, Australia. His primary research interests focus on the areas of learning analytics, natural language processing, and affective computing. Currently, Jionghao is mainly working on applying artificial intelligent technologies to understand and optimize the learning environment. He received his B.E. degree from Jianghan University, China, and Master degree in Data Science from Monash University, Australia.

Shaveen Singh is a Research Fellow at the Centre of Learning Analytics at Monash University. His research interests include the design and deployment of technology to increase the understanding and improve digital learning experiences. More specifically, his work examines the areas of learning analytics, personalized active learning, and building tools for teacher support. Shaveen is currently pursuing his Ph.D. at Monash University.

Lele Sha is a second-year Ph.D. student in the Centre for Learning Analytics at Monash University. His main research interest centres on applying Machine Learning and Natural Language Processing techniques to automatically processing educational forum posts. Specifically, he is focusing on improving model performance by applying extensive feature engineering and sentence embeddings. Before starting his Ph.D., Lele also worked in several software-as-a-service projects on learning management systems, which were successfully deployed to production and currently offering hundreds of online courses on its interactive training platform for Australian students.

Wei Tan is a Doctoral Researcher who studies the cutting-edge machine learning algorithm in Data Science. He specializes in Active Learning that optimize the labelling budget and time for the human annotator. His Ph.D. project is funded by Google Turning point. The aim is to develop the Surveillance System that will enable capture of a more complete set of coded ambulance data relating to SITB, mental health, and AOD attendances to inform policy, practice and intervention. He holds a master's degree from Monash University, and has expertise in analytics design for the social media platform.

David Lang is a doctoral student in the Economics of Education program and an IES Fellow. He graduated from UCLA in 2008 with a B.A. in Economics & a B.S. in Actuarial Mathematics. Prior to his doctoral studies, David worked for five years as a research analyst at the Federal Reserve Bank of San Francisco. His research interests include higher education, online education, and quantitative methods in education research. At Stanford, David also obtained a master's degree in Management Science and Engineering.

Dragan Gašević is Distinguished Professor of Learning Analytics in the Faculty of Information Technology and Director of the Centre for Learning Analytics at Monash University. As the past president (2015–2017) and a co-founder of the Society for Learning Analytics Research, he had the pleasure to serve as a founding program chair of the International Conference on Learning Analytics and Knowledge (LAK) and a founding editor of the Journal of Learning Analytics. His research centres on self-regulated and social learning, higher education policy, and data mining. He is a frequent keynote speaker and a (co-)author of numerous research papers and books.

Dr. Guanliang Chen is serving as a Lecturer in the Faculty of Information Technology, Monash University in Melbourne, Australia. Before joining Monash University, Guanliang obtained his Ph.D. degree at the Delft University of Technology in the Netherlands, where he focused on the research on large-scale learning analytics with a particular focus on the setting of Massive Open Online Courses. Currently, Guanliang is mainly working on applying novel language technologies to build intelligent educational applications. His research works have been published in international journals and conferences including AIED, EDM, LAK, L@S, EC-TEL, ICWSM, UMAP, Web Science, Computers & Education, and IEEE Transactions on Learning Technologies. Besides, he co-organized two international workshops and has been invited to serve as the program committee member for international conferences such as LAK, FAT, ICWL, etc.

3.3 Chapter Summary

Educational researchers have long sought to build dialogue-based ITS that can provide personalised and effective learning experiences to learners [111]. Many prior studies used dialogue acts to analyse communicative patterns from the tutoring dialogues [44, 87, 112]. However, the understanding of delivering effective tutoring is still limited (Gap 1). The existing studies [44, 87, 112] have focused on effective tutorial sessions and have not thoroughly examined ineffective ones to identify the reasons for their failure. Therefore, it is necessary to investigate the role of dialogue acts in the tutoring dialogues to understand what instructional strategies lead to the failure of the tutoring session (RQ2). Driven by this gap, in Chapter 3, we have answered RQ2 by analysing a large corpus of over 15,000 online human-human tutorial sessions, both effective and ineffective, in order to identify common tutoring strategies and use them to predict learners' problem-solving performance.

Through our analysis and evaluations, we identified a set of different communicative patterns that were relevant to educators and learners in different types of tutorial sessions. Our findings have implications for providing guidance to novice educators to consider the use of instructional strategies in tutoring (as described in **Contribution 3**). Additionally, we discovered that the dialogue acts (taken by learners and educators during tutorial sessions) alone were not enough to accurately predict learners' problem-solving performance, and should be considered in conjunction with other factors such as the informativeness of the statements made during the session (as described in **Contribution 4**).

Contribution 3: Instructional strategies used in the tutoring

- Educators often took actions to provide feedback and information to learners and to raise questions to guide learners to solve problems. Correspondingly, learners took more actions in answering questions or expressing agreement with educators and acceptance of the provided explanations or solutions.
- In tutorial sessions where learners delivered correct or partially correct solutions to the problems, educators tended to ask more thought-provoking questions, offered more information hints, and posed less irrelevant statements or questions to learners compared to tutorial sessions with lower problem-solving performance.
- If a learner had made certain progress in solving a problem before entering a tutorial session, the learner was likely to receive fewer hints and thought-provoking questions from an educator but still had a higher chance to successfully solve the problem and receive more positive feedback from the educator.
- Positive feedback expressed by educators can be used as a strong discriminator to differentiate tutoring dialogues of different levels of problem-solving performance.

Contribution 4: The capability of dialogue act features in predicting learner problem-solving performance

- Dialogue-act-related features (dialogue acts and dialogue act patterns) were insufficient to reveal the problem-solving performance of learners and should be utilised together with other relevant factors (e.g., the informativeness, complexity, and the sentimental polarities of the utterance). By training the GTB model on the feature set including dialogue-act-related features and other features, the GTB model demonstrated its potential to identify potentially unsuccessful dialogues before the end of the tutoring dialogue sessions. With these potentially unsuccessful tutoring dialogues detected during the dialogue sessions, we expect that more interventions (e.g., changing an instructional strategy) can be taken to better engage and assist learners.
- We released the dialogue act classifier, which was constructed by applying the state-of-the-art pre-trained language model BERT, to better support researchers for relevant research studies, which is accessible via https://github.com/bertDA/BertDA.

Chapter 4

Enhancing the Identification of Instructional Strategies from Tutoring Dialogues

4.1 Introduction

In Chapter 3, we developed a dialogue act classifier, which obtained promising classification performance (F1-score of 0.74 and Cohen's κ of 0.74) in identifying fine-grained level dialogue acts (i.e., classify 31 dialogue acts). However, we also identified two main limitations (i.e., insufficient efforts on exploring the impacts of discourse contextual information and labour-intensive issues in annotating dialogue acts) about the dialogue act classification. We deemed that the performance of the dialogue act classifier could be further improved to become comparably accurate as the human annotation [38] (i.e., Gap 2 described in Chapter 1). Firstly, inspired by the dialogue act classification in general human communication [93], the discourse contextual information (e.g., the content of preceding utterances from interlocutors) could be included as the input for training the dialogue act classifier which could improve the performance of dialogue act classification. However, the dialogue act scheme designed for general human communication might not fit well with the annotation of the educational dialogue [96]. Therefore, it is still unknown how much contextual information should be incorporated for enhancing the educational dialogue act classification (RQ 3.1). Secondly, training an effective supervised machine learning model was non-trivial, especially for the data-hungry models (e.g., deep learning models), which require a large amount of annotated data to make an accurate prediction. However, annotating data for dialogue act classification commonly needs much manual work and financial expenses which can constrain the process of training the dialogue act classifier. Driven by

this, several prior studies [90, 113, 114] suggested employing statistical active learning methods, which could sample the representative data from the dataset for human annotation. Relying on the statistical active learning methods, human annotators could annotate fewer amounts of samples to obtain an effective supervised machine learning model for dialogue act classification [114]. However, the application of statistical active learning methods in educational dialogue act classification was relatively limited and under-explored (RQ 3.2).

Driven by the concerns raised in Chapter 3, we were motivated to explore the discourse context and investigate the recent active learning methods to enhance the dialogue act classification performance in Chapter 4. Specifically, we first explored the extent to which the incorporation of discourse contextual information could boost a model's prediction performance (RQ3.1). Then, we investigated the potential of the recent statistical active learning methods on the alleviation of the labour-intensive issue (i.e., high demand for annotating labels to train the ML model) for recognising instructional strategies (RQ3.2).

The results of this study have been submitted to the *IEEE Transactions on Learning Technologies*.

- Lin, J., Tan, W., Du, L., Buntine, W., Lang, D., Gašević, D., & Chen, G. (2023). Enhancing Educational Dialogue Act Classification with Discourse Context and Sample Informativeness. *IEEE Transactions on Learning Technologies* (Under Review).

4.2 Publication: Enhancing Educational Dialogue Act Classification with Discourse Context and Sample Informativeness.

Enhancing Educational Dialogue Act Classification with Discourse Context and Sample Informativeness

Jionghao Lin, Wei Tan, Lan Du, Wray Buntine, David Lang, Dragan Gašević, and Guanliang Chen

***Abstract*—Automating the recognition of instructional strategies from a large-scale online tutorial dialogue corpus is indispensable to the design of dialogue-based intelligent tutoring systems. Despite many studies automating the recognition process by machine learning (ML) models, training a well-performed ML model is non-trivial since the sample size is commonly limited for model training. Therefore, our study aimed to investigate the improvement of automating the recognition of dialogue acts, a popular approach to the detection of instructional strategies, from two perspectives. Firstly, we explored whether and to what extent the incorporation of discourse contextual information can boost a model's prediction performance. Then, we investigated the extent to which the recent active (machine) learning methods can alleviate the labour-intensive issues (i.e., high demand of manual annotation) in training the ML model for recognising the instructional strategies. Our study showed that: 1) the ML models trained on the features that included the discourse context achieved better performance than the models excluding it; 2) the effectiveness of the contextual information decayed after the ML model achieved an optimal performance; and 3) compared with the random baseline, active learning methods can select informative samples from the training dataset to train ML models, which can alleviate the labour-intensive issues.**

***Index Terms*—Educational Dialogue Analysis, Dialogue Acts, Intelligent Tutoring Systems, Active Learning, Discourse Context**

I. Introduction

ONE-ON-ONE human tutoring has been widely acknowledged as an effective form of instruction in guiding students across various educational backgrounds [1], [2]. In the tutoring process, a tutor can guide a student by employing a set of instructional strategies (i.e., the approaches or principles used by tutors to guide students toward the expected achievements), such as asking students thought-provoking questions and providing corrective feedback [3]. Traditionally, a tutor instructs a student in a face-to-face tutoring session. Since the outbreak of the COVID pandemic, most face-to-face tutoring sessions have transitioned to online one-on-one tutoring as the online tutoring environment can allow tutors to guide students remotely [4]. However, online one-on-one human tutoring has been considered a time and labour demanding instructional method, especially for the high-enrolment courses where the limited number of tutors need to provide instruction or feedback to a large number of students [5]. In this space, researchers have been working on the development of a dialogue-based intelligent tutoring system (i.e., a computer-based instructional system guiding students in a learning process via conversation) by investigating the effective instructional strategies from tutors in the real-world tutoring sessions [6], [7], [8], [9]. A typical example of a dialogue-based Intelligent Tutoring System (ITS) is AutoTutor [10], which complied the instructional strategies observed by human tutors into the computer system. Despite the existing dialogue-based ITSs having assisted millions of students, they still need to be improved to be effective as human tutors [11], [6], [12]. One potential improvement is that researchers can develop the dialogue-based ITS by analysing the instructional strategies from the real-world human tutoring sessions [13], [14]. Therefore, identifying the instructional strategies from one-on-one human tutoring is important to shed light on the guideline of effective tutoring practice and also the design of dialogue-based ITS.

Informed by existing studies [15], [7], [9], the instructional strategies can be identified in the form of dialogue acts (DAs) which reflect the intention behind the conversational utterances in the tutoring process. As tutors' instructional strategies were employed based on the students' actions (e.g., requesting hints), the students' DAs should also be considered in the analysis process. To identify the DAs from human conversation, researchers commonly relied on the existing DA annotating schemes (e.g., DAMSL [16], DATE [17], and ISO 24617-2 standard [18]), which defined the purpose and rules of annotating DAs. It should be noted that different types of DA schemes present different purposes [19]. For example, the ISO 24617-2 standard scheme can be used to annotate the general human communication [18]. However, according to Bunt *et al.* [20], the ISO 24617-2 standard scheme might not fit well with the purpose of annotating the educational dialogue. Therefore, to analyse instructional strategies, it is necessary to analyse the DAs built upon the educational DA scheme.

Previous studies employed educational DA schemes to investigate the effective instructional strategies in the online tutoring process [15], [21]. These works analysed the effective instructional strategies from the online tutoring dialogue based

J. Lin was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: Jionghao.Lin1@monash.edu.

W. Tan was with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: Wei.Tan2@monash.edu.

L. Du was with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: Lan.Du@monash.edu.

W. Buntine was with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: Wray.Buntine@monash.edu.

D. Lang was with the Graduate School of Education, Stanford University, United States, e-mail: dnlang86@standford.edu.

D. Gašević was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: dragan.gasevic@monash.edu.

G. Chen was with the Centre for Learning Analytics, Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia, e-mail: guanliang.chen@monash.edu.

on the limited number of tutors and students, which might not fully demonstrate the effectiveness of the instructional strategies [15], [21]. Instead, a recent research trend is to analyse the instructional strategies in the form of DAs from a large amount of tutor-student dialogue datasets. Previous studies employed the traditional supervised machine learning models (e.g., Support Vector Machine, Naive Bayes, and Deep Learning Models) to automatically identify the DAs from tutors and students and further conducted an analysis based on the identified results [22], [23], [24], [7], [9]. These studies have thereby contributed many insightful results to the delivery of effective instructional strategies. However, many studies concluded that the performance of DAs recognition was constrained by the limited sample size [9], [22], [25] and the annotation process is commonly time-consuming and cost-demanding. Based on this fact, our study aimed to optimise the process of DA recognition on the limited annotated samples.

The study presented in this paper aimed to address open research questions that are noted as critical to be invested in the literature for the effective process of automating educational DA recognition [9]. Firstly, the number of incorporated content-based contextual information (e.g., content of preceding utterances from interlocutors) could be further explored for the accuracy performance improvement of DAs recognition as a response in the dialogue often depends on the context from preceding utterances. However, it is still unknown how much contextual information should be incorporated for enhancing the DA recognition performance. Secondly, training an effective supervised machine learning model is non-trivial, especially for the data-hungry models (e.g., deep learning models), which required a large amount of annotated data to make an accurate prediction. However, annotating dataset for DA recognition commonly needs much manual work and financial expense which limits the need for scaling up the sample size and further constrains the process of recognising the effective instructional strategies in online one-on-one tutoring. To obtain a well-performed ML model on limited samples, researchers have begun employing active learning (AL) methods, which can sample the representative data from the dataset for human annotation [26]. Existing research has shown that AL methods can alleviate the demand of human annotation to build a classifier on textual dataset [27], [28]. Relying on the AL methods, human annotators can annotate fewer amounts of samples to obtain an effective supervised machine learning model for dialogue act classification [26]. However, the application of AL methods in the education in general and dialogue analysis in particular has been relatively limited, so the efficacy of the AL methods on supporting educational dialogue analysis is still under-explored. To unlock the potentials of machine learning models on classifying the educational dialogue acts, it is worth investigating the extent to which the existing active learning methods can not only alleviate the labour-intensive issues but also maintain the effectiveness of the machine learning model.

To this end, our study aimed to investigate the extent to which the discourse contextual information relates to performance in automatic classification of DAs and evaluate the extent to which the state-of-the-art active learning are effective in deep learning models to minimise the annotation budget for labelling the model training dataset. Specifically, our study aimed to answer the following **R**esearch **Q**uestions:

RQ 1 To what extent does the discourse context relate to the performance of automatic classification of educational dialogue acts?

RQ 2 To what extent can widely-used active learning methods alleviate the labour-intensive issues in automatic classification of educational dialogue acts?

II. Related Work

A. Educational Dialogue Act Classification

Analysing the instructional strategies can contribute to the tutoring practice and the development of dialogue-based intelligent tutoring systems. As the initial stage of discovering instructional strategies, most empirical works developed a coding scheme to annotate the communicative intention behind the utterances from tutors and students (i.e., Dialogue Acts). For example, a tutor's utterance "*Well done!*" can be characterised as the dialogue act `Positive Feedback`. However, manually annotating DAs for large-scale datasets is time-consuming and cost-demanding [29]. To this end, a recent trend is to annotate a certain amount of training dataset and then use supervised machine learning methods to train on the annotated dataset so as to automate the annotating process [29]. To date, many studies in classifying educational DAs have been based on supervised machine learning methods [30], [31], [32], [25], [7], [33], [9]. Boyer *et al.* [30] trained a Logistic Regression model on the lexical and syntactic features (e.g., part of speech taggers and N-grams) and structural features (e.g., hidden dialogue state predictions from Hidden Markov Models). These features were extracted from the annotated utterances in 48 programming tutoring sessions. Samei *et al.* [31] employed Decision Trees and Naive Bayes model to train on 210 annotated tutoring utterances on the science-related topic. They defined two feature sets, including semantic features (e.g., first two and last words of current utterance) and contextual features (e.g., dialogue acts and speaker's role of previous five utterances). Samei *et al.* [32] trained Naive Bayes, Bayesian Networks, Logistic Regression, and Conditional Random Fields on 1,438 annotated tutoring sessions on Algebra and Physics topics. Similar to feature engineering in their previous work [31], they used the first three words for each current utterance and first and last words of the previous three utterances as the input features. Rus *et al.* [25] found that Bayes Nets achieved optimal performance by using the N-gram features on the dataset of 222 online Algebra tutoring sessions. Later on, they employed Conditional Random Field models to train on 500 Algebra tutorial sessions and used more input features, including the leading three words, last words of current and previous utterances, utterance length, N-grams, and dialogue acts of previous utterances [7]. Stasaski *et al.* [33] employed the Long Short Term Memory DL model to train on 741 tutoring exercises and used the GloVe embedding to vectorise the textual content of current utterance as the model's input.

Many studies, as discussed above, have incorporated contextual information into the input feature set [31], [32], [7]. These studies typically used either the DAs from previous utterances as the contextual features (e.g., [31], [7]) or the limited number of words from previous utterances as the features (e.g., [32]) based on the Correlation Feature Selection (CFS) measure. Although these studies successfully trained the DA classifiers, we argue that it is also important to investigate DA classifiers trained on more contents of previous utterances as more information can be provided for machine learning models to determine the labels. In line with our argument, Lin *et al.* [9] employed the BERT model to train a classifier on 50 tutoring sessions, and they used the role of the utterances (i.e., tutor or student), the content of current utterance and content of one previous utterance as the input to the model. However, they draw the conclusion that the deep learning model for DA classification can be further improved by incorporating more content from previous utterances. Inspired by Ribeiro *et al.* [34], they incorporated contextual information from 1 to 5 previous utterances into the input feature and evaluated the effect of contextual information on the performance of the DA classifier. However, as the educational DA annotation is domain-specific [20], it is still unknown to what extent the contextual information can be used to improve the performance of educational DA classification.

B. Active Learning

Though deep learning models have shown considerable success in recent studies of educational DA classification [33], [9], the high demand for the annotated dataset is still an issue for deploying deep learning models. To alleviate this issue, researchers proposed active learning methods to support the deep learning model training process. Active learning (AL) methods can select the representative data samples for human annotation, and AL methods have been built for reducing the annotation budget for labelling the data [35]. Traditionally, there are three typical scenarios of AL methods: *membership query synthesis*, *stream-based sampling*, and *pool-based sampling* [35].

- *Membership query synthesis* can generate artificial data-point for labelling rather than sampling the data-point from the real-world data distribution [35].
- *Stream-based sampling* scans through a sequential stream of unlabelled data-point and make sampling query decision individually [35].
- *Pool-based sampling* selects the most informative samples from the unlabelled data pool and send them to the oracle (e.g., human annotator) for annotation [35].

As the annotated dataset was available in our study and the dataset was not collected in a sequential stream, we considered our study fitting well with the scenario of pool-based sampling. Therefore, we mainly introduced the pool-based scenario in this section. Pool-based sampling scenario have shown great success in recent studies [36], [37], [38], [39] where researchers used a small number of representative examples as a surrogate for the entire dataset to train machine learning models. Theoretically, the pool-based AL methods can both reduce the computational cost of model training and maintain the performance of the model trained on the labelled dataset [35].

Many recent studies achieved promising results in training the machine learning models based on the pool-based AL methods. For example, Yuan *et al.* [38] applied pool-based AL methods with the BERT model to the sentence classification tasks by using many large-scale datasets, including news articles (AG NEWS [40]), sentiment reviews (IMDB [41] and SST-2 [42]), and medical abstracts (PUBMED [43]). They found that the AL methods BADGE and ALPS used less than 10% labelled dataset, achieving closed performance as compared to the model training on the full dataset. Additionally, both AL methods significantly outperformed the random sampling methods. Zhao *et al.* [37] examined the efficacy of pool-based AL methods with the logistic regression model on a small dataset (i.e., UCI User Knowledge [44]). They found that the AL method WMOCU maintained the model's classification accuracy performance with less amount of labelled data and significantly outperformed other sampling methods. Tan *et al.* [39] employed the pool-based AL methods with the DistilBERT model on the task of sentence classification. They evaluated the efficacy of the AL method on the same datasets as done in the work by [38]. Their results demonstrated that the AL method BEMPS outperformed ALPS, BADGE and WMOCU on text classification. Though the efficacy of AL methods has been demonstrated in many recent works, the use of AL methods has rarely been investigated in classification of educational DAs. To fill the gap, our study aimed to examine these recent AL methods in classification of educational DAs.

III. Methods

A. Dataset

TABLE I: The descriptive statistics of the dataset used in the study.

Metric	**All dataet**
1. # Total sessions:	14,562
2. # Total uttrances:	1,216,784
3. # Tutors:	116
4. # Students:	5,165
5. Avg. Session duration (mins):	30.27 $\pm$ 30.66
6. Avg. # Utterances per session:	83.56 $\pm$ 81.05
7. Avg. # Words per session:	647.75 $\pm$ 596.12
8. Avg. % Utterances by tutors:	58.42 $\pm$ 7.86
9. Avg. % Words by tutors:	78.36 $\pm$ 9.10

We obtained the ethics approval from Monash University for the secondary data use (Project ID 26156). The dataset was collected via a mobile phone application and provided by an educational technology company. The company collected the data along with the informed consent from students and tutors allowing the use of the de-identified data for research. By using the mobile phone application, tutors and students can work collaboratively to solve problems in various subjects such as mathematics, physics, and chemistry. In the tutoring process, a student commonly took a picture of an unsolved

problem and sent the picture to the mobile application for requesting help from tutors. Then, the application allocated an experienced tutor with the students to solve the problem by using textual message and images to communicate. It should be noted that the tutors in the educational technology company were trained to give their best instruction to assist students in solving the problem instead of sharing the direct solutions with the students. Hence, the tutorial dialogues in our dataset contained the detailed processes of how tutors scaffold students to solve the problems. In total, the dataset contained 14,562 tutorial dialogues and 92% of these tutorials were related to mathematics tutoring. The details of the dataset are shown in Table I. Additionally, we provided a sample dialogues in the appendix.

B. Educational Dialogue Act Scheme and Annotation

To identify the DAs in tutorial dialogues, existing studies often rely on the educational DA coding scheme [45], [46], [47]. Following the work by [9], we employed the same online tutoring dialogue act scheme built on the work by [15] whose effectiveness in portraying online one-on-one tutoring has been documented in many previous works (e.g., [24], [48], [49]). We characterised the intention behind the tutor-student utterances by using the DA scheme with a two-level structure [15]. As shown in Table II, there were 12 first-level DAs and 31 second-level DAs. The first-level DA tags were used to annotate a broad view of tutor' and students' interaction (e.g., tutors answered students' questions, which was annotated as `Answer`). In comparison, the second-level DA tags were used to discover more fine-grained level information from tutor-student conversation (e.g., tutors used "*Yes*" or "*No*" as the part of the answer, which could be annotated as `Yes-No Answer`). It should be noted that some DAs are specific to the role of tutor or student. For example, the DA tag `Evaluation Question` could only be used by tutors to assess students' understanding, while the tag `Confirmation Question` could only be used by students to seek feedback. Additionally, some DAs could be used for both tutors and students, such as `Correction` (correcting the typos in tutors' or students' previous utterance) and `Acknowledge` (expressing acknowledgement to their interlocutors).

As manually annotating the DAs for the large-scaled dialogue dataset is quite time-consuming and cost-demanding, many existing studies labelled a subset of the whole dataset and used this labelled dataset to train a machine learning classifier to automate the DA annotating process [9], [22], [32]. In our work, we first randomly sampled 50 tutorial dialogues (contained 3,629 utterances from tutors and students) from the original dataset. Then, we pre-processed the dialogue dataset by breaking the utterances into a sentence (an utterance often contained more than one sentence and different sentences might present different DAs) and removing the utterances presenting emoji or symbols. Next, we recruited two human coders to annotate DAs for these sampled tutorial dialogues. Cohen's k for the annotation was 0.77, which indicated a substantial agreement between the two coders. The inconsistent cases were further resolved by inviting a third educational expert to discuss with the two coders.

C. Identifying Educational Dialogue Acts

To answer RQ1, we aimed to explore the contextual information and to evaluate a widely-used supervised machine learning model to classify the educational dialogue acts. Many existing studies in DA classification suggested that incorporating the contextual information into the training process of a model can improve the DA classification accuracy [34], [50], [51], [52]. It should be noted that these existing studies termed the sentence being classified as the current segment and preceding utterances of the current segment as preceding segments (i.e., contextual information). To have a clear understanding of the current segment and preceding segments, we used an annotated sample dialogue as an example (shown in Table III). We assume that a DA classifier identified the DA for the sentence in Row 6 (Table III), which is called the current segment. By considering the preceding segments (i.e., contextual information), the preceding utterances in rows 1-5 would be incorporated with the current segment as the input for the DA classification. It should be noted that Ribeiro *et al.* [34] examined the number of preceding segments from 0 to 5 on the task of DA classification for human general communication dialogue dataset and found that the DA classifier trained on the input with the incorporation of preceding segments (i.e., incorporating the preceding segments from 1 to 5) achieved higher classification accuracy than without. Furthermore, many previous studies found that incorporating one preceding segment in the input was most important to the DA classification and that influence decayed with the increasing number of preceding segments [34], [50], [51]. However, as the general human communication DA scheme might not fit well with the DA annotation of educational dialogue [20], we posit that the influence of the preceding segments in educational DA classification might be varied. To this end, the current study aimed to investigate the preceding segments from 0 to 5 on the tutorial dialogue dataset to observe the impact of contextual information on educational DA classification.

To improve the accuracy of DA classification, we also investigated the wide-used supervised machine learning model. Inspired by the success of the educational DA classification from the work [9], our current study aimed to conduct the DA classification by using the same model, which was the BERT pre-trained language model. Additionally, to embrace the **Green Artificial Intelligence (AI)** concept [53] – i.e., the machine learning study should not only yield novel results but also take the model's computational cost into account so as to reduce the carbon footprint to the human living environment – the current study also selected the ELECTRA model as the comparison. Similar to the BERT model, ELECTRA was a recently developed language model for generating textual data representation, but ELECTRA was pre-trained on a more efficient approach called replaced token detection as compared to the masked language modelling in BERT [54]. It should be noted that the ELECTRA model was demonstrated outperforming BERT on many classification tasks and requiring

TABLE II: The DA scheme used in [9]. The column **Role** indicates whether a DA is only specific to students (S), tutors (T), or both (S&T).

First-level DA	Second-level DA	Role	Sample Utterances
Acknowledge	Acknowledge	S&T	*"Ok!"*
Correction	Correction	S&T	*"*divide"*
Directive	Directive	T	*"Hold on just a moment"*
Hint	Observation	S&T	*"We have 4i-5(-1)."*
	Hint by Image Information	T	*[Image]* *"Now, we multiply both sides by 10."*
Lukewarm Feedback	Lukewarm Feedback	T	*"You're so close!"*
Negative Feedback	Negative Feedback	T	*"Not quite!"*
Positive Feedback	General Positive Feedback Elaborated Positive Feedback	T	*"Great work!"* *"Yes, your third term is correct"*
Request Confirmation	Evaluation Question	T	*"Were you able to understand that?"*
Request Feedback	Request Feedback by Image Confirmation Question	S	*[Image]* *"Are these the correct points ?"*
Answer	Extra Domain Answer WH Answer Yes-No Answer	S&T	*"I'm doing good"* *"I got 64"* *"Yes, I would love to help"*
	Ready Answer	S	*"Yep, ready to go."*
Question	Extra Domain Question	S&T	*"How's it going today?"*
	Information Question Direction Question	S	*"What do you mean to change it?"* *"Then, what's the second?"*
	Factual Question Open Question Operational Question Probing Question Ready Question	T	*"Did you find the measure of angle Y?"* *"What would be our next step?"* *"May I know where you are stuck?"* *"Can those be 0?"* *"Are you ready?"*
Statement	Explanation Extra Domain Other Greeting	S&T	*"Because B = A"* *"The photo is still sending"* *"Hi!"*
	Understanding Not Understanding	S	*"Oh, I see it"* *"No, I don't know."*
	Reassurance	T	*"Let's work with this problem together!"*

TABLE III: An example of the annotated tutorial dialogue. Here are the abbreviation of the DA tags: Oprt-Ques (Operational Question), Y-N-Ansr (Yes-No Answer), Req-Fdbk (Request Feedback), Conf-Ques (Confirmation Question), Pos-Fdbk (Positive Feedback), and G-Pos-Fdbk (General Positive Feedback).

Row	Dialogue	1st lv. DA	2nd lv. DA
1	**T:** Hi, have you tried any work on it?	Question	Oprt-Ques
2	**S:** No idea	Answer	Y-N-Ansr
3	**T:** No worries, I will help you	Statement	Reassurance
4	**T:** So, let us first find the value of x	Hint	Information
5	**S:** Would the value of x be 50?	Req-Fdbk	Conf-Ques
6	**T:** Yes, correct!	?	?

less computing power in the training process [54], which was in line with the idea of **Green AI** [53]. Therefore, the current work evaluated the performance of educational DA classification by both the BERT and ELECTRA models.

D. Active Learning on the Educational Dialogue Act Classification

To answer RQ2, we aimed to investigate the extent to which AL methods can alleviate the labour-intensive issue. In the recent work [39], they developed two AL methods, namely CoreMSE and CoreLOG, which achieved state-of-the-art performance on many textual data classification tasks. The current study included both methods to evaluate their efficacy on the educational DA classification task. Additionally, we also investigated the recent pool-based AL methods which are

TABLE IV: Datasets and the used language model

DA level	#Class	Labelled size	Test size	Lang. Model	Initial labelled size
First level	12	3763	476	ELECTRA	50
Second level	31	3763	476	ELECTRA	50

most related to the CoreMSE and CoreLOG AL methods as the comparison, including MOCU [37], WMOCU [37], BADGE [36], ALPS [38] and the random baseline. Following their originally published algorithms, we re-implemented them in the current study.

MOCU built upon the idea of the metric Mean Objective Cost of Uncertainty which can be used to quantify the influence of uncertainty on classification error [37].

WMOCU is the modified version of the MOCU-based method. The WMOCU incorporated the weights with the metric of MOCU and the study results demonstrated that the WMOCU method achieved better performance in decreasing the classification error and reducing the influence of uncertainty on classification performance than the baseline methods (e.g., Random, MES, and ELR) [37].

BADGE is the method to sample a group of diverse and uncertain points to train machine learning models. The study by [36] demonstrated that the BADGE method generally outperformed other active learning methods (e.g., Entropy, ALBL, and Coreset) on three image datasets and four non-image datasets .

ALPS can select the initial informative samples by using masked language modelling loss, which avoids the issue of randomly selecting the initial sample. This method generally outperformed other AL baseline model (e.g., Entropy, Random, and BADGE) in accuracy and efficiency in many text classification tasks (e.g., review sentiment classification) [38],.

CoreMSE and CoreLOG are the methods of Bayesian estimate of Mean Proper Scores (BEMPS) that enhance the diversity of acquired samples. They are the two variations of arbitrary strictly convex functions to quantify the classification error. CoreMSE uses a mean square error scoring rule, known as a Brier score, and CoreLOG uses a logarithmic scoring rule. The results show that the use of these methods consistently outperform many AL methods in text classification [39].

E. Study Experiments

Datasets We used our annotated dialogue dataset for two different classification tasks: first level task and second-level task, as shown in Table IV. The first level task contained 4K samples of 12 imbalanced classes. The second level task contained 4K samples of 31 imbalanced classes.

Study Setup for RQ1 We implemented the classifier by using the BERT [55] and ELECTRA [54] pre-trained language models. For classifying DA, the labelled sentences were randomly split to *training*, *validation*, and *testing* datasets in the ratio of 80%:10%:10%. We set the maximum sequence length to 128 and fine-tuned on a batch size of 16 for 30 epochs. The AdamW optimizer was used with the learning rate of 2e-5 to optimize the training of the classifier. All experiments were implemented on RTX 3090 and Intel Core i9 CPU processor.

Study Setup for RQ2 We used the optimal DA classifier from RQ1 as the backbone classifier in the pool-based AL experiments for both first-level and second-level tasks. The sample size was set to 50 for the labelled pool of the initial model training. Each AL method was run for five times with different random seeds. The batch size was set to 100. All experiments were run under the same setting described in **Study Setup for RQ1**.

F. Results on RQ1

Evaluation Metrics. For both of the two classification tasks above, we adopted three representative metrics for measuring the competency of the classification models, i.e., Area Under the Curve (AUC), F1 score, and Cohen's κ coefficient (Cohen's κ). We also present the results of classification accuracy as a reference. For the AL performance comparison, we used the learning curve for all the AL methods based on the acquired sample size and F1 score.

G. Batch Analysis by Data Maps

It should be noted that not all instances from the training set equally contribute to the training process of a ML model [56]. Some instances are easy to learn for ML models and these easy-to-learn instances might be redundant when exceeding amount of these instances are allocated for model training [57]. In comparison, the informative instances are commonly hard to learn which should be frequently sampled for model training [57]. As the pool-based AL methods select the training samples based on their informativeness, our study also investigated the extent to which AL methods can sample informative instances from the training dataset for model training by *Data Maps*. The Data Maps is a model-specific tool to inform the difficulty level of machine learning models to learn from the training instances [57]. With the use of Data Maps, each training instance can be assigned a difficulty level score from 0 to 1 and these levels can be further categorised into **Easy** ([0.00, 0.25)), **Medium** ([0.25, 0.50)), **Hard** ([0.50, 0.75)), and **Impossible** ([0.75, 1.00)) [58]. The difficulty levels can be used to represent the informative level of an instance that ML models make the prediction [57].

IV. Results

A. Results on RQ1

For RQ1, we investigated the extent to which the discourse context correlated with the model performance. Figure 1 shows the classification performance of the models evaluated according to four metrics (i.e., classification accuracy, F1 score, Cohen's $kappa$, and AUC). We compared the performance of the ELECTRA and BERT models on two levels of the DA classification task. The solid line in Figure 1 represents the first level DA classification performance whereas the dash line

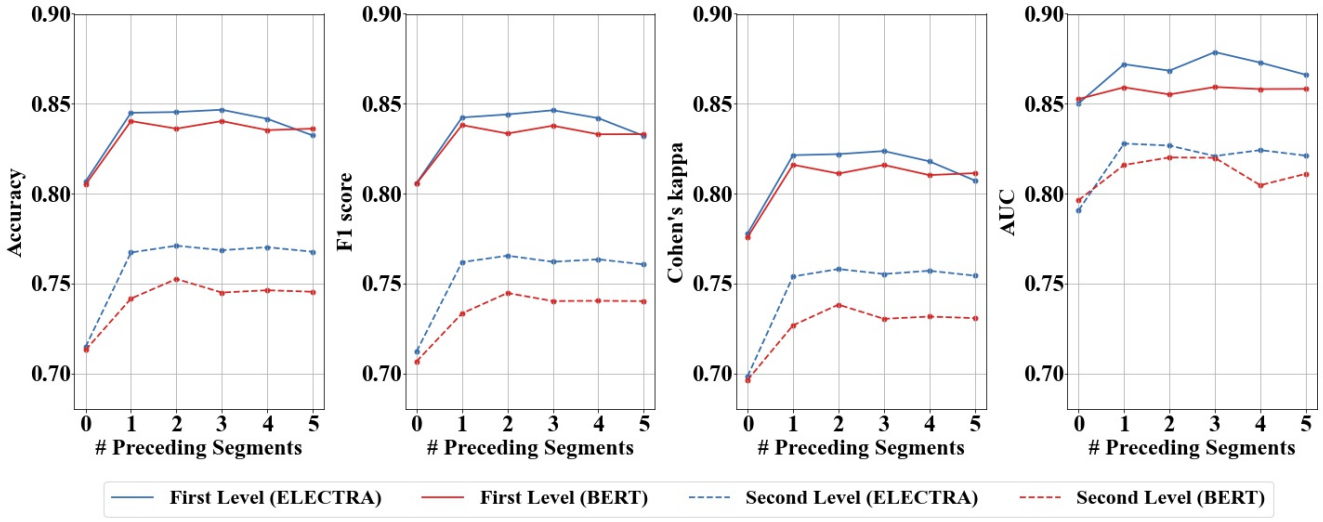


Fig. 1: The models' performance of the first- and second-level dialogue act classification. The red and blue lines represents the performance of BERT and ELECTRA model, respectively. The solid and dash line represents first- and second-level DA classification, respectively.

represents the second level. For each metric in Figure 1, we can see the changes of model performance as the increasing number of preceding segments. We used the averaged values of five experimental results for each line in Figure 1 to reduce the random biased.

In the first level DA classification (solid lines, shown in Figure 1), there were 12 first-level DAs for classification. The ELECTRA model (i.e., solid blue line) generally outperformed the BERT model (i.e., solid red line) across four metrics from 0 to 4 preceding segments. Both models performed worst when the models only used the current segment (i.e., the number of preceding segments equalled to 0). The results indicate that the model performance could be better when including the contextual information than excluding it. It should be noted that the ELECTRA model achieved the best performance when the model incorporated three preceding segments for the first level DA classification. However, by incorporating more than three preceding segments for model training, we found that the model performance decreased. The model performance decayed as the more preceding segments were incorporated after the optimal stage which was in line with the results by [34].

In the second level DA classification (dash lines, shown in Figure 1), there were 31 DA tags for classification, which was more complicated than the first level DA classification. It should be noted that the differences in classification performance between the two models were more observable on the second level DA classification as compared to the first level. The ELECTRA model generally outperformed than the BERT model on the second level DA classification. Additionally, we found that the performance of the ELECTRA model achieved the best classification performance when incorporating two preceding segments. Notably, the more complicated DA classification required fewer preceding segments to reach optimal performance.

B. Results on RQ2

For RQ2, we aimed to investigate to what extent the widely-used active learning methods can alleviate the labour-intensive issues in the educational DA classification tasks. Based on the results from RQ1, we found that the ELECTRA model generally outperformed than BERT model on first- and second-level DA classification. Notably, the Electra model achieved the best performance on first- and second-level DA classification by incorporating three and two preceding segments, respectively. Therefore, we investigated the efficacy of active learning methods based on the best performing ELECTRA models reported in the results of RQ1. In line with the work by [39], our current study investigated the metrics of classification accuracy and F1 score on the performance of the ELECTRA model to evaluate the efficacy of AL methods for each sampling batch. Based on the summary of stopping criterion for active learning [59], we only plotted efficacy of selected AL methods until the result where the best performed AL method reached our desired performance, i.e., 95% of the performance of the model trained on the full training dataset.

Figure 2 shows the efficacy of the seven active learning methods (described in Sec. III-D) on the task of first-level DA classification. The dash line on the top of the Figure 2 represents the performance of the ELECTRA model trained on the full training dataset where the ELECTRA model achieved 0.847 accuracy and 0.846 F1 score. In Figure 2, all the methods demonstrated an increasing pattern as more annotated instances were incorporated. We observed that the CoreLOG method outperformed other methods after acquiring 500 samples. When acquiring 1,400 samples, the ELECTRA model with the support of the CoreLOG method reached our desired performance where the model achieved the performance of 0.819 accuracy and 0.814 F1 score.

The results of the seven active learning methods on the task of second-level DA classification are shown in Figure 3. The

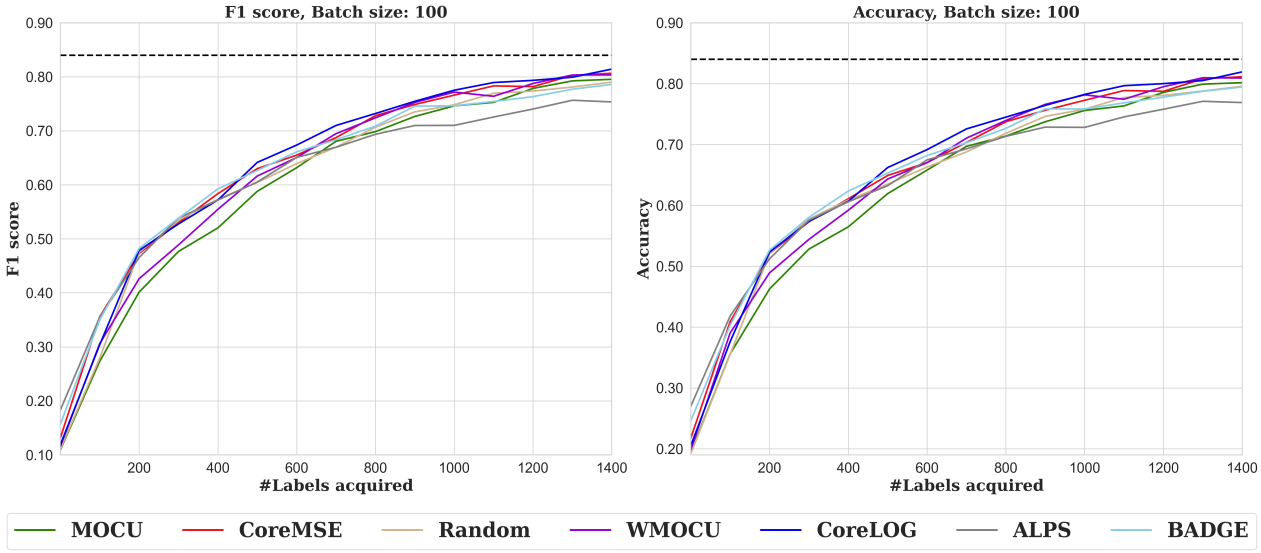


Fig. 2: Learning curves of batch size 100 for first-level DA classification. The dash line on the top represents the performance of the ELECTRA classifier trained on the full training dataset.

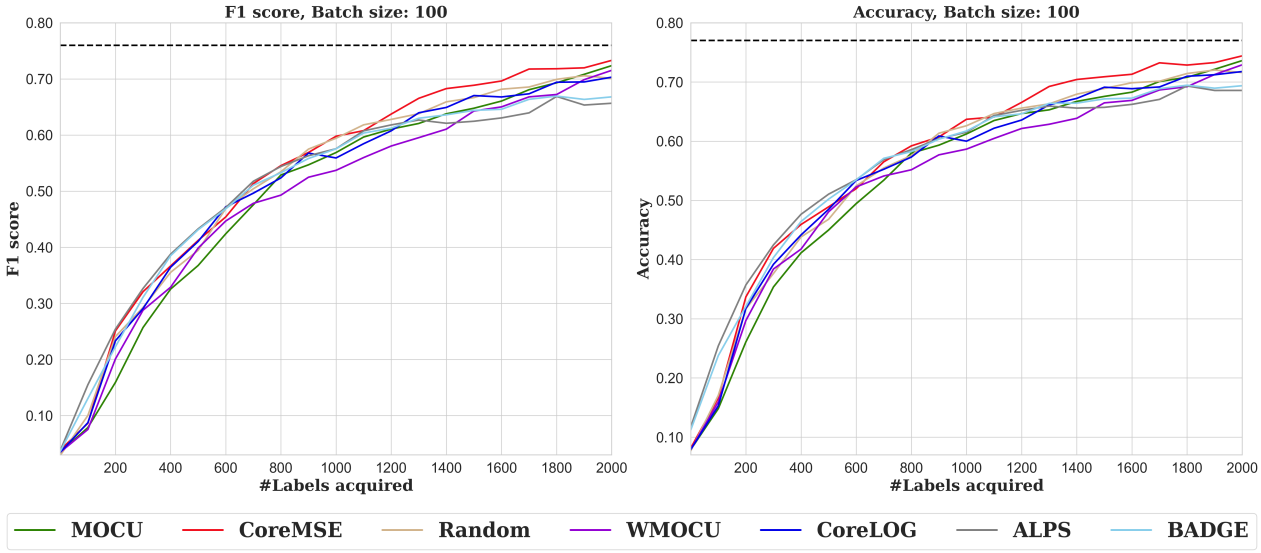


Fig. 3: Learning curves of batch size 100 for second-level DA classification. The dash line on the top represents the performance of the ELECTRA classifier trained on the full training dataset.

dash line on the top of Figure 3 represents the performance of the ELECTRA model on the full training dataset (0.771 accuracy and 0.762 F1 score). We observed that the CoreMSE method generally outperformed the other AL methods after acquiring 1,200 samples. When acquiring 2,000 samples, the ELECTRA model with the support of the CoreMSE method reached our desired performance where the model achieved the performance of 0.744 accuracy and 0.733 F1 score.

Figure 2 and 3 show the efficacy of AL methods CoreLOG and CoreMSE on the first- and second-levels of DA classification, respectively. To gain a better understanding of the difference between both AL methods and the random baseline in the sampling process, we further investigated the distribution of informative levels in each sample batch by *Data Maps*. As introduced in Section III-G, we can obtain the difficulty levels of Easy, Medium, Hard, and Impossible for each instance as the representatives of samples' informativeness. Figure 4 shows the distribution of sample informativeness between random and CoreLOG methods. We found that the difficulty levels of the sampled instances were evenly distributed. This indicates the random method was not selective to informativeness of the samples and some redundant samples were selected for the model training. In comparison, the CoreLOG method demonstrated that it gradually selected more informative samples (i.e., Medium, Hard, and Impossible) as more labels were acquired. When we scrutinised the sampling process on

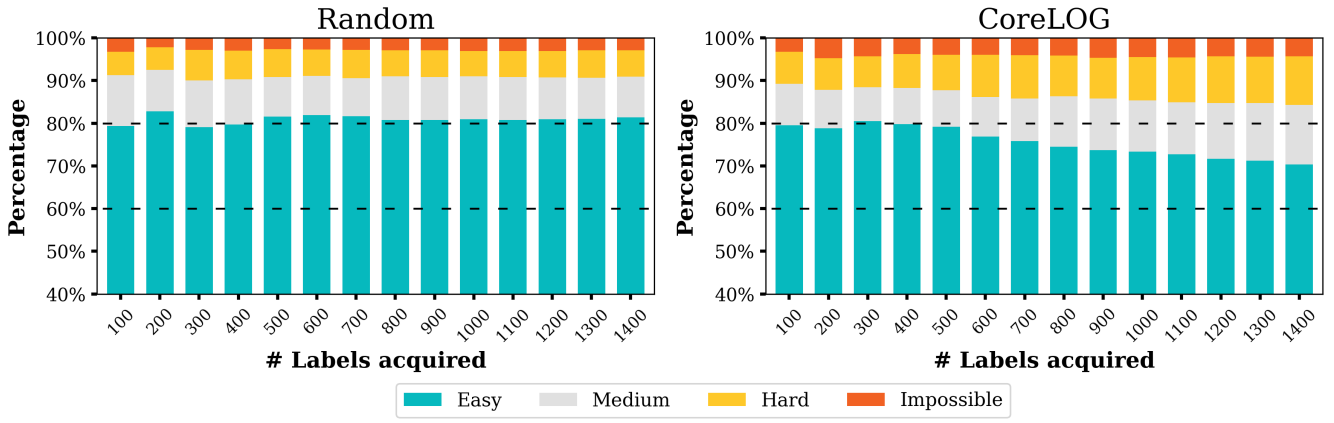


Fig. 4: We visualised the sampling preferences for each training batch between random baseline and CoreLOG method on first-level DA classification task. Compared with the random baseline, CoreLOG method was inclined to sample less easy samples after several batches.

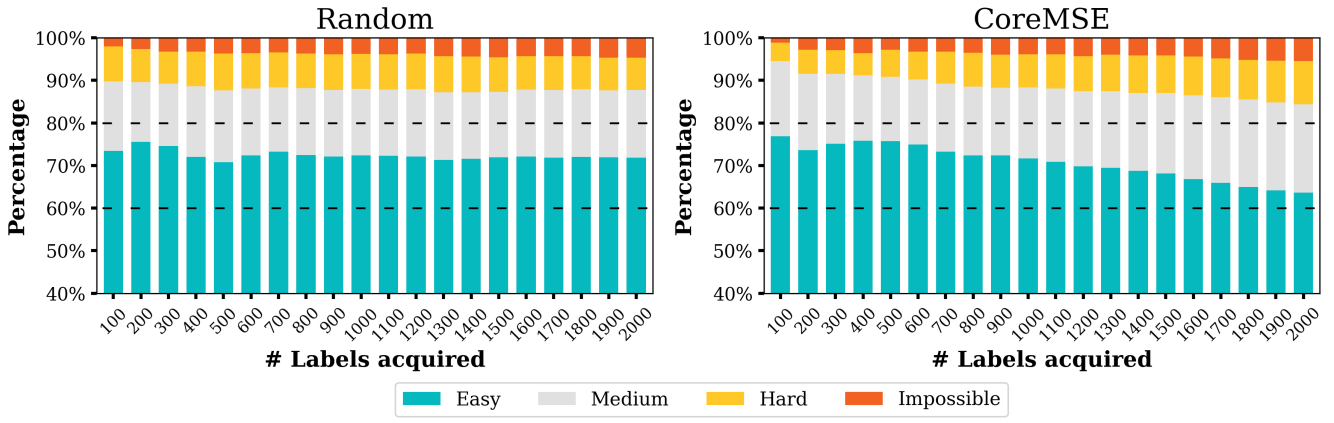


Fig. 5: We visualised the sampling preferences for each training batch between random baseline and CoreMSE method on second-level DA classification task. Compared with the random baseline, CoreMSE method was inclined to sample less easy samples after several batches.

the second-level DA classification, we also observed similar patterns between the random and CoreMSE methods (Figure 5). Figure 5 shows the sampling process of the random and CoreMSE methods on second-level DA classification. The random method selected the samples evenly distributed in each sample batch. In comparison, the CoreMSE method also demonstrated that it gradually selected more informative samples as more labels were acquired. These results of the selection of sample informativeness may explain the reasons why the AL methods are promising to be used to reduce the demand for annotating dataset for DA classification.

V. Discussion and Conclusion

A. Discussion on Educational Dialogue Act Classification

Exploring the content of discourse context was considered a potential direction for the improvement of educational DA classification in the existing literature [31], [9]. Our study contributed to the existing literature on educational DA classification by following aspects. Firstly, our results demonstrated that in our dataset, both ELECTRA and BERT models performed better when including the discourse context than excluding. The reason for this can be in the fact that the machine learning model might capture some keyword information from the previous utterances, which can further support the model to identify the DA in the current segment. Secondly, the improvement of DA classification performance from the discourse context (i.e., preceding segments), however, only sustained to a certain number of preceding utterances and the performance might decrease with more discourse context incorporated, which is in line with the findings of existing works in non-educational domain [34], [50], [51]. The reason for this can be in the fact that when the ML models incorporate the preceding segments more than the optimal amount, the more bias (e.g., non-relevant discourse context) will also be included in the model training process, which might further lower the model's performance. Therefore, we suggested that, to obtain a well-performed educational DA classifier, educational researchers should find the optimal

number of incorporated discourse contexts. In our study, we found that the ELECTRA model achieved optimal performance on first-level DA classification by incorporating three preceding utterances and on second-level DA classification by two preceding utterances. Our results also demonstrated practical implications for the design of dialogue-based ITS. As DA classification is considered an integral part of dialogue-based ITS, building a more reliable DA classifier can support a dialogue-based ITS to better understand students' requests and provide more accurate tutoring responses.

B. Discussion on Active Learning in Educational Research

Another direction for the improvement of educational DA classification is to alleviate the demand of annotated dataset [9]. In the current study, we explored this direction by evaluating the most recent pool-based AL methods and demonstrated how effectively the CoreLOG and CoreMSE methods can support deep learning models to achieve optimal performance while reducing the demand for annotated samples in educational DA classification. In addition, we also demonstrated the possible reason why both AL methods outperformed random baseline in selecting the samples from the dataset (shown in Figure 4 and Figure 5). Our results provide evidence that the CoreLOG and CoreMSE are reliable AL methods that can be of practical use to alleviate the costs of manual annotation on the task of educational DA classification. We expect that both AL methods can be applied to more educational classification tasks that are confronted with the issues of annotation cost (i.e., time and financial expense) for ML model training. For example, automating the process of identifying effective educational feedback may have a great impact on guiding novice tutors to draft feedback [60]. However, the annotation cost has constrained the performance of the practical utility in the existing works [60], [61]. We suggest that these works [60], [61] could consider deploying the pool-based AL methods in the annotation process, which might help to unlock the potentials of ML models by improving the accuracy of the classifiers and minimising the cost of annotation.

C. Limitations and Future Works

The labels of our annotated dataset were imbalanced. Though the AL methods could identify some impossible and hard instances (e.g., *"*divide"* annotated as `Correction`), the DA classifier cannot capture the pattern to successfully classify these instances because the percentage of these DAs was low (less than 1%). Therefore, in future work, it is worth annotating more instances for the DAs with low frequency. Then, many other active learning methods could be further explored. Secondly, our investigation was limited to a single large dialogue dataset. It is worth to evaluate the efficacy of AL methods across various student demographics (e.g., culture and educational background) and teaching subjects (e.g., programming courses). Thirdly, as AL methods are widely applied to the single-class classification, multi-class classification by AL methods is still under-explored. Many existing educational dialogue corpora [62], [63] involved multi-class classification. Therefore, it is worth further investigating the efficacy of active learning methods in these educational data corpora.

APPENDIX A

The example of the tutoring dialogue was given below.

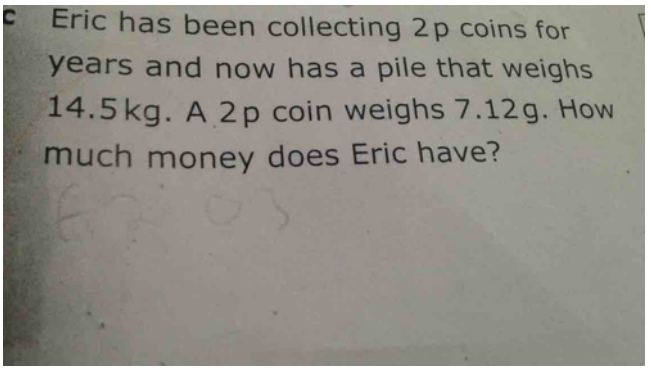


Fig. 6: The Image sent by student

Student: [Image Link] (as shown in Figure. 6)
Tutor: Hi [Anonymized], welcome!:)
Tutor: Please give me a minute to look over your problem.
Student: Ok thank you
Tutor: How much progress have you made so far? Or are you unsure of how to start?
Student: I know that I have to do 1450g divided by 7.12g but I'm a bit stuck after that
Tutor: You're nearly there! Looks like you just need some assistance with the unit conversion. I can certainly help you with that. Remember to participate as much as you can and we'll have this in no time.:)
Tutor: Tell me, how many grams are there in a kilogram?
Student: 1000
Tutor: Excellent! So, how many grams would be 14.5kgs?
Student: 14500
Tutor: Exactly! Now, you have got it right. Why do you think we would now divide 14500 by 7.12?
Student: Because one 2p coin weighs 7.12g and we want to find out how many 2p's are 14500g
Tutor: Very good! Alright, so how many 2p coins would be there?
Student: Well 14500 divided by 7.12 is 2036.52 so 2036?
Tutor: Absolutely correct! This is the number of coins that Eric would have.
Tutor: You can also find the equivalent amount of money in pounds, do you have any ideas on how we can do that?
Student: I'm not sure maybe multiplication???
Tutor: Yes, there can be multiplication. How many pounds do you think equal to 1p?
Student: Well there is 100 pennies on a pound so £20.36
Tutor: Awesome! Good job there.:)
Tutor: Thanks for participating. Is there anything else that I can help you with?
Student: Thank you for your help that's all for now but maybe later
Tutor: Sure, thanks for using our service! Have a good one.:)

REFERENCES

[1] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational researcher*, vol. 13, no. 6, pp. 4–16, 1984.

[2] R. E. Slavin, "Making chapter 1 make a difference," *The Phi Delta Kappan*, vol. 69, no. 2, pp. 110–119, 1987.

[3] R. A. Sottilare, J. A. DeFalco, and J. Connor, "A guide to instructional techniques, strategies and tactics to manage learner affect, engagement, and grit," *Design recommendations for intelligent tutoring systems*, vol. 2, pp. 7–33, 2014.

[4] M. Carlana and E. La Ferrara, "Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic," 2021.

[5] R. Nkambou, J. Bourdeau, and R. Mizoguchi, "Introduction: what are intelligent tutoring systems, and why this book?" in *Advances in Intelligent Tutoring Systems*. Springer, 2010, pp. 1–12.

[6] J. Paladines and J. Ramirez, "A systematic literature review of intelligent tutoring systems with dialogue in natural language," *IEEE Access*, vol. 8, pp. 164 246–164 267, 2020.

[7] V. Rus, N. Maharjan, L. J. Tamang, M. Yudelson, S. Berman, S. E. Fancsali, and S. Ritter, "An analysis of human tutors' actions in tutorial dialogues," in *The Thirtieth International Flairs Conference*, 2017.

[8] N. Maharjan, V. Rus, and D. Gautam, "Discovering effective tutorial strategies in human tutorial sessions," in *The Thirty-First International Flairs Conference*, 2018.

[9] J. Lin, S. Singh, L. Sha, W. Tan, D. Lang, D. Gašević, and G. Chen, "Is it a good move? mining effective tutoring strategies from human–human tutorial dialogues," *Future Generation Computer Systems*, vol. 127, pp. 194–207, 2022.

[10] B. D. Nye, A. C. Graesser, and X. Hu, "Autotutor and family: A review of 17 years of natural language tutoring," *International Journal of Artificial Intelligence in Education*, vol. 24, no. 4, pp. 427–469, 2014.

[11] A. Almasri, A. Ahmed, N. Almasri, Y. S. Abu Sultan, A. Y. Mahmoud, I. S. Zaqout, A. N. Akkila, and S. S. Abu-Naser, "Intelligent tutoring systems survey for the period 2000-2018," 2019.

[12] A. Alkhatlan and J. Kalita, "Intelligent tutoring systems: A comprehensive historical survey with recent developments," *arXiv preprint arXiv:1812.09628*, 2018.

[13] B. Du Boulay and R. Luckin, "Modelling human teaching tactics and strategies for tutoring systems: 14 years on," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 1, pp. 393–404, 2016.

[14] R. A. Sottilare, A. Graesser, X. Hu, and B. Goldberg, *Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management*. US Army Research Laboratory, 2014, vol. 2.

[15] A. K. Vail and K. E. Boyer, "Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes," in *Intelligent Tutoring Systems*, 2014.

[16] M. G. Core and J. Allen, "Coding dialogs with the damsl annotation scheme," in *AAAI fall symposium on communicative action in humans and machines*, vol. 56. Boston, MA, 1997, pp. 28–35.

[17] M. Walker and R. J. Passonneau, "DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems," in *Proceedings of the First International Conference on Human Language Technology Research*, 2001.

[18] H. Bunt, J. Alexandersson, J.-W. Choe, F. Chengyu, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum, "ISO 24617-2: 2012 language resource management–semantic annotation framework (SemAF)–part 2: Dialogue acts: 2012 language resource management–semantic annotation framework (SemAF)–part 2: Dialogue acts," 2012.

[19] D. R. Traum, "20 questions on dialogue act taxonomies," *Journal of semantics*, vol. 17, no. 1, pp. 7–30, 2000.

[20] H. Bunt, V. Petukhova, E. Gilmartin, C. Pelachaud, A. Fang, S. Keizer, and L. Prevot, "The ISO standard for dialogue act annotation," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 549–558.

[21] D. Morrison, B. Nye, B. Samei, V. V. Datla, C. Kelly, and V. Rus, "Building an intelligent pal from the tutor. com session database phase 1: Data mining," in *Educational Data Mining 2014*. Citeseer, 2014.

[22] V. Rus, N. Maharjan, and R. Banjade, "Dialogue act classification in human-to-human tutorial dialogues," in *Innovations in smart learning*. Springer, 2017, pp. 185–188.

[23] T. Rasor, A. Olney, and S. D'Mello, "Student speech act classification using machine learning," in *Twenty-Fourth International FLAIRS Conference*, 2011.

[24] A. Ezen-Can, J. F. Grafsgaard, J. C. Lester, and K. E. Boyer, "Classifying student dialogue acts with multimodal learning analytics," in *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 2015, pp. 280–289.

[25] V. Rus, R. Banjade, N. Maharjan, D. Morrison, S. Ritter, and M. Yudelson, "Preliminary results on dialogue act classification in chatbased online tutorial dialogues," in *Proceedings of the 9th international conference on educational data mining*, 2016, pp. 630–631.

[26] R. M. Felder and R. Brent, "Active learning: An introduction," *ASQ higher education brief*, vol. 2, no. 4, pp. 1–5, 2009.

[27] R. Hu, B. Mac Namee, and S. J. Delany, "Active learning for text classification with reusability," *Expert systems with applications*, vol. 45, pp. 438–449, 2016.

[28] K. Yu, S. Zhu, W. Xu, and Y. Gong, "Trnon-greedy active learning for text categorization using convex ansductive experimental design," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 635–642.

[29] B. D. Nye, D. M. Morrison, and B. Samei, "Automated session-quality assessment for human tutoring based on expert ratings of tutoring success." *International Educational Data Mining Society*, 2015.

[30] K. Boyer, E. Y. Ha, R. Phillips, M. Wallis, M. Vouk, and J. Lester, "Dialogue act modeling in a complex task-oriented domain," in *Proceedings of the SIGDIAL 2010 Conference*, 2010, pp. 297–305.

[31] B. Samei, H. Li, F. Keshtkar, V. Rus, and A. C. Graesser, "Context-based speech act classification in intelligent tutoring systems," in *International conference on intelligent tutoring systems*. Springer, 2014, pp. 236–241.

[32] B. Samei, V. Rus, B. Nye, and D. M. Morrison, "Hierarchical dialogue act classification in online tutoring sessions." in *EDM*, 2015, pp. 600–601.

[33] K. Stasaski, K. Kao, and M. A. Hearst, "CIMA: A large open access dialogue dataset for tutoring," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA â†' Online: Association for Computational Linguistics, Jul. 2020, pp. 52–64.

[34] E. Ribeiro, R. Ribeiro, and D. M. de Matos, "The influence of context on dialogue act recognition," *arXiv preprint arXiv:1506.00839*, 2015.

[35] B. Settles, 2012.

[36] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *International Conference on Learning Representations*, 2020.

[37] G. Zhao, E. Dougherty, B.-J. Yoon, F. Alexander, and X. Qian, "Uncertainty-aware active learning for optimal Bayesian classifier," in *International Conference on Learning Representations, ICLR 2021*, 2021.

[38] M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-start active learning through self-supervised language modeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7935–7948.

[39] W. Tan, L. Du, and W. Buntine, "Diversity enhanced active learning with strictly proper scoring rules," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[40] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[41] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[42] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[43] F. Dernoncourt and J. Y. Lee, "Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2017, pp. 308–313.

[44] H. T. Kahraman, S. Sagiroglu, and I. Colak, "The development of intuitive knowledge classifier and the modeling of domain dependent data," *Knowledge-Based Systems*, vol. 37, pp. 283–295, 2013.

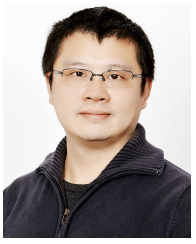
[45] J. C. Marineau, P. M. Wiemer-Hastings, D. Harter, B. A. Olde, P. Chipman, A. Karnavat, V. Pomeroy, S. Rajan, and A. Graesser, "Classification of speech acts in tutorial dialog," 2000.

[46] R. Pilkington, *Analysing educational discourse: The DISCOUNT scheme*. University of Leeds, Computer Based Learning Unit, 1999.

[47] S. Hennessy, S. Rojas-Drummond, R. Higham, A. M. Márquez, F. Maine, R. M. Ríos, R. García-Carrión, O. Torreblanca, and M. J. Barrera, "Developing a coding scheme for analysing classroom dialogue across educational contexts," *Learning, Culture and Social Interaction*, vol. 9, pp. 16 – 44, 2016.
[48] A. K. Vail, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Predicting learning from student affective response to tutor questions," in *ITS*. Springer, 2016, pp. 154–164.
[49] J. F. Grafsgaard, J. B. Wiggins, A. K. Vail, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring," in *Proceedings of the 16th International Conference on Multimodal Interaction*, 2014, pp. 42–49.
[50] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using context information for dialog act classification in dnn framework," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2170–2178.
[51] E. Ribeiro, R. Ribeiro, and D. M. de Matos, "Deep dialog act recognition using multiple token, segment, and context information representations," *Journal of Artificial Intelligence Research*, vol. 66, pp. 861–899, 2019.
[52] B. Noble and V. Maraev, "Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning," in *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, 2021, pp. 166–172.
[53] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.
[54] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.
[55] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3606–3611.
[56] K. Vodrahalli, K. Li, and J. Malik, "Are all training examples created equal? an empirical study," *arXiv preprint arXiv:1811.12569*, 2018.
[57] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9275–9293.
[58] S. Karamcheti, R. Krishna, L. Fei-Fei, and C. D. Manning, "Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 7265–7281.
[59] Y. Zhang, W. Cai, W. Wang, and Y. Zhang, "Stopping criterion for active learning with model stability," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 2, pp. 1–26, 2017.
[60] A. P. Cavalcanti, A. Diego, R. F. Mello, K. Mangaroska, A. Nascimento, F. Freitas, and D. Gašević, "How good is my feedback? a content analysis of written feedback," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, ser. LAK '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 428–437.
[61] A. P. Cavalcanti, R. F. L. de Mello, V. Rolim, M. André, F. Freitas, and D. Gaševic, "An analysis of the use of good feedback practices in online learning courses," in *2019 IEEE 19th international conference on advanced learning technologies (ICALT)*, vol. 2161. IEEE, 2019, pp. 153–157.
[62] A. Caines, H. Yannakoudakis, H. Edmondson, H. Allen, P. Pérez-Paredes, B. Byrne, and P. Buttery, "The teacher-student chatroom corpus," *Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2020)*, p. 10, 2020.
[63] S. Hennessy, C. Howe, N. Mercer, and M. Vrikki, "Coding classroom dialogue: Methodological considerations for researchers," *Learning, Culture and Social Interaction*, vol. 25, p. 100404, 2020.

Jionghao Lin is a Ph.D. student in the Centre for Learning Analytics at Monash University, Melbourne, Australia. His primary research interests focus on the areas of learning analytics, natural language processing, and affective computing. Currently, Jionghao is mainly working on applying artificial intelligent technologies to understand and optimize the learning environment. He received his B.E. degree from Jianghan University, China, and Master degree in Data Science from Monash University, Australia.

Wei Tan Wei Tan is a Doctoral Researcher who studies the cutting-edge machine learning algorithm in Data Science. He specializes in Active Learning that optimize the labelling budget and time for the human annotator. His PhD project is funded by Google Turning point. The aim is to develop the Surveillance System that will enable capture of a more complete set of coded ambulance data relating to SITB, mental health, and AOD attendances to inform policy, practice and intervention. He holds a master's degree from Monash University, and has expertise in analytics design for the social media platform.



Lan Du Dr. Lan Du is currently a senior lecturer in the faculty of information technology, Monash University, Australia. He completed his post-doctoral work at Macquarie University in the computational linguistic group. His research interest lies at the nexus of machine learning and natural language processing and their applications in different domains. His major publication outlets cover top conferences in machine learning, natural language processing, including NeurIPS, ICML, ACL, EMNLP, etc.

Wray Buntine Wray Buntine is full professor and Director of Computer Science at VinUniversity in Hanoi, and was previously Director of the Machine Learning Group at Monash University. He is known for his theoretical and applied work and in probabilistic methods for document and text analysis, with over 200 academic publications, 2 patents and some software products.

David Lang David Lang is a doctoral student in the Economics of Education program and an IES Fellow. He graduated from UCLA in 2008 with a B.A. in Economics a B.S. in Actuarial Mathematics. Prior to his doctoral studies, David worked for five years as a research analyst at the Federal Reserve Bank of San Francisco. His research interests include higher education, online education, and quantitative methods in education research. At Stanford, David also obtained a master's degree in Management Science and Engineering.

Dragan Gašević is Distinguished Professor of Learning Analytics in the Faculty of Information Technology and Director of the Centre for Learning Analytics at Monash University. As the past president (2015–2017) and a co-founder of the Society for Learning Analytics Research, he had the pleasure to serve as a founding program chair of the International Conference on Learning Analytics and Knowledge (LAK) and a founding editor of the Journal of Learning Analytics. His research centres on self-regulated and social learning, higher education policy, and data mining. He is a frequent keynote speaker and a (co-)author of numerous research papers and books.

Guanliang Chen is serving as a Lecturer in the Faculty of Information Technology, Monash University in Melbourne, Australia. Before joining Monash University, Guanliang obtained his Ph.D. degree at the Delft University of Technology in the Netherlands, where he focused on the research on large-scale learning analytics with a particular focus on the setting of Massive Open Online Courses. Currently, Guanliang is mainly working on applying novel language technologies to build intelligent educational applications. His research works have been published in international journals and conferences including AIED, EDM, LAK, L@S, EC-TEL, ICWSM, UMAP, Web Science, Computers & Education, and IEEE Transactions on Learning Technologies. Besides, he co-organized two international workshops and has been invited to serve as the program committee member for international conferences such as LAK, FAT, ICWL, etc.

4.3 Chapter Summary

In this Chapter, we investigated the methods to enhance the educational dialogue act classification (Gap 2). We explored the influence of discourse context in improving the dialogue act classification (RQ 3.1) and evaluated the statistically active learning methods to alleviate the labouring-intensive issues (RQ 3.2). To answer RQ 3.1, our study provided evidence of positive impacts by incorporating discourse contextual information into the input of the dialogue act classifier, which was summarised as **Contribution 5**. We suggest that future educational research related to dialogue act classification should include discourse context information to train a model which could potentially improve the dialogue act classifier performance.

Contribution 5: Impacts of discourse context on enhancing dialogue act classification.

- Overall, the dialogue act classifier performed better when including than when excluding the discourse context.
- The improvement of dialogue act classification performance from the discourse context only sustained to a certain number of preceding utterances and the model's performance decreased with more preceding utterances incorporated. In our study, the ELECTRA model achieved the optimal performance on fine-grained level dialogue acts classification (i.e., classify 31 dialogue acts) by two preceding utterances.

Additionally, we conducted a comparative study on the most recent statistical active learning methods with respect to their efficacy in alleviating labour-intensive issues on the task of educational dialogue act classification (RQ 3.2). We demonstrated the potential of employing the active learning methods on the data demanding classification tasks (i.e., dialogue act classification in our study), which was summarised as **Contribution 6**. We suggest that educational researchers may consider using active learning methods in their works to mitigate the cost of time and human labour.

Contribution 6: The efficacy of statistical active learning on facilitating dialogue act classification.

- The active learning method CoreLOG and CoreMSE demonstrated their efficacy in reducing the demand for annotated samples in first-level and second-level dialogue act classification, respectively. Both CoreLOG and CoreMSE methods presented that they can select more informative samples than the random baseline.

Chapter 5

Exploring the Politeness of Instructional Strategies from the Tutoring Dialogues

5.1 Introduction

After examining the role of politeness (in Chapter 2) and dialogue acts (in Chapter 3), we continued this line of research to further investigate the communicative patterns in tutoring dialogues (Gap 1 described in Chapter 1). In the tutoring dialogue, the expressions of politeness by educators are closely related to their dialogue acts (e.g., providing corrective feedback and asking thought-provoking questions) [66, 67]. As informed by prior research [61], most learners prefer to receive instructional information in feedback expressed in a polite manner. However, the excessive use of polite expressions in instructional information can have negative effects on the learning process [65–67]. Both benefits and hindrances of guiding learners politely coexist in the instructional communication process of tutoring dialogues, which requires further efforts to investigate the extent to which educators express politeness in instructional communication (Gap 1 described in Chapter 1). However, few studies have examined this relationship in online one-on-one tutoring dialogues (RQ 4).

To address RQ 4, we built on the research reported in Chapters 2 and 3. In Chapter 2, we demonstrated the potential of using the politeness scoring tool developed by Niu and Bansal [83] to measure the politeness levels of sentences in tutoring dialogue. We found that there was no evident correlation between the politeness of the educators' expressions and the learners' performance in non-instructional communication whereas the learners would benefit more from the educators' direct expressions than polite expressions in instructional communication. In Chapter 3, we categorised tutoring dialogue into dialogue acts, which can be used to distinguish

non-instructional and instructional communication. With the use of the instructional dialogue acts (e.g., positive feedback, information hint, and probing question) identified in Chapter 3, we answered RQ 4 by investigating the politeness levels of instructional dialogue acts.

This research has been published in the Proceedings of the *12th International Learning Analytics and Knowledge Conference*.

- Lin, J., Raković, M., Lang, D., Gašević, D., & Chen, G. (2022). Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 282-293).

5.2 Publication: Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues.

Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues

Jionghao Lin
jiongh.lin@gmail.com
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Mladen Raković
Mladen.Rakovic@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

David Lang
dnlang86@standford.edu
Graduate School of Education,
Stanford University
California, United States

Dragan Gašević
Dragan.Gasevic@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Guanliang Chen*
guanliang.chen@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

ABSTRACT

Existing research indicates that students prefer to work with tutors who express politely in online human-human tutoring, but excessive polite expressions might lower tutoring efficacy. However, there is a shortage of understanding about the use of politeness in online tutoring and the extent to which the politeness of instructional strategies can contribute to students' achievement. To address these gaps, we conducted a study on a large-scale dataset (5,165 students and 116 qualified tutors in 18,203 online tutoring sessions) of both effective and ineffective human-human online tutorial dialogues. The study made use of a well-known dialogue act coding scheme to identify instructional strategies, relied on the linguistic politeness theory to analyse the politeness levels of the tutors' instructional strategies, and utilised Gradient Tree Boosting to evaluate the predictive power of these politeness levels in revealing students' problem-solving performance. The results demonstrated that human tutors used both polite and non-polite expressions in the instructional strategies. Tutors were inclined to express politely in the strategy of providing positive feedback but less politely while providing negative feedback and asking questions to evaluate students' understanding. Compared to the students with prior progress, tutors provided more polite open questions to the students without prior progress but less polite corrective feedback. Importantly, we showed that, compared to previous research, the accuracy of predicting student problem-solving performance can be improved by incorporating politeness levels of instructional strategies with other documented predictors (e.g., the sentiment of the utterances).

*Corresponding Author

LAK22, March 21–25, 2022, Online, USA

ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506904

CCS CONCEPTS

• **Applied computing** → **Education**; **E-learning**; **Distance learning**.

KEYWORDS

Learning Analytics, Educational Dialogue Analysis, Politeness, Student Performance, Prediction

ACM Reference Format:
Jionghao Lin, Mladen Raković, David Lang, Dragan Gašević, and Guanliang Chen. 2022. Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3506860.3506904

1 INTRODUCTION

Tutoring is an effective form of instruction that can support students in different learning tasks across all educational levels [4, 35, 37]. The tutoring process often involves many instructional activities employed by tutors to guide student learning (e.g., providing corrective feedback and hints) and these activities can be conceptualised as instructional strategies, i.e., the approaches and principles employed by instructors to bring students towards the instructional goals [13, 34]. Instructional strategies can be manifested as tutors' words or sentences [35, 37]. There are many types of instructional strategies including hints (e.g., *"You may consider using the second condition"*), open questions (e.g., *"Why do you use this formula?"*), and corrective feedback (e.g., *"Yes, that is right"* and *"No, the second step is not correct"*) [8]. However, some strategies (e.g., directive hints and negative feedback) should be carefully used by tutors because these strategies might impose negative feelings on their students [7, 8], which might further impede students' motivation and acceptance of feedback [31]. For example, tutors can trigger students' feelings of restricted autonomy (e.g., by requesting students to take certain steps, *"Solve the value of x first"*) or neglect to show appreciation to students' efforts (e.g., by providing explicit negative feedback on students' work, *"You're wrong. Try it again"*) [8]. These expressions that can harm students' self-efficacy are called direct expressions [6].

To mitigate the negative feelings that direct expressions can impose on students, human tutors often express their words in a polite

manner (e.g., *"Let us work together to solve the problem*"). In this way, the tutors demonstrate solidarity and build rapport with students in the tutoring process [7, 32], which has been demonstrated to benefit students' learning performance [23–26, 38–40]. Several previous studies highlighted the benefits of polite expressions in the tutoring practice [24, 26, 36, 38, 39] and pointed out the necessity of including those polite expressions in the dialogue-based intelligent tutoring system (i.e., a computer-based tutoring system that guides students in a learning process, via natural conversation) [19, 22]. However, tutors' excessive use of polite utterances in tutoring dialogue can have negative effects on learning process [7, 8, 30]. For instance, some human tutors might spend too much time working on polite expressions which can in turn hinder the tutoring process [7, 8, 30]. Moreover, the previous research demonstrated that students with high prior knowledge often prefer to receive instruction directly expressed [24].

Therefore, both benefits and hindrances of guiding students politely coexist in instructional communication. To maintain the effectiveness of one-on-one online tutoring, it is worth investigating the extent to which tutors express politeness in different instructional strategies (e.g., providing hints and corrective feedback). A promising way to investigate how tutors effectively communicate with students is to analyse instructional strategies based on student performance (e.g., whether students successfully completed learning tasks) [35]. Consequently, the appropriateness of expressing politely in the instructional communication can be reflected by investigating the relationship between the use of politeness in instructional strategies and student performance. However, few studies have examined this relationship in online one-on-one tutoring. As we posited the use of appropriate politeness can better facilitate the online one-on-one tutoring process, it is worth investigating this relationship to enhance the effectiveness of the instructional communication and further support students towards their learning goals.

To this end, our study aimed to analyse online human-human tutorial dialogues to shed light on the use of politeness in different instructional strategies. Specifically, we investigated different types of instructional strategies and their politeness levels in relation to the student prior progress (i.e., the progress on the problem that students made before joining a tutoring session) and the students' problem-solving performance. The study also aimed to examine the extent to which the politeness levels of instructional strategies can be used to predict students' problem-solving performance. Formally, our study was guided by the following two research questions:

- **RQ 1** To what extent do tutors express politeness in different instructional strategies in online human-human tutoring?
- **RQ 2** To what extent does the politeness level of instructional strategies predict students' performance in online human-human tutoring?

To answer RQ1, we firstly employed a widely-used dialogue act scheme (proposed by Vail and Boyer [37]) to identify tutor actions behind the utterances in a tutorial dialogue (e.g., asking a question and providing hints); these tutor actions were represented as instructional strategies. Then, we relied on the linguistic theories of politeness proposed by Brown and Levinson [6] to analyse the politeness levels in the tutors' utterances in online one-on-one tutoring. Building upon dialogue actions and politeness levels identified through the analysis of tutors' utterances, we investigated the politeness levels used in different instructional strategies. For RQ2, the politeness levels of instructional strategies were further used as input for training a well-established machine learning model —Gradient Tree Boosting (GTB) [9] —to predict student performance based on the statistical features (i.e., *mean*, *standard deviation* and *median*) of the politeness levels of instructional strategies in each tutoring session.

Our results revealed that human tutors used both polite and direct expressions in the instructional strategies. Most tutors were inclined to express directly in comprehensive-gauging questions (e.g., *"Does that make sense?"*) and negative feedback (e.g., *"No, it is incorrect"*) and politely in positive feedback (e.g., *"Well done, your result is correct"*). Compared to the students who had made some prior progress on the problem before the start of a tutoring session, tutors were inclined to show more polite expressions in open questions but less polite in corrective feedback (i.e., negative feedback and positive feedback) to the students without prior progress. Lastly, by incorporating the politeness levels of instructional strategies with other documented factors (e.g., sentiment, time on task, and task complexity) [21], the GTB model achieved better performance of predicting student problem-solving performance compared to our previous work [21].

2 RELATED WORK

2.1 Facework in Instructional Communication

The central concept of Brown and Levinson's politeness theory is *Face*, which is the feeling of being respected by others [6]. In instructional communication, students' face can inevitably be threatened by tutors' utterances since tutors often restrict students' autonomy (e.g., direct students how to proceed to next step, *"Solve the value of x first"*) and may neglect students' need of appreciation (e.g, negative feedback, *"No, it's wrong"*) [8]. These face threatening utterances are called face-threatening acts (FTAs) [6]. If the tutors do not redress the FTAs in the instructional communication, tutors will instruct the students through direct expressions. In contrast, tutors can also express politely to mitigate FTAs to maintain students' face. For example, instead of directly requesting students to proceed to the next step (e.g., *"Solve the value of x first"*), tutors can use more polite guidance (e.g., *"Shall we work on the value of x first?"*) to alleviate the feeling of restricting the autonomy.

2.2 Effectiveness of Politeness in Instructional Communication

Existing research suggests that polite tutors are more effective than direct tutors in guiding students' problem solving and to learn new concepts in different educational contexts [23–26, 38–40]. For example, Wang et al. [39] found that students guided by polite instruction (e.g., *"Do you want to save the value now"*) had better performance than those guide by direct instruction (e.g., *"Save the value now"*) in solving industrial engineering problem. The effectiveness of polite instruction on students' performance was further confirmed in learning foreign language [38]. In line with the findings presented in [39], McLaren et al. [24] also examined the effects of politeness in a web-based tutoring system and found that

students who received polite tutorial instructions performed better than those who received direct tutorial instructions. Schneider et al. [36] showed that students who were given politely written task instructions outperformed their peers who were given direct instructions. Recently, Mikheeva et al. [26] found that the politely written feedback improved student performance in solving complex mathematical problems at the university level.

Despite many documented benefits of polite expression in previous work, tutors should avoid excessively expressing politeness in instructional communication [7, 8, 18, 30]. The reasons can be summarised from two perspectives. First, excessive polite expressions might lead to negative effects on instructional efficiency. The earliest research of politeness in instructional communication pointed out that tutors should avoid rephrasing certain instructions completely in a polite expression as it might hinder the tutors' ability to give adequately informative feedback to students [30]. This claim was further confirmed by Brummernhenrich and Jucks [7]. They found that compared with the online tutoring sessions guided by polite tutors, non-polite tutors provided more hints, requests and answers to students' questions [7]. Second, the positive effect of politeness was conditioned on the levels of students' prior knowledge. Several empirical studies found that polite expression was only effective in supporting the students who had low prior knowledge (e.g., mastery levels on a learning subject) [24, 38, 39]. It should be noted that the students with high prior knowledge might feel more difficult to work with polite tutors than direct tutors on problem-solving tasks [24].

To make the online tutoring sessions effective, tutors should make a trade-off between instructing politely and directly [8]. Exploring the appropriate use of politeness in instructional expressions is therefore critical for tutors to maintain the productive communication. However, it is still unclear the extent to which tutors express appropriate politeness in different instructional strategies in online one-on-one tutoring. The studies by [7, 8] concluded that there was a clear connection between politeness and instructional strategies but they did not reveal how the tutors strike the balance between the direct and polite expressions in the tutoring process. The study presented in this paper, instead, investigated the extent to which tutors expressed politeness in instructional communication on a large scale tutorial dialogue dataset. Secondly, the studies by [7, 8] did not investigate the changes of the politeness levels of instructional strategies as the tutoring sessions progressed. In contrast, our previous work found that in comparison to ineffective tutoring sessions, in effective tutoring sessions where students successfully collaborated with tutors to solve problems, tutors tended to use polite expressions at the start of the online tutoring sessions and gradually used more direct expressions to instruct students as the tutoring sessions progressed [19]. However, it is still unclear whether the effectiveness of tutoring associated with the politeness levels of instructional communication as the tutoring progressed.

3 METHODS

3.1 Dataset

We analysed the tutor-student dialogue dataset provided by an online educational company. Our research was approved by the Human Research Ethics Committee of Monash University under Project ID 26156. The dataset contained data about 5,165 K-12 students and 116 qualified tutors. The students and tutors were working collaboratively on solving problems in different STEM subjects such as mathematics, chemistry and physics. A tutoring session was commonly initiated by a student by sharing unsolved problems and an experienced tutor was then allocated to work collaboratively with the student, rather than to directly provide students with a simple solution. Therefore, the communication between tutors and students contained fine-grained details about the scaffolding process in solving problems. The dataset originally contained 18,203 dialogue sessions but some short tutoring sessions did not have sufficient tutoring guidance. Therefore, we discarded the dialogue sessions that concluded within the first minute and contained less than 10 utterances. We also discarded the sessions where the students' prior progress was not available and the details of the process followed to determine the students' prior progress is described in Sec. 3.2. After removal, the final dataset contained 14,562 tutorial sessions and 92% of them were related to the maths tutoring.

3.2 Data Pre-processing

Yang and Li [42] suggested that tutors should select appropriate instructional strategies with the consideration of the students' prior progress which was positively associated with students' self-efficacy and problem-solving performance. In the stage of data preparation, we first manually annotated the students' prior progress based on the tutor-student utterances. To this end, we identified students' prior progress in each session by observing the first several utterances from tutors and students. At the start of a tutorial session, tutors often asked students about the progress that they made before joining the tutoring sessions. For instance, a tutor might ask *"What you have done on this problem?"*, and a student might answer *"Nothing"* or *"Yes, let me show you"* and then described the progress they had made. By observing the first few tutor-student utterances of a tutorial dialogue, we annotated the tutoring sessions as either *With-Prior-Progress* or *Without-Prior-Progress*. The annotation process was first implemented by two human coders. Each dialogue was labelled by two coders independently. The percentage of agreement between two human coders was 0.847. The Cohen's κ score of 0.735 was derived as a measure of inter-rater reliability of the data labelling, which demonstrated a substantial level of agreement. The disagreement between the two coders was resolved by a third independent coder. It should be noted that we had no access to the information indicating students' prior knowledge on the mastery levels of a specific learning subject (e.g., mathematics). Instead, we included the student prior progress (i.e., the progress that students made on the consulted problem before joining a tutoring session) as the proxy of student prior knowledge, which was in line with the definition of student prior knowledge (i.e., student experiences on the learning contents before they having a tutoring session [14]).

3.3 Descriptive Statistics of the Dataset

To answer research questions RQ1 & 2, we identified the students' performance for each tutoring session since incorporating the students' performance in the analysis can better inform the appropriateness of using politeness in instructional strategies. To this end, we annotated each tutorial session based on the student's performance at the end of a session. Instead of annotating each session as

Table 1: The descriptive statistics of the dataset. WP and WoP denote *With Prior Progress* and *Without Prior Progress*, respectively. Mann-Whitney tests were applied to examine differences (Rows 5-9) between any pair of the Gap-clarified, Gap-explained, and Gap-bridged categories in which students had the same level of prior progress. All differences were significant ($p < 0.01$).

Metric	All	Gap-clarified		Gap-explained		Gap-bridged	
		WP	WoP	WP	WoP	WP	WoP
1. # total sessions:	14,562	1,203	1,302	1,255	1,931	4,482	4,389
2. # total uttrances:	1,216,784	31,014	30,128	78,575	113,099	475,849	488,119
3. # tutors:	116	92	96	98	99	110	106
4. # students:	5,165	763	962	908	1,419	1,800	2,168
5. Avg. Sess Dur (mins):	30.27 ± 30.66	10.55 ± 7.64	9.75 ± 7.21	25.94 ± 19.03	22.88 ± 18.05	37.78 ± 32.17	38.60 ± 37.37
6. Avg. # Uttr / Sess:	83.56 ± 81.05	25.78 ± 16.68	23.14 ± 14.92	62.61 ± 43.79	58.57 ± 42.73	106.17 ± 87.62	111.21 ± 93.70
7. Avg. # Words / Sess:	647.75 ± 596.12	201.62 ± 131.81	198.13 ± 134.44	524.09 ± 351.18	489.56 ± 346.82	807.46 ± 649.08	845.28 ± 675.05
8. Avg. % Uttr by tutors:	58.42 ± 7.86	53.95 ± 9.49	56.46 ± 9.51	58.68 ± 7.82	60.25 ± 7.77	58.03 ± 7.07	59.75 ± 6.94
9. Avg. % Words by tutors:	78.36 ± 9.10	74.32 ± 11.80	80.54 ± 10.20	78.87 ± 8.59	82.21 ± 7.96	76.09 ± 8.69	79.30 ± 7.81

effective and ineffective (e.g., [35, 37]), our study provided a more fine-grained approach to distinguish the performance level of a student, which was listed in below three categories:

- **Gap-clarified**: a tutor identified the problem but was unsure whether the student made any progress at the end of the tutoring session;
- **Gap-explained**: a tutor identified the problem and assisted the student to make certain progress, but the student did not completely solve the problem;
- **Gap-bridged**: a tutor identified the problem and guided the student to successfully solve the problem.

Each tutorial dialogue was labelled by an educational expert employed by the data provider (i.e., the educational technology company). To validate the reliability of their annotations, we randomly selected 500 tutorial dialogues from the whole dataset and annotated them independently by using the same coding rules. The result of validation reached Cohen's κ score of 0.787. We provided sample dialogues for each of the above three categories in an electronic appendix, which is accessible via https://bit.ly/lak22_tutor_appendix.

The descriptive statistics of the dataset are shown in Table 1. We observed that most of the dialogues were of the Gap-bridged sessions (8,871, 60.9%), followed by Gap-explained (3,186, 21.9%), and then Gap-clarified (2,505, 17.2%). This indicated that more than half of the students solved their problems at the end of the sessions. Additionally, we observed that more than 47% of the students made some progress on the problem before joining a tutoring session Table 1. Rows 5-7 of Table 1 showed that tutors and students invested more effort in the Gap-bridged sessions than the other two. An explanation for this might be the more efforts tutors and students invested in tutorial sessions, the better problem-solving performance achieved by student.

3.4 Inferring the Instructional Strategies from Instructional Communication

To answer RQ1, we identified the instructional strategies as shown in the tutors' utterances. We drew from the work by Brummernhenrich and Jucks [8], who investigated the politeness in seven types of commonly used instructional strategies, which were presented in the first column of Table 2. Brummernhenrich and Jucks [8] identified the instructional strategies by relying on the efforts of experienced human coders, which was quite time-consuming and cost-demanding to identify the instructional strategies on a large amount of tutors' utterances. Instead, in our study, we automated the process of identifying the instructional strategies. We characterised instructional strategies by capturing the actions (i.e., dialogue acts) hidden behind tutors' utterances; this was informed by prior research [27, 35, 37]. To automatically identify the dialogue acts, our previous work [21] applied the BERT language model [1] to train a dialogue act classifier, which was built on 3,629 utterances (*training*, *validation*, and *testing* datasets in the ratio of 80%:10%:10%) from tutors and students. It should be noted that the dialogue acts classifier was trained on the same dataset as the current study. After tuning the classifier with the optimal hyper-parameters, the classifier achieved sufficient performance (F1 score 0.742 and Cohen's κ score of 0.735) on identifying 31 dialogue acts (composed of non-instructional and instructional communication) from tutors and students, which were built upon the well-known dialogue acts scheme [37]. For the current study, we only focused on the tutors' dialogue acts which matched instructional strategies proposed in [8]. Compared with the the work by Brummernhenrich and Jucks [8], our identified instructional strategies presented a more fine-grained level (as shown in the second column of Table 2). For instance, the strategy `Hint` was differentiated as `Information hint` (i.e., prompt students work on the next step) and the `Observation hint` (i.e., remind students of known information).

3.5 Inferring the Politeness from Instructional Communication

A tutor's utterance could be interpreted as a certain type of instructional strategy and be reflective of a certain level of politeness. The recognition of politeness levels could further help researchers understand the extent to which human tutors expressed politeness in instructional strategies. To this end, we employed the well-established politeness level scoring tool (available in the `Github` repository[1]) by Niu and Bansal [29] to quantify the politeness levels of tutors' utterances. This tool was originally developed to classify textual data into direct or polite categories. In addition to assigning the polite or direct categories, this tool also generated a score to indicate the politeness level from 0 (most direct) to 1 (most polite) [29].

[1]https://github.com/WolfNiu/polite-dialogue-generation

Table 2: The description of the instruction strategies derived from Brummernhenrich and Jucks [8] and Lin et al. [21]

Instructional Strategy in [8]	Instructional Strategy in Our Work [21]	Sample Utterances
Hints	Information Hint Observation Hint	*"It can be any one of the cards in the deck."* *"We have 80."*
Closed question (prompt)	Probing Question	*"How many options can it be?"*
Open-ended Questions	Open Question	*"What do you think we could try next?"*
Comprehension-gauging questions	Evaluation Question	*"Does that make sense?"*
Explanation	Explanation	*"The straight line is the line on the bottom."*
Positive Feedback	General Positive Feedback Elaborated Positive Feedback	*"Correct!"* *"Your formula for period is correct!"*
Negative Feedback	Negative Feedback	*"No, it is incorrect."*

The generated politeness scores, hence, helped us understand the politeness levels for tutors' utterances in the instructional communication. Since this tool was originally used to classify the textual data as either polite or direct, we validated the reliability of this tool on our dialogue dataset as follows. We recruited two human coders to independently annotate 500 utterances (randomly sampled from our dialogue data) into polite and direct labels. The percentage of agreements between two human coders was 0.724 and the Cohen's *kappa* score was 0.531, which was enough for analysis [28]. The inconsistency between two human coders was then resolved by the third human coder. Then, we used our annotated results to compare the classified results by the tool [29] and obtained Cohen's k score of 0.639, which presented a sufficient score to use this tool for analysis.

To infer the politeness levels of instructional strategies, we first used the dialogue act classifier developed in our previous work [21] to identify the hidden instructional strategies from tutors' utterances. Then, we used the politeness level scoring tool by [29] to identify the politeness levels of the same tutors' utterances. Finally, we paired the instructional strategies and the politeness levels based on the same tutors' utterances to represent the politeness levels of instructional strategies.

3.6 Predicting Student Performance

3.6.1 Prediction Model. To answer RQ2, we trained a Gradient Tree Boosting (GTB) [9] machine learning model to predict the student performance (i.e., Gap-clarified, Gap-explained, and Gap-bridged) and a Random Forest model as the comparison on the prediction. The Random Forest model was built upon the idea of the ensemble method [12] which claimed that combining many predictors to make predictions on the same task can obtain better performance than an individual predictor. [15] found that the Random Forest model had generally better performance compared to 179 commonly used machine learning models (e.g., Naive Bayes, Decision Trees, and Support Vector Machines) on 121 different classification tasks. Therefore, our study adopted Random Forest classifier to answer RQ2. Similar to the Random Forest model, GTB also processed the data by a collection of decision tree models as the predictors, and the final prediction was made by considering the predictions from all constructed predictors. It should be noted that each decision tree predictor was trained on a sample randomly selected from the data so each decision tree was distinct from others. These distinct decision trees work collaboratively to capture patterns in the data to make predictions. The main difference between GTB and Random Forest was the collaborative approach of predictors. For Random Forest, the predictors were constructed independently and many of them could produce the same type of prediction error. In contrast, the predictors in GTB were constructed sequentially so that the prediction errors made by previous predictors can be corrected by the subsequent predictors. We selected GTB model to answer research RQ2 due to its accuracy in our previous study [21] in predicting student performance on the same dataset.

3.6.2 Feature Engineering. To train the GTB classifier, we engineered a total of 27 statistical features (i.e., *mean*, *standard deviation*, and *median* suggested by [41]) about strategy politeness, which included: (i) average politeness level of the nine instructional strategies used in our study as shown in Table 2; (ii) standard deviation of politeness level of nine instructional strategies; and iii) median values of politeness level of nine instructional strategies. We denoted these features as *Strategy-politeness Feat.* We assumed that students' prior progress on a problem solving task might be beneficial in enhancing the model performance. However, we did not include the students' prior progress into the feature set because we aimed to build a predictive model for real-time use (i.e., the input features to the model should be automatically generated from the observed data) and the progress level was not readily available.

The current study mainly focused on the politeness features of instructional strategies produced by tutors. However, as student performance could be influenced by other factors (e.g., student sentiment [17]) in the tutoring process, we also included those non-politeness features in our final model. Building upon our previous work [21], we engineered additional 543 non-politeness features to enhance our feature set, which included:

- **Effort**, which measured three activities from tutor-student utterances in a tutoring session (i.e., the duration of the tutorial session, the number of utterances, and the total number of words in utterances).
- **Informativeness**, which counted the number of concepts and unique words in the tutor and student utterances.
- **Complexity**, which calculated the Flesch readability score [11] from the utterances tutors and students.
- **Responsiveness**, which measured the average time that students spent waiting for the tutor's reply after the request for help had been made.

- **# Questions**, which counted the number of questions that tutors and students created.
- **Entrainment**, which measured the level of lexical consistency [5] that a tutor's and student's utterances reached in a tutoring session.
- **Sentiment**, which described the sentiment levels, ranging from -1 (most negative sentiment) to +1 (most positive sentiment), detected in utterances from tutors/students. The sentiment levels were generated by the VADER model [16].
- **Experience**, which described the number of tutorial sessions that students and tutors participated in before the current one.
- **Dialogue Acts**, which described the intention behind the tutors/students' utterances (e.g., asking questioning). This group of features was composed by the number of specific dialogue acts made by students and tutors in a dialogue, the fraction of specific dialogue acts made by students and tutors in a dialogue, and the number of sequential dialogue act patterns made by students and tutors in a tutoring session.
- **N-grams**, which contained the top 100 most frequent unigrams, bigrams, and trigrams lexicons extracted from the tutor-student utterances.

3.6.3 Study Setup. Our study evaluated the prediction performance of the GTB model by comparing with the RF model (baseline model). Before training both models, we randomly allocated the full annotated dataset into *training set*, *validation set*, and *testing set* with the proportion of 80%:10%:10%, respectively. Then, both models were trained using the Python package `scikit-learn`[2]. The hyperparameters (e.g., the maximum depth of the tree) in both models were tuned by using grid search on the validation dataset. Finally, both models' performance was evaluated on the testing set. To evaluate models' performance, we employed four representative metrics, i.e., classification accuracy, F1-score, Area Under the Curve (AUC), and Cohen's k score.

3.6.4 Ablation Test. To gain a better understanding of the predictive power of different types of features, we also measured the respective contributions made by all features (both strategy-politeness and non-politeness features) by conducting an ablation test, which is a widely-used method by researchers to assess contributions made by features on the performance of a prediction model [20, 21]. The contribution of a feature was measured by calculating the difference between the prediction performance of a model including the feature and the model excluding the feature.

4 RESULTS

4.1 Results on RQ1

The average politeness levels of instructional strategies are presented in Table 3. These instructional strategies are sorted based on the values of their average politeness levels calculated in the whole dataset. The results indicated that, in real-world online one-on-one tutoring, tutors tended to express different politeness levels of instructional strategies in the tutoring process to guide students. The three most direct instructional strategies in our dataset were `Evaluation Question` (e.g., *"Does that make sense?"*), `Negative Feedback` (e.g., *"No, it is incorect"*), and `Probing Question` (e.g., *"What if you add the value of y?"*), and the three most polite ones were `General Positive Feedback` (e.g., *"Correct"*), `Elaborated Positive Feedback` (e.g., *"Yes, you use the function correctly"*), and `Information Hints` (e.g., *"Hint: you might use the third condition"*).

We investigated the difference of politeness levels of the instructional strategies between any pair of the Gap-clarified, Gap-explained, and Gap-bridged sessions. A pair-wise comparison was conducted between students with and without prior progress and the differences were examined by the Mann-Whitney test; the significant results are marked with the same symbol (i.e., ◇, †, ♣) in Table 3. We found that tutors in the Gap-bridged sessions presented lower average politeness levels of the strategies (`Probing Question`, `Observation Hint`, and `General Positive Feedback`) than in the other two categories of sessions (i.e., Gap-explained and Gap-clarified) in both groups of students with and without prior progress. This finding indicated that tutors in the most effective sessions (i.e., Gap-bridged) tended to use direct expressions of these instructional strategies to guide students. Then, we scrutinised the politeness levels of the instructional strategies between *With-Prior-Progress* and *Without-Prior-Progress* for each session category. We found that the politeness levels of the instructional strategies `Negative Feedback` and `Elaborated Positive Feedback` when used with the *With-Prior-Progress* students were statistically higher than those used with the *Without-Prior-Progress* students in the Gap-bridged session categories. This result indicated that the tutors tended to give explicit and direct expressions of feedback to the *Without-Prior-Progress* students. Additionally, we also found that *Without-Prior-Progress* students received more polite expression of `Open Questions` from tutors compared to their *With-Prior-Progress* counterparts in the Gap-bridged sessions.

We further investigated the distribution of politeness in instructional communication (shown in Fig. 1). Building upon the suggestions by Niu and Bansal [29], we categorised the politeness levels of tutors' utterances into three politeness groups (i.e., the politeness levels of tutors' utterances in the range (0.8, 1.0] as *Polite*, [0.2, 0.8] as *Neutral*, and [0, 0.2) as *Direct*). These politeness groups were used to demonstrate the variance of instructional strategies' politeness levels. Taking the strategy `Evaluation Question` (i.e., `Eval Ques` in Fig. 1) as an example, less than 5% of the `Evaluation Question` strategies were presented in the polite form (e.g., *"I hope this make sense?"*), 40% in direct (e.g.,*"Do you understand what I mean?"*), and more than 50% in neutral (e.g.,*"Would that make sense?"*), which is shown for the three different types of sessions in Fig. 1). Additionally, the percentage of polite `General Positive Feedback` (e.g., *"Brilliant!"*) in the Gap-bridged session was less than the other two sessions. Instead, tutors in Gap-bridged sessions used more neutral form (e.g., *"Yes, that is right"*) to inform the correctness. It should be noted that tutors from three different sessions rarely used the direct form of `General Positive Feedback` (e.g., *"Exactly as what I got"*).

Inspired by the variation of politeness existed within each instructional strategy, we further investigated the variation of instructional strategies' politeness as the tutorial progressed and in particular at the early stage of the tutoring which is important to build the rapport with students [3]. Based on our previous work [21], some instructional strategies (e.g., `Negative Feedback` and

[2]https://scikit-learn.org/

Table 3: The politeness level of instructional strategies sorted according to the values of their average politeness levels in the dataset in an ascending order, i.e., the column All. WP and WoP denote *With Prior Progress* and *Without Prior Progress*, respectively.Mann-Whitney test was applied to examine the difference between WP and WoP for each session categories (significant results were in bold font, $p < 0.001$), and the difference between any two of the Gap-clarified, Gap-explained, and Gap-bridged categories in which students had the same level of prior progress (significant results were marked with the same symbol[3] in a row, $p < 0.001$).

	Strategy	All	Gap-clarified		Gap-explained		Gap-bridged	
			WP	WoP	WP	WoP	WP	WoP
1.	Evaluation Question	0.22	0.24	0.22	0.22	0.22	0.22	0.22
2.	Negative Feedback	0.39	0.43	◇ 0.42	**0.40**	◇ **0.37**	**0.40**	**0.38**
3.	Probing Question	0.44	♣ 0.46	♣ 0.48	† 0.45	† 0.46	† ♣ 0.43	† ♣ 0.44
4.	Open Question	0.49	**0.52**	◇ ♣ **0.63**	0.48	◇ 0.52	**0.47**	♣ **0.49**
5.	Observation Hint	0.56	♣ 0.60	◇ ♣ 0.61	† 0.58	◇ † 0.58	† ♣ 0.56	† ♣ 0.56
6.	Explanation	0.57	♣ 0.61	0.60	0.58	0.57	♣ 0.57	0.57
7.	Information Hint	0.62	◇ ♣ 0.65	◇ ♣ 0.65	◇ 0.62	◇ † 0.63	♣ 0.62	† ♣ 0.62
8.	Elaborated Positive Feedback	0.63	0.64	0.65	0.65	† 0.65	**0.63**	† **0.62**
9.	General Positive Feedback	0.72	♣ 0.74	◇ ♣ 0.75	† 0.73	◇ † 0.74	† ♣ 0.72	† ♣ 0.72

[3] We used three symbols to mark the statistical differences. ◇ marked the difference between Gap-clarified and Gap-explained, † marked the difference between Gap-explained and Gap-bridged, ♣ marked the difference between Gap-clarified and Gap-bridged

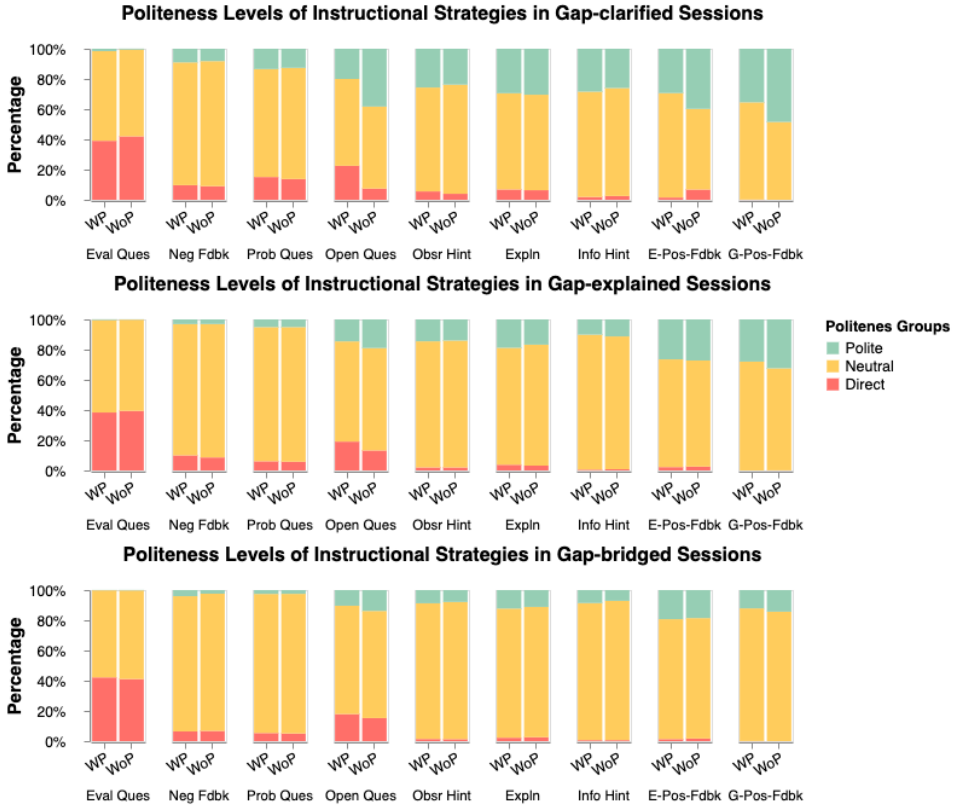


Figure 1: The distribution of the politeness groups (i.e., *Polite*, *Neutral*, and *Direct*) for each instructional strategy in different session categories. WP and WoP denote *With Prior Progress* and *Without Prior Progress*, respectively. The abbreviations include Eval-Ques (`Evaluation Question`), Neg-Fdbk (`Negative Feedback`), Prob Ques (`Probing Question`), Open Ques (`Open Question`), Obsr Hint (`Observation Hint`), Expln (`Explanation`), Info Hint (`Information Hint`), E-Pos-Fdbk (`Elaborated Positive Feedback`), and G-Pos-Fdbk (`General Positive Feedback`).

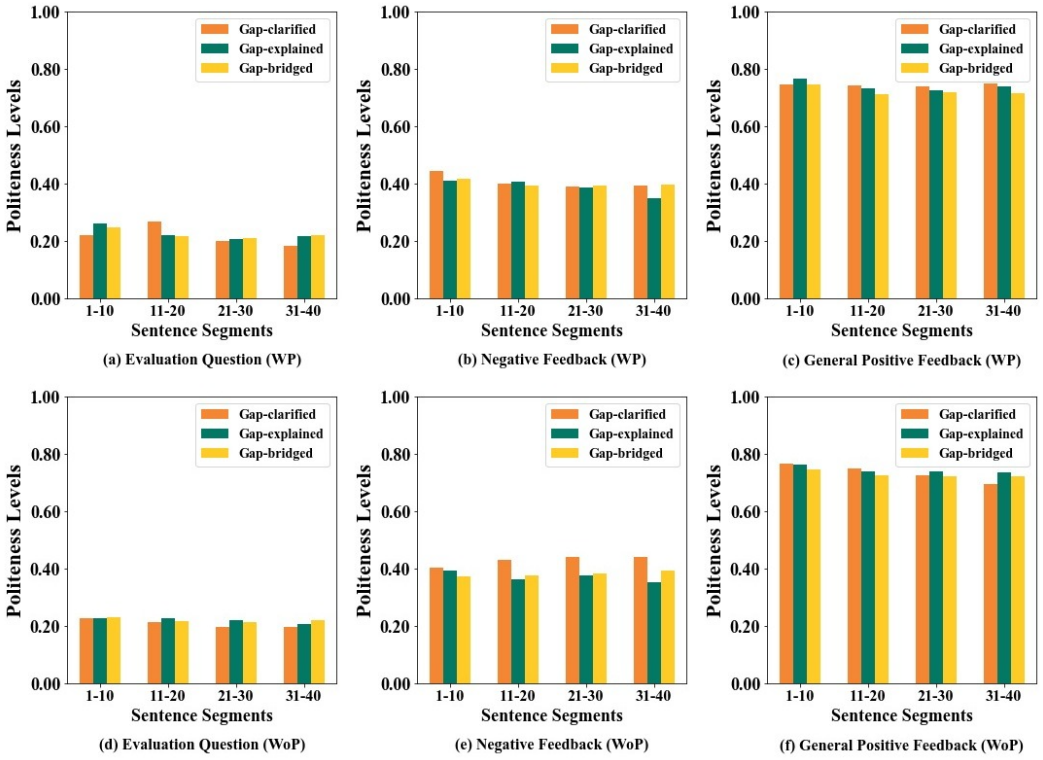


Figure 2: The distribution of tutors' sentences associated with different average politeness levels of instructional strategies as tutorial sessions progressed. WP and WoP denote *With Prior Progress* and *Without Prior Progress*, respectively.

`Evaluation Question`) were not used frequently by tutors so the occurrence of these strategies was also less frequent compared with the commonly-used instructional strategies (e.g., `General Positive Feedback` and `Probing Question`) as the tutoring progressed. To obtain less biased results, we computed the average politeness levels of every instructional strategy in the first [1, 10], [11, 20], [21, 30], and [31, 40] utterance segments for the three session categories. Due to the limited space, we only presented the results for the three instructional strategies (i.e., `Evaluation Question`, `Negative Feedback`, and `General Positive Feedback`) in Fig. 2. The results of other instructional strategies can be found in the digital appendix[4]. To examine the changes of strategies' politeness as tutoring progressed, Mann-Whitney U tests were applied to measure the difference (significant difference was measured by $p < 0.001$) between the two consecutive segments in the same session category. In the sessions where students had prior progress, tutors in Gap-bridged and Gap-explained sessions asked more polite `Evaluation Questions` in the utterance segment [1, 10] than in [11, 20] (Fig. 2 (a)) but no difference between segments [11, 20] and [21, 30] and between segments [21, 30] and [41, 40] were found. In comparison, in the sessions where students had no prior progress (Fig. 2 (d)), we did not find significant differences in politeness levels of the `Evaluation Question` strategy between the consecutive utterance segments across the three session categories. When scrutinising the strategy `Negative Feedback` (Fig. 2 (b) and (c)), we did not observe significant differences in politeness levels between the consecutive utterance segments in the three session categories. Regarding the strategy `General Positive Feedback` in Fig. 2 (c) and (f), tutors in Gap-bridged and Gap-explained sessions delivered this strategy more politely in the utterance segment [1, 10] than in [11, 20] but no significant differences were found between segments [11, 20] and [21, 30] and between segments [21, 30] and [41, 40].

4.2 Results on RQ2

We answered RQ2 from two different perspectives. Firstly, we investigated the effectiveness of the feature set (i.e., strategy-politeness features and non-politeness features) designed in Sec. 3.6 in predicting student performance. To this end, we decided to use the first N utterances (where N ϵ [5, 10, 15, 25, 30, 35, 40, *All*] utterances from the whole dialogues) to extract the features and those features were then used as the input for both Random Forest and GTB algorithms. As shown in Fig. 3 (solid lines), GTB outperformed Random Forest in predicting student performance when using all utterances from the tutoring sessions as the input and GTB achieved the performance of 0.810, 0.805, 0.812, and 0.643 as measured by Accuracy, F1 score, AUC, and Cohen's κ, respectively. To further examine the power of strategy-politeness features in predicting students' performance, we used the non-politeness features as the comparison to train both algorithms. The results (dash lines in Fig.

[4]https://bit.ly/lak22_tutor_appendix

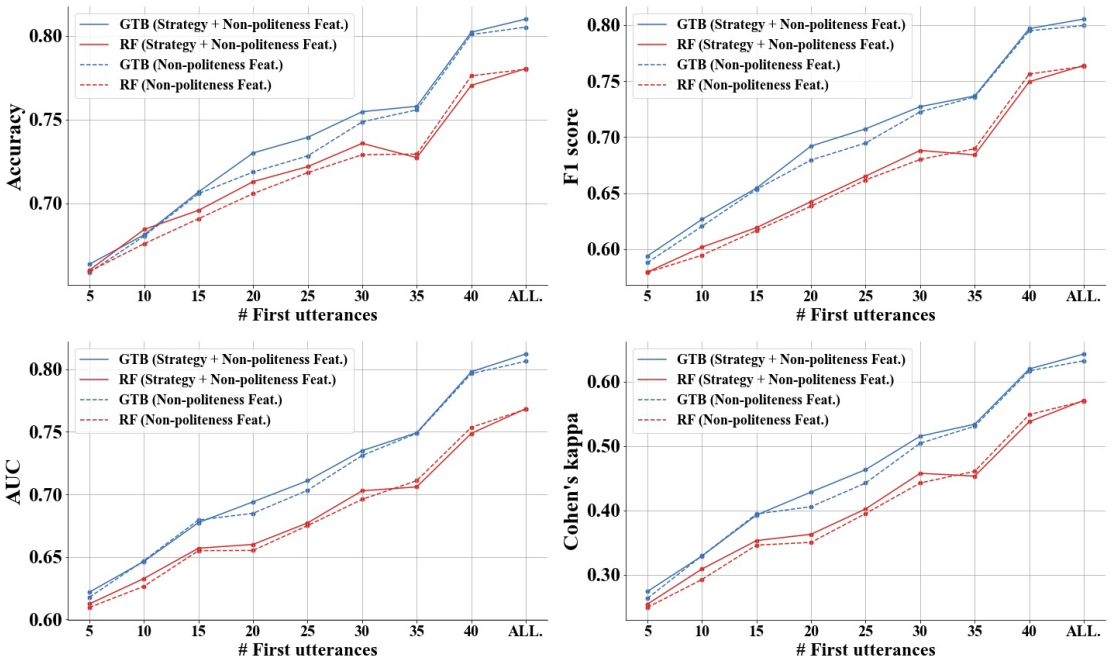


Figure 3: The performance of GTB and Random Forest in predicting student performance in solving problems.

3) indicated that the GTB model trained on non-politeness features achieved similar prediction performance at the earlier stage of the tutoring (i.e., $N = 5, 10, 15$) compared to that of the model trained on both strategy-politeness features and non-politeness features. This might be the reason that tutors commonly used more non-instructional communication (e.g., Greeting) at the outset of the tutoring [27]. Some instructional strategies in tutoring sessions might not be frequently used at the outset of the tutoring such as evaluation question (e.g., *"Does that make sense?"*). Then, it should be noted that the GTB model trained on the strategy-politeness features and non-politeness features outperformed than the model on the non-politeness features when N was 20, 25, and 30. This suggests that strategy-politeness features might help improve the prediction of students' problem-solving performance in the middle of the tutoring process.

To gain a better understanding of the predictive power of different types of features, we measured the contribution of features to the prediction by conducting the ablation test, which is described in Sec.3.6.4. As the instructional strategies were mostly employed in the middle of the tutoring process [27], we decided to investigate the ablation results when only considering the first 20 utterances as input. Here, instead of focusing on each feature, we were more interested in analysing the contributions made by each group of features. It should be noted that our previous study [21] had demonstrated contributions of non-politeness features. Hence, the current study did not expand the analysis on specific features in the non-politeness features. For the strategy-politeness features, not only do we present the feature importance of strategy-politeness features (i.e., Row 3, Table 4) but also break down the strategy-politeness features into three sub-groups of features (i.e., Row 4-6, Table 4) as described in Sec. 3.6.2. Though the non-politeness features contributed to the prediction performance the most, the strategy-politeness features also demonstrated the potentials to improve the model's performance in four metrics.

5 DISCUSSION AND CONCLUSION

The current study contributed with the following main findings that are of relevance to human-human online tutoring practice and the design of dialogue-based intelligent tutoring systems:

- Overall, tutors were inclined to use more direct expressions in the instructional strategies of comprehensive-gauging questions and negative feedback and use less direct expressions in the positive feedback.
- In tutorial sessions where students solved problems, tutors tended to ask more polite open questions and deliver more direct corrective feedback (negative feedback and elaborated positive feedback) for the students without prior progress than the students with prior progress.
- By incorporating the features of instructional strategies politeness with the features proposed in [21] (e.g., the sentiment levels of the utterance, complexity, and informativeness), the GTB model achieved better model performance of predicting student performance compared to our previous work [21],

5.1 Implications

The current study investigated the correlation between politeness and instructional communication from real-world human-human online dialogue tutoring. Our study provided guidelines for human tutors to consider the extent to which politeness can be expressed in instructional communication. Firstly, the findings suggested that

Table 4: The ablation test results of the GTB model when only considering the first 20 utterances in dialogue. Here, Row 1 of Table 4 presents the performance of the GTB model taking all strategy-politeness and non-politeness features as the input, and Rows 2-6 present the performance of the GTB model without taking a specific type of feature as input. The percentages in the brackets were calculated by taking the model performance of considering all features (i.e., Row 1) as the comparison. The results with the maximum performance decrease are in bold.

Feature	Accuracy	F1 score	Cohen's k	AUC
1. Strategy + Non-politeness Feat.	0.730	0.692	0.428	0.694
2. w/o Non-politeness Feat.	**0.687 (-5.89%)**	**0.623 (-9.97%)**	**0.292 (-31.78%)**	**0.627 (-9.65%)**
3. w/o Strategy-politeness Feat.	0.722 (-1.10%)	0.682 (-1.45%)	0.413 (-3.50%)	0.689 (-0.72%)
4. w/o Strategy-politeness Feat. (*Mean*)	0.728 (-0.27%)	0.692 (0.00%)	0.427 (-0.23%)	0.694 (0.00%)
5. w/o Strategy-politeness Feat. (*Median*)	0.726 (-0.55%)	0.689 (-0.43%)	0.421 (-1.64%)	0.692 (-0.29%)
6. w/o Strategy-politeness Feat. (*Std.*)	0.723 (-0.96%)	0.684 (-1.16%)	0.413 (-3.50%)	0.687 (-1.01%)

tutors could use direct expression for the strategies `Evaluation Questions` and `Negative Feedback` as long as the instruction that tutors deliver meets the students' needs of solving problems, which is in line with the claim by [8]. As human tutors were reluctant to impose unpleasant feelings on students [7], they often spent much time on the word formulation of some instructional communication (e.g., providing negative feedback and directive hints) in tutoring which might hinder the efficiency of the tutoring [7, 8]. Our study found that the tutors commonly guided students in less polite expressions for the instructional strategies `Evaluation Questions` and `Negative Feedback`. The strategy `Negative Feedback` was used to help the students identify errors (e.g., *"No, it is wrong"*). It is inevitable to be direct when using this strategy in order to point out the students' errors even though this tone of expression might appear to neglect the students' expectation for the appreciation [8].

Secondly, the current study showed that tutors asked more polite open questions to guide students without prior progress than the ones with prior progress. The strategy `Open Question` was used to gauge students' understanding and progress on the problem (e.g., *"Why do you use this formula?"*). Such questions could elicit students' self-explanation (i.e., students explain their understanding of the discussed topics) which proven to have benefits for learning [2, 10]. It should be noted that students with low prior knowledge are not confident to demonstrate self-explanation activities [2] and they commonly engaged more with polite tutors [24, 32]. As our study investigated the student prior progress as the proxy of student prior knowledge, we suggested that tutors should provide more polite open questions (e.g., *"What do you think we could try next?"*) to the students without prior progress compared to the students with prior progress.

Thirdly, our results (Fig. 1) showed that most tutors delivered positive feedback in a non-direct form. The strategy `Positive Feedback` was used to verify the students' correctness and encourage them to make more effort on the problems. Providing positive feedback in a direct form (e.g, *"Right, that's exactly what you should get"*) might mitigate the advantages of positive feedback in building a friendly relationship with students since students expected receiving the praise and approval from tutors when students made correct steps [40]. In the future, it is worth further investigating the relationship between the use of politeness in positive feedback and students' learning experience by designing randomised controlled experiments.

The results of the current study shed some light on the design of future dialogue-based intelligent tutoring system (ITS) about how to express politely in the instructional process. Most of the existing studies have only examined instructional politeness in a binary form (i.e., either polite or direct) [23, 24, 26, 36, 38, 39] but neglected that instructional communication, in real-world tutoring, could be expressed in different surface forms, which can represent different politeness levels [33]. Based on the results of the current study, we submit that dialogue-based ITSs should not only focus on a single form of polite or direct instruction. In particular, it is important to make the trade-off between polite and direct expression in the tutoring since the polite tutors are more approachable for students[6, 8] and direct tutors made the tutoring sessions more efficient [7, 8, 30]. To this end, it was necessary to understand the politeness levels of instructional strategies. Our study demonstrated the politeness levels of instructional strategies by observing human expert instruction, which was a typical method to further develop the design of ITS [13].

Finally, our prediction results demonstrated that the strategy-politeness features had the potential to predict the student's performance in the middle of the tutoring. In particular, by incorporating the strategy-politeness features in our feature set, the GTB model obtained the improvement of model performance in four evaluation metrics (e.g., F1 score,the AUC, and Cohen's κ) when considering the first 20 utterances as input. The results indicated that the use of politeness in the instructional strategies might correlate with the students' performance. Therefore, we suggested that the politeness of instructional strategies should be incorporated in the future work of predicting student performance.

5.2 Limitations

We acknowledge there are several limitations in the current study. Firstly, our results were limited to the United States cultural background. The tutors in our dataset were trained to teach students politely followed the policy of the educational technology company based in US. However, the students from different cultural backgrounds (e.g., those Asian countries) might have different perceptions of politeness. It is necessary to conduct a study investigating the variance of politeness levels in tutoring communication from the perspective of other cultures. Then, we did not analyse politeness across different subject areas (e.g., maths, chemistry, and physics). Our dataset contained the maths tutoring in more than

92% of the tutoring sessions, so our results might demonstrate the patterns in relation to maths tutoring. In our future work, it is worth to collect more dialogue sessions in other subjects (e.g., physics and chemistry) and analyse these dialogues separately regarding the different subjects. Thirdly, our study only examined the politeness levels of instructional strategies from the tutors' perspective. It still remains unknown whether the students prefer collaborating with the tutor who expressed the politeness in the instructional communication based on our analysed results. To further guide the practice of tutoring and the development of dialogue-based ITS, a randomised experiment should be conducted to investigate the students' preference on the politeness of instructional communication. Lastly, since we measured students' prior progress from the first few utterances, it may happen that some students started talking about this progress outside of that window, i.e., after the utterances that have been taken for the analysis.

5.3 Future Work

Our study obtained the results by examining real-world human-human tutoring. The results demonstrated that most tutors expressed politely in some instructional strategies such as positive feedback. However, some tutors might deliver positive feedback in a direct form (e..g, *"Right, that's what you should get"*), which might lead to negative effects on tutoring. Therefore, we plan to construct a politeness transfer tool to automatically convert non-polite tutors' expressions into polite forms. Then, as indicated in Sec. 3.6.2, we did not incorporate students' prior progress in the feature set for student performance prediction. In the future work, we plan to build a machine learning model to automatically identify the students' prior progress in the tutoring process and include it as an additional feature for predicting student performance.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*. 3606–3611.

[2] Kirsten Berthold, Tessa HS Eysink, and Alexander Renkl. 2009. Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science* 37, 4 (2009), 345–363.

[3] Tasha Bleistein and Marilyn Lewis. 2014. *One-on-one language teaching and learning: Theory and practice*. Springer.

[4] Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher* 13, 6 (1984), 4–16.

[5] Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition* 22, 6 (1996), 1482.

[6] Penelope Brown and Stephen Levinson. 1987. Politeness: some universals in language usage. Cambridge University Press, Cambridge, UK.

[7] Benjamin Brummernhenrich and Regina Jucks. 2013. Managing face threats and instructions in online tutoring. *Journal of Educational Psychology* 105, 2 (2013), 341.

[8] Benjamin Brummernhenrich and Regina Jucks. 2016. "He shouldn't have put it that way!" How face threats and mitigation strategies affect person perception in online tutoring. *Communication Education* 65, 3 (2016), 290–306.

[9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 2016 Conference on KDD*. 785–794.

[10] Jennifer L Chiu and Michelene TH Chi. 2014. Supporting self-explanation in the classroom. *Applying science of learning in education: Infusing psychological science into the curriculum* (2014), 91–103.

[11] Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165, 2 (2014), 97–135.

[12] Thomas G Dietterich et al. 2002. Ensemble learning. *The handbook of brain theory and neural networks* 2 (2002), 110–125.

[13] Benedict Du Boulay and Rosemary Luckin. 2016. Modelling human teaching tactics and strategies for tutoring systems: 14 Years on. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 393–404.

[14] Judi Simmons Estes. 2017. Teacher preparation programs and learner-centered, technology-integrated instruction. In *Handbook of research on learner-centered pedagogy in teacher education and professional development*. IGI Global, 85–103.

[15] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research* 15, 1 (2014), 3133–3181.

[16] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[17] Samantha Jiménez, Reyes Juárez-Ramírez, Victor H Castillo, and Juan José Tapia Armenta. 2018. *Affective Feedback in Intelligent Tutoring Systems: A Practical Approach*. Springer.

[18] Regina Jucks, Lena Päuler, and Benjamin Brummernhenrich. 2016. "I need to be explicit: You're wrong": Impact of face threats on social evaluations in online instructional communication. *Interacting with Computers* 28, 1 (2016), 73–84.

[19] Jionghao Lin, David Lang, Haoran Xie, Dragan Gašević, and Guanliang Chen. 2020. Investigating the Role of Politeness in Human-Human Online Tutoring. In *Artificial Intelligence in Education*. Springer International Publishing, Cham, 174–179.

[20] Jionghao Lin, Shirui Pan, Cheng Siong Lee, and Sharon Oviatt. 2019. An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. ACM, New York, NY, USA, 2069–2072.

[21] Jionghao Lin, Shaveen Singh, Lele Sha, Wei Tan, David Lang, Dragan Gašević, and Guanliang Chen. 2022. Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems* 127 (2022), 194–207.

[22] Richard E Mayer, W Lewis Johnson, Erin Shaw, and Sahiba Sandhu. 2006. Constructing computer-based tutors that are socially sensitive: Politeness in educational software. *International Journal of Human-Computer Studies* 64, 1 (2006), 36–42.

[23] Bruce M McLaren, Krista E DeLeeuw, and Richard E Mayer. 2011. Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education* 56, 3 (2011), 574–584.

[24] Bruce M McLaren, Krista E DeLeeuw, and Richard E Mayer. 2011. A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies* 69, 1-2 (2011), 70–79.

[25] Bruce M McLaren, Sung-Joo Lim, David Yaron, and Kenneth R Koedinger. 2007. Can a polite intelligent tutoring system lead to improved learning outside of the lab? *Frontiers in Artificial Intelligence and Applications* 158 (2007), 433.

[26] Maria Mikheeva, Sascha Schneider, Maik Beege, and Günter Daniel Rey. 2019. Boundary conditions of the politeness effect in online mathematical learning. *Computers in Human Behavior* 92 (2019), 419–427.

[27] Donald M Morrison, Benjamin Nye, Vasile Rus, Sarah Snyder, Jennifer Boller, and Kenneth Miller. 2015. Tutorial dialogue modes in a large corpus of online tutoring transcripts. In *International Conference on Artificial Intelligence in Education*. Springer, 722–725.

[28] Kimberly A Neuendorf. 2017. *The Content Analysis Guidebook*. SAGE Publications.

[29] Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the ACL* 6 (2018), 373–389.

[30] Natalie K Pearson, Roger J Kreuz, Rolf A Zwaan, and Arthur C Graesser. 1995. Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and instruction* 13, 2 (1995), 161–188.

[31] Edd Pitt and Lin Norton. 2017. 'Now that's the feedback I want!'Students' reactions to feedback on graded work and what they do with it. *Assessment & Evaluation in Higher Education* 42, 4 (2017), 499–516.

[32] Kaśka Porayska-Pomsta, Manolis Mavrikis, and Helen Pain. 2008. Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction* 18, 1-2 (2008), 125–173.

[33] Kaśka Porayska-Pomsta and Chris Mellish. 2004. Modelling politeness in natural language generation. In *International Conference on Natural Language Generation*. Springer, 141–150.

[34] Kaska Porayska-Pomsta and Helen Pain. 2004. Providing cognitive and affective scaffolding through teaching strategies: applying linguistic politeness to the educational context. In *International conference on intelligent tutoring systems*. Springer, 77–86.

[35] Vasile Rus, Nabin Maharjan, Lasang Jimba Tamang, Michael Yudelson, Susan Berman, Stephen E Fancsali, and Steve Ritter. 2017. An analysis of human tutors' actions in tutorial dialogues. In *The Thirtieth International Flairs Conference*.

[36] Sascha Schneider, Steve Nebel, Simon Pradel, and Günter Daniel Rey. 2015. Mind your Ps and Qs! How polite instructions affect learning with multimedia. *Computers in Human Behavior* 51 (2015), 546–555.

[37] Alexandria Katarina Vail and Kristy Elizabeth Boyer. 2014. Identifying Effective Moves in Tutoring: On the Refinement of Dialogue Act Annotation Schemes. In *Intelligent Tutoring Systems*.
[38] Ning Wang and W Lewis Johnson. 2008. The Politeness Effect in an intelligent foreign language tutoring system. In *ITS*. Springer, 270–280.
[39] Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International journal of human-computer studies* 66, 2 (2008), 98–112.
[40] Ning Wang, W Lewis Johnson, Paola Rizzo, Erin Shaw, and Richard E Mayer. 2005. Experimental evaluation of polite interaction tactics for pedagogical agents. In *Proceedings of the 10th international conference on Intelligent user interfaces*. 12–19.
[41] Qi Wang and Liping Shen. 2018. Student proficiency prediction on CNMOOC data. In *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*. 15–18.
[42] Fan Yang and Frederick WB Li. 2018. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education* 123 (2018), 97–108.

5.3 Chapter Summary

Prior research has documented the benefits of expressing politeness in feedback to the learning process [60–64] and hindrances of excessive polite expressions to tutoring efficacy [65–67]. There is a shortage of understanding about the extent to which the use of politeness in instructional communication can contribute to learners' achievement. Given that we found the importance of politeness (Chapter 2) and dialogue acts (Chapter 3) in the tutoring dialogues, we further investigated the use of politeness in instructional communication to maintain the effectiveness of tutoring dialogues. We first employed the dialogue act classifier developed in Chapter 3 to identify dialogue acts for educators' sentences. Secondly, we selected the educators' sentences which were identified as instructional dialogue acts (e.g., positive feedback and probing question) based on the work by Brummernhenrich and Jucks [67]. Then, we employed the politeness scoring tool developed by [83] to identify the politeness levels for the selected educators' sentences. By doing so, we can investigate the politeness levels in different instructional dialogue acts, which was summarised as the **Contribution 7**. In line with Chapters 2 and 3, we also evaluated the predictive power of politeness in instructional dialogue acts and our results found that the politeness of instructional dialogue acts could enhance the prediction of the learner's performance, which was summarised as the **Contribution 8**. The results in Chapter 5 can provide guidelines for educators to consider how politeness can be expressed in instructional communication and shed light on the design of dialogue-based ITSs about how to politely communicate with learners.

Contribution 7: Track the use of politeness in instructional communication.

1. Educators were inclined to use more direct expressions in the instructional strategies of evaluation questions (e.g., "*Does that make sense?*") and negative feedback (e.g., "*No, it is wrong*") and use less direct expressions in the positive feedback (e.g., "*Well done*").
2. In tutorial sessions where learners solved problems, educators tended to ask more polite open questions (e.g., "*What do you think*") and deliver more direct corrective feedback (negative feedback and elaborated positive feedback) for the learners without prior progress than the learners with prior progress.

Contribution 8: The capability of using politeness in instructional communication on the prediction of learner problem-solving performance.

1. By incorporating the features of instructional strategies politeness with the features proposed in [38] (e.g., the sentiment levels of the sentence, complexity, and informativeness), the GTB model improved its reliability of predicting learner problem-solving performance compared to our previous work [38].

Chapter 6

Analysing the Communicative Patterns in Assessment Feedback

6.1 Introduction

Given the importance of communicative patterns (politeness and dialogue acts) as reported in Chapters 2–5, we also aimed to investigate the communicative patterns in assessment feedback. As discussed in Gap 3, assessment written feedback is a widely-used form of feedback in higher education [27]. However, we argue that the prior works [32–35] of analysing communicative patterns in assessment written feedback are insufficient to demonstrate the insights on the design of dialogic feedback in assessment written assessments. The reason is that the analysed feedback in most prior works was given in the one-way form (i.e., without allowing learners to submit their improved work based on the feedback they received). However, dialogic feedback is often given in a two-way form and often places the emphasis on the interactive process of communication and exchange of ideas between the educator and the learner [1, 22]. Additionally, the existing discussions on how effective feedback content can be constructed to support learners were still insufficient [33–35]. Distinct from the prior works [32–35], in Chapter 6, we aimed to address Gap 3 by employing a learner-centred feedback framework proposed by [115] to analyse communicative patterns in assessment feedback (RQ 5).

The learner-centred feedback framework [115] has summarised the most well-known feedback research in the past decade and identified attributes of communicative patterns in feedback. To answer RQ 5, we collected the educators' feedback on two consecutive assignments offered by an introductory to data science course at the postgraduate level. The first assignment (namely, Assignment I) was about drafting a report proposal and the second assignment (namely, Assignment II) was the final version of the report. On the basis of Assignment I, we could measure the learner grade changes on Assignment II, i.e., whether learners' grades increased or not on

Assignment II. Inspired by the work [32], we used the learner grade changes as the proxy of the learning outcomes. Relying on the Ryan et al. framework [115], we engineered and extracted features from the feedback content on Assignment I and further examined the differences in these features based on learning outcomes. Furthermore, we used the extracted features as inputs for training machine learning models used in the previous works [32–35] to predict learning outcomes. Beyond the prediction, we also provided the interpretation of predicted results by using the **SH**apley **A**dditive ex**P**lanations (SHAP) framework [104].

This research has been published in the Proceedings of the *13th International Learning Analytics and Knowledge Conference*.

- Lin, J., Dai, W., Lim, L., Tsai, Y., Mello, R., Khosravi, H., Gašević, D., & Chen, G. (2023). Learner-centred Analytics of Feedback Content in Higher Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. (100-110).

6.2 Publication: Learner-centred Analytics of Feedback Content in Higher Education.

Learner-centred Analytics of Feedback Content in Higher Education

Jionghao Lin
jionghao.lin1@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Wei Dai
wei.dai1@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Lisa-Angelique Lim
lisa-angelique.lim@uts.edu.au
University of Technology Sydney
Sydney, Australia

Yi-Shan Tsai
yi-shan.tsai@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Rafael Ferreira Mello
rafaelflmello@gmail.com
CESAR School
Recife, Brazil
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Hassan Khosravi
h.khosravi@uq.edu.au
The University of Queensland
St Lucia, QLD, Australia

Dragan Gašević
dragan.gasevic@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

Guanliang Chen*
guanliang.chen@monash.edu
Centre for Learning Analytics,
Monash University
Melbourne, Australia

ABSTRACT

Feedback is an effective way to assist students in achieving learning goals. The conceptualisation of feedback is gradually moving from *feedback as information* to *feedback as a learner-centred process*. To demonstrate feedback effectiveness, feedback as a learner-centred process should be designed to provide quality feedback content and promote student learning outcomes on the subsequent task. However, it remains unclear how instructors adopt the learner-centred feedback framework for feedback provision in the teaching practice. Thus, our study made use of a comprehensive learner-centred feedback framework to analyse feedback content and identify the characteristics of feedback content among student groups with different performance changes. Specifically, we collected the instructors' feedback on two consecutive assignments offered by an introductory to data science course at the postgraduate level. On the basis of the first assignment, we used the status of student grade changes (i.e., students whose performance increased and those whose performance did not increase on the second assignment) as the proxy of the student learning outcomes. Then, we engineered and extracted features from the feedback content on the first assignment using a learner-centred feedback framework and further examined the differences of these features between different groups of student learning outcomes. Lastly, we used the features to predict student learning outcomes by using widely-used machine learning models and provided the interpretation of predicted results by using the SHapley Additive exPlanations (SHAP) framework. We found that 1) most features from the feedback content presented significant differences between the groups of student learning outcomes, 2) the gradient boost tree model could effectively predict student learning outcomes, and 3) SHAP could transparently interpret the feature importance on predictions.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Feedback, Learning Analytics, Content Analysis, Interpretability

ACM Reference Format:
Jionghao Lin, Wei Dai, Lisa-Angelique Lim, Yi-Shan Tsai, Rafael Ferreira Mello, Hassan Khosravi, Dragan Gašević, and Guanliang Chen. 2023. Learner-centred Analytics of Feedback Content in Higher Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13–17, 2023, Arlington, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576050.3576064

1 INTRODUCTION

Feedback has been acknowledged as an effective component in promoting students' learning in higher education [5, 11, 17, 41]. The conceptualisation of feedback in the existing literature has gradually shifted from feedback as information to feedback as a learner-centred process [41]. The earlier definition (i.e., feedback as information) of effective feedback focuses on the information

*Corresponding Author

LAK 2023, March 13–17, 2023, Arlington, TX, USA

ACM ISBN 978-1-4503-9865-7/23/03…$15.00
https://doi.org/10.1145/3576050.3576064

and timeliness [16], whereas the recent understanding (i.e., feedback as the learner-centred process) of feedback focuses more on the student's learning process [41]. Many previous studies found that feedback as a learner-centred process could lead to more effective feedback design to promote learners' achievements compared with the feedback as information [1, 2, 12]. To better understand the construction of learner-centred feedback, Dawson et al. [11] interviewed students and instructors about their experience on the feedback effectiveness. Students indicated that the quality of feedback content (e.g., actionable information for improvement on subsequent tasks) was the most effective aspect of feedback, whereas instructors indicated that feedback should be provided in the context of connected tasks (i.e., the feedback provided for the first task is connected with the subsequent task) [11]. These indications of effective feedback were further affirmed by a recent learner-centred feedback framework [35], which demonstrated a comprehensive set of effective feedback attributes (e.g., comments that provide actionable information for future performance) for textual feedback in higher education.

Effective feedback should deliver high-quality content and demonstrate the effect on student improvement such as improvement in learning outcomes and conceptual understanding [11, 35]. To understand how feedback content affects student improvement, researchers employed Learning Analytics (LA) approaches to analyse the correlation between feedback content and the students' learning outcomes. For example, Nicoll et al. [30] extracted textual features (e.g., N-grams) from feedback on the first assignment and measured student grade changes (e.g., grade increase and decrease) on the subsequent assignment as the proxy of the students' learning outcomes. Then, they [30] used LA approaches to analyse the correlation between the textual features and student grade changes. Though Nicoll et al. [30] demonstrated a method to understand the effect of feedback content on student learning outcomes, we argued that there was no explicit discussion between the textual features in their work [30] and the theoretically grounded attributes in the learner-centred feedback framework [35]. Additionally, the feedback effect (i.e., improvement in students' learning outcomes on the subsequent task) was not substantial since the connection between the two assignments was unclear in their work. Therefore, our study aimed to identify the textual features of feedback content based on a learner-centred framework and examine the use of these features among different groups categorised by students' subsequent learning outcomes on the consecutive assignment.

To facilitate the delivery of effective feedback, LA researchers also adopted machine learning models to automate the process of evaluating feedback quality [8, 30, 31]. Though previous works achieved promising model prediction performance [8, 30, 31], it is important to note that, beyond the prediction, interpreting how textual features contribute to the predicted result is also important as the interpretability of results could further enhance human trust in the analysis results[23, 24, 34]. The interpretation for the predicted results can be provided from the perspective of local and global level [18, 23]. The global interpretation can provide feature importance to the model prediction performance, whereas local interpretation can demonstrate feature effects on individual predictions [18]. Most previous LA studies on feedback demonstrated global interpretations on predicting effective feedback measures [8, 30, 31]. However, the local interpretation was under-explored in these works. According to [23], local interpretation can show the extent to which the features had positive or negative effects on individual predictions, which can enable human users to better understand why a model made a decision instead of blindly trusting the model's prediction output [23]. We deemed that the interpretation at the local level can allow a better understanding for instructors to decide which feedback comments should be modified. Therefore, our study aimed to advance existing knowledge by providing the interpretation of the predicted results at both global and local levels.

To this end, our study aimed to analyse textual features of feedback content based on a learner-centred feedback framework [35], examine the use of textual features in different groups based on students' learning outcomes on the subsequent assignment, and interpret the prediction results of student grade changes. Formally, we aimed to answer the following **R**esearch **Q**uestions:

- **RQ 1** To what extent are features of feedback related to the student grade changes on the consecutive assignment?
- **RQ 2** To what extent do the features of feedback content contribute to the predictions of student grade changes?

Our study answered the above research questions by using a dataset containing feedback on two connective assignments in a higher education course. We analysed feedback content on the first assignment and observed student grade changes (i.e., increase or not increase) on the subsequent assignment, which is connected to the first assignment. To answer **RQ1**, we first employed the learner-centred feedback framework [35] to extract features about feedback content. Then, we examined the use of the features among different groups based on student grade changes on the subsequent assignment. Our results revealed that the use of some extracted features presented significant differences in different student groups regarding the student's grade changes. Additionally, the student performance on the first assignment was considered the most important indicator for differentiating improvement on the second connective assignment. For **RQ2**, the extracted features were further used as inputs for training machine learning models used in the previous works [7, 8, 30, 31] to predict students' grade changes on the subsequent assignment. Then, we used the interpretable framework **SHAP** (**SH**apley **A**dditive ex**P**lanations) [24] to provide the interpretation of predicted results. We found that the **GTB** (**G**radient **T**ree **B**oosting) model achieved better performance in predicting student grade changes compared to other selected models. The **SHAP** framework can transparently interpret the feature effect at global and local levels. For example, with the use of the **SHAP** framework, our study identified the positive emotional words in the feedback as one of the most significant features for model prediction performance (i.e., global-level interpretation) and also demonstrated that the high frequency of positive emotional words in feedback negatively correlated with the student grade increase on the subsequent assignment (i.e., local-level interpretation).

2 BACKGROUND

2.1 Conceptualisation of Feedback

Feedback has been widely acknowledged as a crucial part of students' learning achievements in higher education [11, 13, 17, 20,

32, 35, 39]. When examining the effectiveness of feedback, it is significant to understand the existing literature on feedback practice so as to identify potential indicators [39]. One popular definition of feedback by Hattie and Timperley [16] conceptualised effective feedback as the information which can help students to minimise the gap between current and expected performance [41]. They [16] proposed a well-known *Feedback Model* to characterise effective feedback into three types (i.e., *feeding back*, *forward* and *up*) and further into four-level focuses (i.e., *task*, *process*, *self-regulatory*, and *self* focus). The *Feedback Model* [16] places emphasis on the provision of task-relevant information. The recent literature on feedback has been gradually shifting to feedback as a learner-centred process (e.g., *Mark 2 model* [2], *Dialogic Triangle model* [42], *Feedback Literacy Framework* [6], and the *Typology of Feedback Impact* [17]) where students could make sense of the information on their work, and use feedback to further improve the quality of their future work [32, 41]. The reason behind this shift is that providing feedback as a learner-centred process is more beneficial to learners than as information [35, 41].

To better understand the design of the learner-centred feedback framework, Ryan et al. [35] reviewed the feedback publications from the past decade. They [35] summarised a comprehensive framework of learner-centred feedback, which can be organised into two main design layers: *context* and *artefact*. The *context* focuses on the feedback design related to the process of feedback provision (e.g., frequency and timeliness), while the *artefact* focuses on the attributes of the feedback content. As the focus of this study is on the feedback content, we mainly focused on the *artefact* layer, which involved nine attributes as shown in Table 1.

2.2 Content Analysis on Feedback

Recent works have employed content analysis to examine textual feedback content [7, 8, 30, 31]. For example, Cavalcanti et al. [7] analysed the quality of 1,000 feedback collected from an online course provided by a Brazilian higher education institution. They [7] used the textual features (e.g., features from LIWC and Coh-Metrix) to train a Random Forest model to identify whether the feedback contained the seven principles of good feedback practices proposed by Nicol and Macfarlane-Dick [29], and their work [7] achieved 0.75 classification accuracy and 0.20 Cohen's κ on the testing dataset. Following the study [7], Cavalcanti et al. [8] claimed that the seven principles [29] were too general to annotate the written feedback. Instead, they further annotated the same feedback dataset from [7] based on the Hattie and Timperley [16] four feedback focuses. To automate the process of identifying four feedback focuses, Cavalcanti et al. [8] used the same textual features from [7] to train a Random Forest model and they [8] reached 0.75 and 0.32 for the averaged classification accuracy and Cohen's κ, respectively, on the testing dataset. Later on, Osakwe et al. [31] further extended this work [8] by using a more powerful machine learning model (i.e., Gradient Tree Boosting [9]) and reached 0.83 and 0.39 for the averaged accuracy and Cohen's κ, respectively.

Despite demonstrating the automated process of evaluating feedback quality based on the feedback practice (i.e., seven principles [29] and four level focuses [16]), these studies [7, 8, 31] draw limited conclusions on evaluating feedback quality based on the feedback effect (e.g., student performance changes on future tasks), which can shed light on the feedback design and automatic feedback generation [30]. To this end, Nicoll et al. [30] examined the correlation between the textual features and student grade changes on the subsequent tasks. They extracted textual features (e.g., N-grams and sentiment scores) from feedback and used these features to train a logistic regression model to predict students' grade changes (e.g., increase, no change, and decrease) on the subsequent task [30]. The model reached 0.79 and 0.76 for averaged classification accuracy and AUC, respectively [30]. Though Nicoll et al. [30] demonstrated a method for understanding the impact of feedback content, further insights are still missing regarding feedback practice. First, textual features (e.g., N-grams) in their work [30] were engineered from the linguistic perspective, which lacked a connection with the existing feedback literature. Additionally, the features (e.g., N-grams) used for training the machine learning model were dataset dependent, which might lead to overfitting issues [8, 19, 31]. Thirdly, the connection between the first and subsequent assignments in their work [30] was unclear. According to [17], feedback might not influence student improvement when the first and subsequent assignments are different. Lastly, the interpretation of feature importance in [30] did not demonstrate whether these features were positively or negatively correlated with the student grade changes, which might be hard for educational researchers to understand the impact of feedback content.

To advance existing literature, our study aimed to identify the features of feedback content given on the first assignment and investigate the use of these features in different groups categorised by the student grade changes on the subsequent connective assignment. Instead of using the data-dependent features, we aimed to select textual features in relation to the *artefact* attributes based on a comprehensive learner-centred feedback framework developed in [35]. Furthermore, we aimed to employ the widely-used interpretable framework by [24] to demonstrate feature importance on the model prediction and the features' correlation polarity (i.e., positive or negative) on the output.

3 METHOD

3.1 Data Preparation

Our study obtained ethical approval (ID:29874) from Monash University and collected instructors' rubric feedback from an introductory data science course taught in English at the postgraduate level. This course was offered for two semesters throughout the year, and students were required to work on several assignments (e.g., coding practice and writing a report) for each semester. The marks of all assignments accounted for 40% of the course, and each assignment was marked separately.

Inspired by the learner-centred feedback framework, i.e., effective feedback should be designed in a connective form (i.e., the comments of the current feedback can help students make improvements in the subsequent assignment) [11, 35], our study mainly investigated the feedback on the report assignment which required students to write a report proposal and a final report. To clarify the difference, we named the reported proposal **Assignment I** (counted 5% marks) and the final report **Assignment II** (counted 10% marks). Students were required to submit a report proposal for

Table 1: The description of *artefact* attributes based on the learner-centred feedback framework developed by Ryan et al. [35]

Row ID	Artefact Attributes	Description
Attr.1	**Strengthen teacher and learner relationships**	builds social relationship with learners
Attr.2	**Encourage learner agency**	influences the way where learners attend to the comment details
Attr.3	**Encourage positive learner affect**	shows the emotional response and care on learners' feelings
Attr.4	**Highlight strengths of performance**	acknowledges learners' strength in their works to enhance learners' self-efficacy
Attr.5	**Provide critiques about performance**	demonstrates critiques on learners' works to develop learners' evaluative judgement
Attr.6	**Provide actionable information for future performance**	helps learners achieve expected learning goals, develop learning strategies, and obtain improvement on the similar learning task
Attr.7	**Promote learner independence**	rephrases comments as questions or suggestions rather than statements to encourage learners to think for themselves
Attr.8	**Usable for learners**	rephrases comments in clear language, easy for learners to understand, relevant to the marking criteria, and explanatory to learners
Attr.9	**Invite dialogue about feedback**	encourages learners to ask questions and seek help from others (e.g., family and peers) to have feedback dialogues

Table 2: Students' demographics information

Year	Semester	# Students	# Female	# Male	#Domestic	#International
2021	S2	145	47	98	17	128
2022	S1	127	36	91	10	117
Total	-	**272**	**83**	**189**	**27**	**245**

the assignment I to introduce a data science problem to be solved and describe relevant application background and types of business models. After students submitted assignment I, instructors marked the students' assignment I and provided feedback to help students work on assignment II. In assignment II, students could use the feedback from assignment I to complete the remaining parts of the report, such as the description of the data analysis. After students submitted assignment II, instructors graded students' work and provided corresponding feedback. Both assignments were assessed using the same rubric[1]. However, as assignment I only required students to complete part of report writing, the instructor also used the subset of the rubric for assessing assignment I. Our study investigated the feedback content on assignment I (i.e., report proposal). We collected the feedback data from Semester 2 (Jul–Oct) in 2021 and Semester 1 (Feb–May) in 2022 since the assignment descriptions and rubric were the same for both semesters. In total, we had 288 student records collected by the learning management system. We filtered out 16 data records with two issues: 1) the records missed grades and feedback for the assignment I, and 2) duplicate records. Then, we obtained 272 records (83 female and 189 male). The students' demographics in our dataset are presented in Table 2.

3.2 Representation of Students' Grade Changes

Nicoll et al. [30] proposed a method to observe student grade changes by subtracting the students' grades in the first task from the grades in the second task. Inspired by their work [30], we also measured student grade changes by subtracting the students' grades in the assignment I from their grades in assignment II. The students' assignments I and II were marked separately by the same instructor. According to the course assessment policy, each assignment was marked into five categories from low to high: Fail (**N**, scoring from 0-49), Pass (**P**, 50-59), Credit (**C**, 60-69), Distinction (**D**, 70-79), and High Distinction (**HD**, 80-100). We encoded five categories into a numerical scale (i.e., **N** = 0, **P** = 1, **C** = 2, **D** = 3, and **HD** = 4) to calculate the grade changes between two assignments. The positive values indicated students achieved performance increase on assignment II, and we encoded the records of the positive values as the `Increase` group. Whereas, the other calculated results were encoded as the `Not Increase`. However, it should be noted that 60 students were graded **HD** marks in both assignments I and II. We argue that maintaining the outstanding performance (i.e., **HD**) on the subsequent assignment is also an improvement but the performance increase could not be directly observed from the students' subsequent performance change for these high-performing students. The feedback for these 60 high-performing students could be more concise (e.g., a line of comment, "*Well done!*") compared with other students, which might influence the analysis results. Thus, we decided to filter out the feedback on these 60 students and planned to investigate them in future work. As a result, we obtained 212 records for analysis where 66 records were in the `Increase` and 146 in `Not Increase` group.

3.3 Mapping Artefact Level Features

To answer RQ1, we used a set of textual features extracted from feedback content to map against the *artefact* attributes of the learner-centred feedback framework by [35] as shown in Table 3. It should be noted that we did not find an automatic tool to identify the feedback attribute **Attr.9 Invite dialogue about feedback** so we decided to annotate this attribute manually. However, we did not observe the **Attr.9** in our feedback dataset because the invitation of dialogue feedback (i.e., weekly one-on-one consultation) was commonly sent by the instructors in an oral form during the class or by the forum in the learning management systems. Thus, we only analysed eight attributes in our study.

3.3.1 Linguistic Politeness. Expressing politeness in textual language can show respect and care for the interlocutors' feelings [3], which are important to enhance the relationship between learners

[1]The marking rubric can be accessed by https://github.com/jionghaolin/LAK2023

Table 3: Mapping the artefact attributes from Ryan et al. [35] learner-centred feedback framework with the textual features extracted by software and domain experts. Agree, K, and Freq represent Agreement scores, Cohen's κ, and Frequency, respectively

Row ID	Artefact Attributes	Textual Features	# Features	Method	Agree	K	Freq (%)
Attr.1	**Strengthen teacher and learner relationships**	Politeness Strategies	39	`Politeness` (R Package)	N/A	N/A	N/A
		Relational Impact	4	LIWC	N/A	N/A	N/A
Attr.2	**Encourage learner agency**	Cognitive Impact	10				
Attr.3	**Encourage positive learner affect**	Affective Impact	5				
		Self Focus	1	Manual Annotation	0.99	0.98	55.88%
Attr.4	**Highlight strengthen of performance**	Task Focus (Positive Feedback)	1		0.98	0.96	58.46%
Attr.5	**Provide critiques about performance**	Task Focus (Negative Feedback)	1		0.96	0.92	29.04%
Attr.6	**Provide actionable information for future performance**	Task Focus (Non-corrective)	1		0.93	0.79	77.94%
		Process Focus	1		0.86	0.72	62.13%
		Feeding Forward	1		0.91	0.78	88.60%
Attr.7	**Promote learner independence**	Self-regulatory Focus	1		0.91	0.78	25.74%
Attr.8	**Usable for learners**	Feeding Up	1		0.98	0.95	69.49%
		Feeding Back	1		0.98	0.95	72.79%
		Writing Metrics	77	Coh-Metrix	N/A	N/A	N/A

and instructors [21, 40] and design effective feedback [11, 20, 32]. Therefore, we deemed that linguistic politeness can be mapped to the **Attr.1** [17] "*...development of healthy relationships between the teacher and the learner...*". To examine the politeness in feedback, our study employed the politeness tool[2] [43] (built on the Brown and Levinson [3] politeness theory) to extract 39 politeness strategies (e.g., `Hedges` strategy, "*Adding a diagram **might** be better*") from the feedback. The description of these politeness strategies was documented in an electronic appendix, which is accessible via https://github.com/jionghaolin/LAK2023. The extracted politeness strategies were counted on their frequency from the feedback.

3.3.2 Linguistic Inquiry and Word Count (LIWC). The LIWC dictionary has been widely used in content analysis of educational feedback to characterise written words into many psychological categories such as cognitive processes and emotions [7, 8, 13, 31], and these categories can reflect the writers' psychological states [33]. As suggested by Derham et al. [13], they selected 19 categories from the LIWC dictionary, which could manifest the impact of feedback (i.e., relational, cognitive, and affective impacts [17]) on students. By scrutinising the description of these impacts, we found these impacts could potentially be mapped to the *artefact* attributes as shown in Table 3. Firstly, the *relational impact* might influence students' engagement with the feedback (e.g., seeking feedback and making actions), which was related to the **Attr.1** [35]. Then, the *cognitive impact* might influence students' thinking about how to process information, memorise the learning information, and form concepts [13, 17], which might be related to the **Attr.2** [35] "*...and encouraging learners to engage in further independent study...*". Lastly, the *affective impact* might influence students' affective states (e.g., happy and stress), which is related to the **Attr.3** [35] "*...it may be beneficial for teachers to think about the potential affective impact...*". In line with the work of [13], our study also extracted 19 categories of features from the LIWC dictionary to represent relational (four categories), cognitive (ten categories), and affective (five LIWC categories) impacts. The examples of each category were shown in the electronic appendix, which is accessible via https://github.com/jionghaolin/LAK2023.

3.3.3 Feedback Model by Hattie and Timperley [16]. Inspired by the success of previous works [8, 31], it is promising to automate the process of identifying effective feedback practice (e.g., Hattie and Timperley four-level feedback focuses [16]). However, existing automatic tools for feedback analysis could not be applied to automate content analysis. Instead, we decided to manually annotate the feedback type and feedback focus for each instructor's feedback from our dataset and aimed to build a classifier in future work to automatically identify the components of feedback models. During the annotating process, we recruited two domain experts to annotate the entire feedback dataset by following the definition of feedback models introduced in [16]. The agreement score and Cohen's κ were shown in Table 3, which demonstrated promising annotation results. The inconsistent cases were further resolved by a third feedback expert. We also demonstrated the frequency of each feedback model feature in Table 3.

[2] https://cran.r-project.org/web/packages/politeness/

By scrutinising the description of the learner-centred feedback framework [35], we deemed that the feedback models could be mapped to the *artefact* attributes. In terms of the four feedback focuses, we divided task-level focus into three sub-categories, i.e., positive feedback, negative feedback, and non-corrective comment (i.e., instruction-related information) as suggested in [13]. We posited that the positive feedback at the task focus to be related to the **Attr.4** [35] "*...information which highlights what the learner has done well can be valuable...*", negative feedback at task focus to be related to the **Attr.5** [35] "*...comments that provide critiques of a learner's performance...*", and non-corrective comment at task focus to be related to the **Attr.6** "*...to improve on similar task; to achieve learning outcomes...*". Then, the feedback on the process focus leads to more direct information about the learning process underlying a task [16], which is potentially related to the **Attr.6** [35]. Feedback on self-regulatory focus aims to promote students' capabilities including self-monitoring, self-direction, and self-control to achieve learning goals, which is posited to be related to the **Attr.7** [35] "*...phrase feedback comments as suggestions or questions rather than statements...*". Feedback on self focus involved the positive evaluation and affective components in the feedback, which is posited to be related to the **Attr.3** [35] "*...to aim for positive emotional responses...*". In terms of feedback type, the feeding forward can inform students to determine the next steps in the subsequent tasks [16], which is assumed to be related to the **Attr.6** [35]. Then, the feeding up was used to clarify the goals and criteria of the assessment [14, 16], which was related to the **Attr.8** [35] "*..the information should be relevant to the assessment criteria...*". The feeding back informs the learners' progress towards the learning goals and responds to learners' work [14, 16], which is posited to be related to the **Attr.8** [35] "*...clearly related to particular aspects of the performance...*".

3.3.4 Coh-Metrix. To quantify the quality of feedback writing, we adopted the computational linguistic system Coh-Metrix, which was built by a set of metrics to calculate the complexity, cohesion, and readability of the written text [27]. The features extracted by Coh-Metrix have been widely used in the feedback content analysis to evaluate the quality of feedback writing [7, 8, 31]. By scrutinising the description of each metric in Coh-Metrix [27], we decided to select a subset of features from Coh-Metrix. According to the **Attr.8** [35], feedback should be easy to read, avoid the use of complex terms, and consider the detailed level of feedback. Firstly, to quantify the extent to which textual feedback can be easily read, previous work suggested that the traditional readability measures (e.g., Flesch Reading ease scores) and text cohesion-related metrics such as referential cohesion (i.e., measuring the writing cohesion level) in Coh-Metrix can be included [27]. Then, regarding the use of complex terms, as complex terms are rare in the students' reading experience [27], we decided to use the metric word information (i.e., built upon the corpus of commonly used words [27]) to measure the use of complex terms. Thirdly, to measure the detailed level of feedback, we selected several metrics such as descriptive measures (i.e., overview of the feedback characteristics, for example, number of words per feedback) and lexical diversity (i.e., the richness of words) based on the Coh-Metrix handbook [27]. In total, we have 77 features which were detailed in the electronic appendix https://github.com/jionghaolin/LAK2023.

3.3.5 Learners' Knowledge Level. Learners' knowledge level might relate to the learners' perception and sensemaking of the instructor feedback [17, 22]. For example, high-knowledge learners who obtain high performance on the assignment may engage more with feedback compared with low-knowledge learners [20, 39]. Therefore, our study decided to include the learners' knowledge level in the data analysis and treated the students' performance on Assignment I as a proxy of their knowledge level.

3.4 Data Analysis

3.4.1 Statistical Analysis. To answer our **RQ1**, we calculated the mean values and standard deviation for each feature in different student grade change groups (i.e., `Increase` and `Not Increase`). Then, we examined the use of each feature among the different student grade change groups by using statistical tests. Regarding the Feedback Model features (e.g., *Process* and *Feeding Up*), we counted their appearance in a binary form (i.e., exist or not exist) for each feedback. Thus, we adopted the Chi-square test to examine the association between the Feedback Model features and student grade changes and used Cramer's Phi to measure the effect size [38]. The other features were generated in numerical values, so we adopted Mann–Whitney U test and used Rank-Biserial (i.e., r_{rb}) to measure the effect size [38].

3.4.2 Interpretation on Predictive Analysis. To answer **RQ2**, we first investigated the capability of machine learning models on our task, i.e., use the *artefact* features extracted from feedback on assignment I to predict whether students obtained `Increase` or `Not Increase` on assignment II. As discussed in Section 2.2, previous works demonstrated the effectiveness of applying the machine learning models (e.g., Logistic Regression [30], Random Forest [8] and Gradient Tree Boosting [31]) to the prediction tasks (e.g., feedback quality prediction [8, 31] and student performance prediction [30]). Inspired by these works [8, 30, 31], the current study also used these models to predict whether or not the student can achieve improvement (i.e., `Increase` or `Not Increase`) on the subsequent task (i.e., assignment II). **Logistic Regression (LR)** model is one of the widely used statistical models in education studies, which can make binary classification on the task [30]. **Random Forest (RF)** model is a type of ensemble learning method which combines many predictors to predict the results [8]. Each predictor learns the patterns from a random sample of the data records from the training dataset. The final prediction of the RF model is made by averaging the predictions of each individual predictor. **Gradient Tree Boosting (GTB)** model is also an ensemble machine learning method that makes the prediction based on many predictors [9]. The difference between the GTB and RF models is in the way of combining the internal predictors. The predictors of the RF model make the prediction independently and vote on the final prediction results. There might be many errors in some predictions during the training process. In contrast, the predictions of the GTB model are built in a sequential manner, and the prediction errors from the previous predictors can be fixed by the subsequent predictors.

Then, we adopted the well-established interpretable framework **SHAP** (**SH**apley **A**dditive ex**P**lanations) [24] to understand the contribution of features on model prediction performance. The SHAP framework was developed based on the game theory to quantify

Table 4: The comparison of selected features between the `Increase` group and the `Not Increase` group. Features marked with † were examined by Chi-square test and Cramer's Phi effect size whereas the remaining features were examined by Mann-Whitney U test and Rank-Biserial (i.e., r_{rb}) effect size

Row ID	Artefact Attributes	Feature Clusters	Features	Increase		Not Increase		Difference	
				M	SD	M	SD	P-val	E.S
		Knowledge Level	*1st Assgmt grades*	1.5	1.03	3.16	1.10	***	0.72
Attr.1	**Strengthen teacher and learner relationships**	Politeness [43]	*Negative.Emotion*	0.74	0.79	0.36	0.64	***	-0.27
			Impersonal.Pronoun	1.98	2.18	1.40	2.28	**	-0.23
			Hedges	0.97	1.55	0.51	1.13	**	-0.19
			Give.Agency	0.24	0.50	0.08	0.26	**	-0.14
		Relational Impact (LIWC)	*affiliation*	0.24	0.51	0.13	0.45	*	-0.10
Attr.2	**Encourage learner agency**	Cognitive Impact (LIWC)	*interrog*	1.69	2.11	0.98	1.95	***	-0.31
			cause	1.45	1.71	1.33	2.64	*	-0.16
			QMark	0.58	1.05	0.37	1.02	*	-0.13
Attr.3	**Encourage positive learner affect**	Affective Impact (LIWC)	*posemo*	5.76	2.31	10.89	12.17	***	0.48
Attr.4	**Highlight strengths of performance**	Feedback Model [16]	† *Task (Pos)*	0.61	0.49	0.56	0.50	N.S	0.04
Attr.5	**Provide critiques about performance**		† *Task (Neg)*	0.44	0.50	0.25	0.43	**	0.19
Attr.6	**Provide actionable information**		† *Process*	0.85	0.36	0.60	0.49	***	0.25
Attr.7	**Promote learner independence**		† *Self-regulate*	0.38	0.49	0.24	0.43	*	0.14
			† *Feeding Up*	0.91	0.29	0.62	0.49	***	0.29
Attr.8	**Usable for learners**	Writing Metrics (Coh-Metrix)	*DESWC*	73.03	22.55	52.04	35.43	***	-0.48
			WRDADJ	24.86	9.41	16.99	12.50	***	-0.42
			DRAP	19.92	6.51	14.70	10.05	***	-0.41
			DESWLsyd	0.93	0.20	0.76	0.27	***	-0.41

Note: **M** = mean; **SD** = standard deviation; **P-val** =. p-values; **E.S** = effect size; *1st Assgmt grades* = student grades on Assignment I; *posemo* = positive emotional words; *interrog* = interrogatives words; *cause* = causal language; *QMark* = question mark; *Task (Neg)* = negative feedback at task focus; *Task (Pos)* = Positive feedback at task focus; *DESWC* = number of words; *WRDADJ* = adjective incidence; *DRAP* = adverbial phrase incidence; *DESWLsyd* = standard deviation of the word length; In the columns of **P-val**, *** p<0.001; ** p<0.01; * p<0.05; N.S = not significant;

the features' contribution to the model's prediction [24]. According to [23, 24, 34], the SHAP framework shows the feature contribution not only at the global level (i.e., which features contributed most to the model prediction performance) but also at the local level (i.e., which features presented the positive and negative impact on individual prediction). It should be noted that the interpretability from SHAP does not indicate causality.

3.4.3 Study Setup. To evaluate the model prediction performance, we randomly split the dataset into *training set*, *validation set*, and *testing set* with the ratio *70%:10%:20%*, respectively. The models were trained using the Python package `scikit-learn`[3] and the models' hyper-parameters (e.g., *n_estimators*) were further tuned by applying grid search on the *validation set*. Finally, all the models' performances were evaluated on the *testing set* by using four representative metrics, i.e., F1-score, Area Under the Curve (AUC), Cohen's κ, and classification accuracy.

4 RESULTS

4.1 Results on RQ1

We reported the statistical results based on the *artefact* attributes (shown in Table 4) and the results within each attribute were sorted based on their effect size[4]. We found that the values of feature *1st Assgmt grades* (i.e., student grades on Assignment I) in `Increase` group were significantly lower than the `Not Increase` group. Additionally, the effect size of the feature *1st Assgmt grades* presented a strong association [10] with the student grade changes.

[3]https://scikit-learn.org/

[4]Due to the space limit, we presented the five most significant results for each *artefact* attribute.

For the **Attr.1** in Table 4, we investigated the linguistic politeness features and relational impact where the use of *Negative.Emotion*, *Impersonal.Pronoun*, *Hedges*, *Give.agency*, and *affiliation* were more frequent in the `Increase` group than they were in the `Not Increase`. It should be noted that the *affiliation* feature is related to the sentences fostering a sense of encouragement (e.g., "*It can **help** the analysis*"), *Negative.Emotion* is related to the use of negative emotional words (e.g., "*it is **difficult** to explain*"), *Impersonal.Pronoun* is related to the non-person referents (e.g., "***That** is not a data science work*"), *Hedges* is related to the indirect voice tone (e.g., "*The focus **might** need to be narrowed down*"), and *Give.Agency* is related to the sentences fostering a sense of suggestion (e.g., "*It would be great if you can find a dummy dataset*"). These significant results might indicate that though the comments were crafted with more negative emotional words in the `Increase` than `Not Increase` group, instructors also made more effort to maintain the relationship in `Increase` group by using expressions (i.e., *affiliation*, *Impersonal.Pronoun*, *Hedges*, and *Give.agency*).

For the **Attr.2** in Table 4, we investigated the cognitive impact features where the use of *interrog*, *cause* and *QMark* features in the `Increase` group were more frequent than those in the `Not Increase`. The use of *QMark* (e.g., "*?*") and *interrog* (e.g., "***What** is the role of data scientist.*") were related to the questions in feedback and the significant results indicated that instructors might have posed more questions in their feedback to the `Increase` group than they did in feedback for the `Not Increase` group. Then, the *cause* feature could demonstrate comments expressed in the sense of causation (e.g., "***because** you need to address these in next assignment*") and the results indicated that the students might have received more comments contained *cause* in the `Increase` group than they did in the `Not Increase`.

For **Attr.3** in Table 4, we investigated the affective impact features where the use of *posemo* feature in the `Not Increase` group was more frequent than those in the `Increase`. The *posemo* feature is related to the adjective expressing positive emotion, which could capture most praise (e.g., "***Good** job!*", "***Excellent!***") in feedback. Thus, the significant result might be the reason that the `Not Increase` students received more praise from instructors than the `Increase` since the students in the `Not Increase` had higher grades than those in the `Increase` on Assignment I.

In Table 4, four *artefact* attributes were captured by the Feedback Model. For the **Attr.4**, the *Task (Pos)* (i.e., positive feedback in task level focus) identified the feedback highlighted strengths of student performance (e.g., "*Good topic with a clear discussion of project goals*"). The significant difference was not found between the `Increase` and `Not Increase` groups. For the **Attr.5**, the *Task (Neg)* feature (i.e., negative feedback in task level focus) identified the feedback contained the instructor's criticism on students' submission (e.g., "*You did not introduce the role of data science*"). The significant results indicated that the use of *Task (Neg)* features was significantly correlated with the student grade change. For the **Attr.6**, the *Process* feature (i.e., the appearance of process level focus) identified the feedback with concrete instruction to improve on the subsequent task and achieve the expected goals (e.g., "*It would be great if you add the flow chart presenting the overall structure of the project*"). The significant differences indicated that the use of the *Process* features was significantly correlated with the student grade change. For the **Attr.7**, the *Self-regulate* (i.e., self-regulatory level focus) identified the feedback with the questions or statements to encourage learners to think more about improvement on their subsequent assignment. (e.g., "*What are the data science roles?*"). The significant differences indicated that the use of *Self-regulate* features was significantly correlated with the student grade change.

We investigated the **Attr.8** by using the feature *Feeding Up* and features related to Writing Metrics. Compared with the `Not Increase` group, `Increase` group had higher feature values of *Feeding Up*, *DESWC* (i.e., number of words per feedback, which indicated the informativeness level of the feedback [27]), *WRDADJ* (i.e., adjective incidence, which indicated the density level of using adjectives [27], "***Good** topic with **detailed** overview*"), *DRAP* (i.e., the adverbial phrase incidence, which indicated the density level of using adverbial phrases [27], "*Describe the data aspect **in the goal***"), and *DESWLsyd* (i.e., standard deviation of the word length, which indicated the level of text variation in terms of the word length [27]). The *Feeding Up* feature identifies feedback containing marking criteria or expected goals (e.g., "*You should clearly describe the business benefits in your report*" where the clear description of the business benefits is in the marking criteria). The significant differences indicated that the use of *Feeding Up* features was significantly correlated with the student grade change. Regarding the writing metrics, the significant results of the features *DESWC*, *WRDADJ*, *DRAP*, and *DESWLsyd* indicated that the students in `Increase` group might receive the feedback with more descriptive information compared with `Not Increase`.

4.2 Results on RQ2

Table 5: The performance of LR, RF, and GTB in predicting student grade changes on assignment II.

Model	Accuracy	F1-score	AUC	Cohen's κ
Logistic Regression (**LR**)	0.81	0.78	0.77	0.56
Random Forest (**RF**)	0.80	0.76	0.75	0.51
Gradient Tree Boosting (**GTB**)	**0.84**	**0.80**	**0.80**	**0.61**

The model evaluation results were shown in Table 5. We found that the performance of **GTB** model outperformed the other two models (i.e., **LR** and **RF**) across four metrics. Therefore, we decided to use the **GTB** model to analyse the contribution of *artefact* features to the model's prediction. By applying the SHAP framework to the trained GTB model, we presented the feature importance at the global level (shown in the bar chart Fig 1 (a)) and local level (shown in the beeswarm plot Fig 1 (b)). The features were sorted from the most significant to the less significant ones in both Fig 1. Due to the space limit, we only presented the top 10 most significant features. In Fig 1 (a), we observed that the *1st Assgmt grades* feature was considered the most significant feature by **GTB** model. Then, three significant features were extracted from the *LIWC* dictionary (i.e, *interrog*, *posemo*, and *ppron*) and six features calculated by *CohMextrix* (i.e., *DESWLsyd* – standard deviation of the word length; *DESSL* – average length of sentences; *WRDCNCc* – concreteness level of the text; *SMCAUSlsa* – measures of verb overlap calculated by Latent Semantic Analysis (LSA); *LSAGN* – LSA-based cohesion

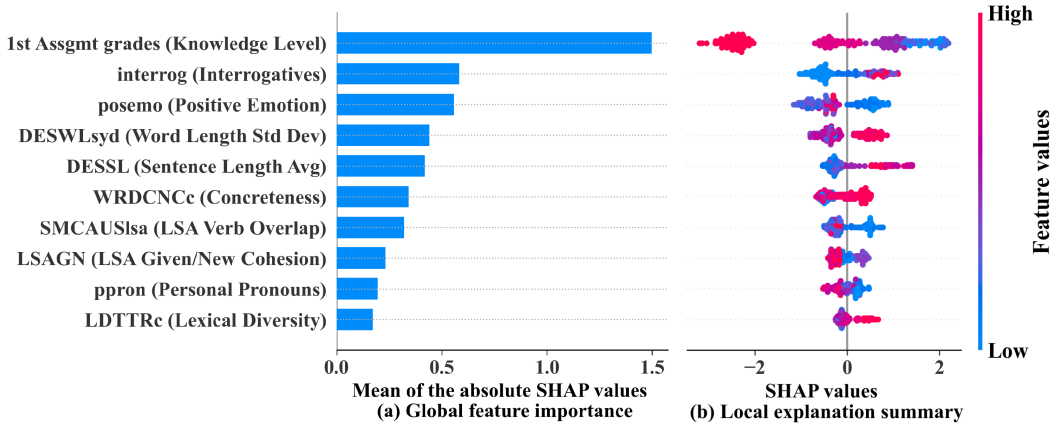


Figure 1: Top 10 most significant features on the GTB model' prediction. Fig 1 (a) and (b) shared the same rank of feature importance. In Fig 1 (a), the bar chart ranked the features from the most important to the less important. The feature importance was measured by the mean of absolute SHAP values. In Fig 1 (b), the beeswarm plot demonstrated the distribution and direction of the feature effect at the instance level. Each line has an equal number of scatters, and the number of scatters in a line is the same as the number of instances in our training dataset. The position of each scatter in a line was determined by the SHAP value. The positive SHAP values are indicative of the `Increase` group, while negative are indicative of the `Not Increase` group. The colour of the points represents the feature values from low to high.

measure, which reflects the ratio of given and new information in the text; *LDTTRc* – lexical diversity ratio, measured by the number of unique words divided by the total words in a feedback comment.)

We further investigated the interpretation of feature importance at the local level, which presents the distribution of instances and direction of the feature effect. For example, in Fig 1 (b), the feature *1st Assgmt grades* presents a cluster of red scatters on the left side of the central axis, which indicates many instances with the high grades on assignment I were predicted as `Not Increase`. In contrast, the instances with low grades, concentrated on the right, tended to be predicted as `Increase`. The observed patterns of feature *1st Assgmt grades* demonstrated that the student grade on assignment I was a strong indicator to distinguish between `Increase` and `Not Increase`, which was in line with the result in Table 4. Similarly, we found that the high feature values of *interrog* concentrated on the right side, which indicates a positive correlation between the high feature values of *interrog* and the prediction of `Increase`. In contrast, most instances with low values of *interrog* concentrated on the left side, which indicates that the low values of *interrog* positively correlated with `Not Increase`. Then, we found that the high feature values of *posemo* concentrated on the left side. This result indicates a positive correlation with the prediction of `Not Increase`. The high value of *posemo* could be a strong predictor for the GTB model to predict `Not Increase`. However, the instances with low values of the *posemo* were dispersed from the positive SHAP values to the negative. It indicates *posemo* feature values may not have a clear correlation with either of the two groups – `Increase` or `Not Increase`. Additionally, we also observed that the low feature values of *DESWLsyd*, *WRDCNCc*, and *LDTTRc* converged towards the left, but high values of these features were dispersed. The results indicate that the low feature values of *DESWLsyd*, *WRDCNCc*, and *LDTTRc* positively correlated with the `Not Increase`.

5 DISCUSSION

5.1 Implications

Students with a lower grade on the first assignment were more likely to obtain a grade increase on the subsequent assignment. Unsurprisingly, the students with lower grades had more room to improve on the subsequent task compared to the students who had already obtained higher grades. Then, we observed many differences of *artefact* attributes between the group of `Increase` and `Not Increase`, which might shed light on the design and evaluation of the feedback based on the learner-centred framework.

Critiquing students' work as part of feedback is unavoidable, and crafting feedback comments in polite language for low knowledge students is recommended. Existing research also found that expressing feedback comments politely could help build rapport with students [28, 40] and also potentially enhance student performance [28], especially for the low-knowledge students who were more engaged with the polite comments [26]. Compared to the students in `Not Increase`, we observed that students in `Increase` group with lower grades on assignment I received more negative emotional words (e.g., "*Your reference style is* ***incorrect***") and also more polite elements such as encouraging student engagement and rewording the comments into suggestions (e.g., "*It would be great if you add a flow chart*") instead of commands. Although more negative emotional words were used in the feedback, more polite elements were detected as well, which showed instructors trying to direct students to identify where to improve and meanwhile also gave the care to maintain the student and instructor relationship.

Positively framed feedback content might be a good indicator to distinguish between the student grade changes. Compared with the group of `Not Increase`, students in the `Increase` received fewer positive emotional words, which were related to comments encouraging positive affect and highlighting learners' strengths.

This might be the reason that students in the Increase group obtained lower grades on the first assignment than those in the Not Increase so the students in the Increase were less likely to receive the praise (e.g., "***Excellent** Work!*") from instructors [25]. Prior works also suggested that using praise for low knowledge students needs to be carefully delivered because the praise might undermine students' own responsibility to identify their mistakes or errors [25]. Instead, using feedback to highlight the specific aspect of student strengths (e.g., "*Your report is well structured*") is recommended as this positive feedback can affirm students' understanding of success criteria, enhance student self-efficacy, maintain the student and teacher relationship, and mitigate the potential harshness of criticism.

To assist students' improvements on the subsequent task, it is important to demonstrate comments in questions or suggestions. We noted that the questions in feedback mostly correlated with the use of the self-regulatory focus (e.g., "*What about other challenges?*") and suggestions demonstrated the process focus (e.g., "*You should make sure that you have access to details about the datasets used here*). The students in the Increase group received the feedback with more use of self-regulatory and process focus in feedback compared to those in the Not Increase. The existing literature found that the use of self-regulatory and process focus in feedback had a positive impact on promoting learner independence and grade improvement [15]. However, the self-regulatory focus was rarely provided in the feedback compared to the use of process focus (shown in Table 3). Thus, we recommend that instructors should make more effort in rephrasing statements into questions in the feedback.

Usable feedback comments should be both clear and detailed. Compared with the students in the Not Increase group, the students in the Increase group received more comments on reminding the success criteria of the assignment. The reason might be that the students in Increase group obtained lower grades than Not Increase on assignment I, which indicates the learning gap between the current and the expected grades was larger in the Increase group than the Not Increase group. Thus, instructors reminded more frequently of the success criteria for the low performance students to help them minimise the learning gap and make improvements on the subsequent assignment. Moreover, instructors provided more detailed feedback for the students in the Increase group. The reason might also be in relation to the students' grades on the first assignments. As discussed before, low performance students received more comments on critiques, instructional questions and suggestions than the high performance students did, which led to more detailed feedback for the low performance students. Therefore, we recommend that instructors might consider reminding the low performance students of the successful goals and making an effort to elaborate detailed comments.

Lastly, our results demonstrated the effectiveness of the GTB model, which was in line with the results in [31]. To compensate for the GTB model's low interpretability, our study employed the SHAP framework to interpret how the GTB model learned the patterns from the distribution of engineered features to predict student grade changes. Our results showed the potential of using the SHAP framework to deliver trustworthy analytic results. At the global level of interpretation, the SHAP framework identified some significant features corroborated what we observed when answering RQ1 such as *1st Assgmt grades* (student grade on the first assignment) and *posemo* (positive emotional words) showing significant differences between Increase and Not Increase. Then, the SHAP framework could present the direction (i.e., positive or negative) of the correlation between the significant feature values and the student grade changes at the local level, which provided transparent interpretations for the GTB model prediction. Since the features were built upon the feedback literature [35], we expected that the interpretations might help instructors observe the influence of using different feedback comments and further design their feedback on the basis of supporting students. As suggested in [34], both global and local level interpretation can help enhance the users' trust since the human users are inclined to trust the predicted results if they can understand why the model makes the prediction. However, we believe that future studies with instructors are needed to validate this claim. It is necessary to understand which exactly of the features are valued by the instructors and the extent to which the use of SHAP increases their trust.

5.2 Limitations and Future Work

We acknowledge that there are some limitations in the current study. *First*, our study was based on the feedback dataset from one course, which might not represent feedback in other courses. *Secondly*, some feedback content might be repeated since the same instructors might have given the same feedback to the students who encountered similar issues in their reports. *Thirdly*, we noted that the student's grade changes on the subsequent task might not fully demonstrate the feedback impact since many factors (e.g., peer discussions) might occur in the loop of feedback provision, which could in turn confound associations of feedback with the performance changes [17]. Future research should investigate the correlation between feedback content and other related factors (e.g., student learning strategies). *Lastly*, though the **GTB** model effectively predicted students' grade changes in our study, many falsely predicted cases still existed. The reasons for mistakenly predicted cases could be multi-faceted. For example, prediction accuracy could be improved by using more sophisticated machine learning models. Given the success of the deep neural network models in processing different tasks, it is worthwhile to investigate the value of applying deep neural networks to our task when we collect sufficient datasets for training such models. Another reason for falsely predicted cases might be related to the students' engagement with feedback. Since students might not always read feedback content, feedback might not be effective if students do not engage with it [4, 36]. Therefore, it is worthwhile to further incorporate students' engagement activities (e.g., students' posts in the forum [37]) before and after the feedback provision into the model training process.

Furthermore, an extension of our study would be to design a system that can automatically evaluate instructors' feedback based on the attributes from the learner-centred feedback framework and provide interpretable recommendations to the instructors about how feedback can be improved. For example, when the provided feedback lacks positive highlights of strength and contains many negative comments, the system can make a suggestion for the instructors to include positive emotional words to encourage students.

6 CONCLUSION

This paper illustrated how the learner-centred feedback framework could be used for analysing the feedback content by using the learning analytics approaches. We also demonstrated the potential of using machine learning models to predict student achievements on subsequent assignments and using the well-established framework SHAP to provide transparent interpretations of the predicted results. The implementations of the learner-centred analytics of feedback content in our study have important implications for the future design of automated feedback systems and the practice of feedback provision.

ACKNOWLEDGMENTS

This research was supported by the Australian Government through the Australian Research Council (DP220101209).

REFERENCES

[1] Rola Ajjawi, David Boud, Michael Henderson, and Elizabeth Molloy. 2019. Improving feedback research in naturalistic settings. In *The Impact of Feedback in Higher Education*. Springer, 245–265.
[2] David Boud and Elizabeth Molloy. 2013. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education* 38, 6 (2013), 698–712.
[3] Penelope Brown and Stephen Levinson. 1987. Politeness: some universals in language usage. Cambridge University Press, Cambridge, UK.
[4] David Carless. 2006. Differing perceptions in the feedback process. *Studies in higher education* 31, 2 (2006), 219–233.
[5] David Carless. 2019. Feedback loops and the longer-term: towards feedback spirals. *Assessment & Evaluation in Higher Education* 44, 5 (2019), 705–714.
[6] David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education* 43, 8 (2018), 1315–1325.
[7] Anderson Pinheiro Cavalcanti, Rafael Ferreira Leite de Mello, Vitor Rolim, Máverick André, Fred Freitas, and Dragan Gaševic. 2019. An analysis of the use of good feedback practices in online learning courses. In *2019 IEEE 19th ICALT*, Vol. 2161. IEEE, 153–157.
[8] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. How Good is My Feedback? A Content Analysis of Written Feedback. In *Proceedings of the LAK (LAK '20)*. ACM, New York, NY, USA, 428–437.
[9] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 2016 Conference on KDD*. 785–794.
[10] Jacob Cohen. 2016. A power primer. (2016).
[11] Phillip Dawson, Michael Henderson, Paige Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. 2019. What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education* 44, 1 (2019), 25–36.
[12] Phillip Dawson, Michael Henderson, Tracii Ryan, Paige Mahoney, David Boud, Michael Phillips, and Elizabeth Molloy. 2018. Technology and feedback design. *Learning, design, and technology* (2018).
[13] Cathrine Derham, Kieran Balloo, and Naomi Winstone. 2021. The focus, function and framing of feedback information: linguistic and content analysis of in-text feedback comments. *Assessment & Evaluation in Higher Education* (2021), 1–14.
[14] Douglas Fisher and Nancy Frey. 2009. Feed up, Back, Forward. *Educational Leadership* 67, 3 (2009), 20–25.
[15] John Hattie. 2012. *Visible learning for teachers: Maximizing impact on learning*. Routledge.
[16] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.
[17] Michael Henderson, Rola Ajjawi, David Boud, and Elizabeth Molloy. 2019. Identifying feedback that has impact. In *The impact of feedback in higher education*. Springer, 15–34.
[18] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. 2022. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence* (2022).
[19] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the 6th LAK*. 15–24.
[20] Lisa-Angelique Lim, Shane Dawson, Dragan Gašević, Srecko Joksimović, Abelardo Pardo, Anthea Fudge, and Sheridan Gentili. 2021. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. *Assessment & Evaluation in Higher Education* 46, 3 (2021), 339–359.
[21] Jionghao Lin, Mladen Rakovic, David Lang, Dragan Gasevic, and Guanliang Chen. 2022. Exploring the Politeness of Instructional Strategies from Human-Human Online Tutoring Dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 282–293.
[22] Anastasiya A Lipnevich, David AG Berg, and Jeffrey K Smith. 2016. Toward a model of student response to feedback. In *Handbook of human and social conditions in assessment*. Routledge, 169–185.
[23] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.
[24] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Proceedings of the 31st NeurIPS* 30 (2017).
[25] Effie Maclellan. 2005. Academic achievement: The role of praise in motivating students. *Active learning in higher education* 6, 3 (2005), 194–206.
[26] Bruce M McLaren, Krista E DeLeeuw, and Richard E Mayer. 2011. A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies* 69, 1-2 (2011), 70–79.
[27] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
[28] Maria Mikheeva, Sascha Schneider, Maik Beege, and Günter Daniel Rey. 2019. Boundary conditions of the politeness effect in online mathematical learning. *Computers in Human Behavior* 92 (2019), 419–427.
[29] David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education* 31, 2 (2006), 199–218.
[30] Serena Nicoll, Kerrie Douglas, and Christopher Brinton. 2022. Giving Feedback on Feedback: An Assessment of Grader Feedback Construction on Student Performance. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. 239–249.
[31] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gašević, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. 2022. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence* (2022).
[32] Berry M O'Donovan, Birgit den Outer, Margaret Price, and Andy Lloyd. 2021. What makes good feedback good? *Studies in Higher Education* 46, 2 (2021), 318–329.
[33] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
[34] Filipe Dwan Pereira, Samuel C Fonseca, Elaine HT Oliveira, Alexandra I Cristea, Henrik Bellhäuser, Luiz Rodrigues, David BF Oliveira, Seiji Isotani, and Leandro SG Carvalho. 2021. Explaining Individual and Collective Programming Students' Behavior by Interpreting a Black-Box Predictive Model. *IEEE Access* 9 (2021), 117097–117119.
[35] Tracii Ryan, Michael Henderson, Kris Ryan, and Gregor Kennedy. 2021. Designing learner-centred text-based feedback: a rapid review and qualitative synthesis. *Assessment & Evaluation in Higher Education* 46, 6 (2021), 894–912.
[36] Shirley V Scott. 2014. Practising what we preach: towards a student-centred definition of feedback. *Teaching in Higher Education* 19, 1 (2014), 49–57.
[37] Lele Sha, Mladen Raković, Jionghao Lin, Quanlong Guan, Alexander Whitelock-Wainwright, Dragan Gašević, and Guanliang Chen. 2022. Is the Latest the Greatest? A Comparative Study of Automatic Approaches for Classifying Educational Forum Posts. *IEEE TLT* (2022), 1–14.
[38] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences* 1, 21 (2014), 19–25.
[39] Yi-Shan Tsai, Rafael Ferreira Mello, Jelena Jovanović, and Dragan Gašević. 2021. Student appreciation of data-driven feedback: A pilot study on OnTask. In *LAK21: 11th international learning analytics and knowledge conference*. 511–517.
[40] Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International journal of human-computer studies* 66, 2 (2008), 98–112.
[41] Naomi Winstone, David Boud, Phillip Dawson, and Marion Heron. 2022. From feedback-as-information to feedback-as-process: a linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education* 47, 2 (2022), 213–230.
[42] Min Yang and David Carless. 2013. The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education* 18, 3 (2013), 285–297.
[43] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness Package: Detecting Politeness in Natural Language. *R Journal* 10, 2 (2018).

6.3 Chapter Summary

The evaluation of effective communicative patterns in assessment feedback is an important research topic in the learning analytics literature. Despite prior works revealing insights on the construct of effective assessment feedback [32–35], it is still necessary to investigate other communicative patterns for crafting effective feedback (Gap 3). Recent work by Ryan et al. [8] proposed a learner-centred feedback framework that summarised the effective communicative patterns in assessment feedback based on the prior literature on feedback. However, the communicative patterns from a learner-centred feedback framework to evaluate feedback is underexplored (RQ 5).

In the study reported in the current chapter, we focused on the learner-centred feedback framework [8] and used well-established data analytic methods to investigate the communicative patterns in the assessment feedback. The contributions of our research to the feedback literature are two-fold. First, we provided a set of recommendations for feedback practice (**Contribution 9**). Second, we evaluated the predictive power of extracted features in predicting learner task performance on the subsequent assignment. Beyond the prediction, we employed the interpretable framework SHAP to present the importance of features in the predictions. Our results suggested that the utilisation of the GTB model and SHAP framework would shed light on the design of automated feedback evaluation systems. (as described in **Contribution 10**).

Contribution 9: Feedback practice of communicative patterns in assessment feedback.

1. Critiquing learners' work as part of feedback is unavoidable, and crafting feedback comments in polite language for low-performing learners is recommended.

2. High-performing learners often received more positively framed feedback content (e.g., positive emotional words related to comments encouraging positive affect) than low-performing learners. However, the comments involved highlighting the strengths of performance (e.g., "*Your report is well structured*") were almost equally distributed among the low- and high-performing learners. Thus, crafting feedback to highlight the learner's strengths about their working efforts is recommended since the positive feedback can affirm learners' understanding of success criteria, enhance learners' self-efficacy, maintain the learner-educator relationship, and mitigate the potential harshness of criticism [8].

3. Compared with the high-performing learners, low-performing learners received more comments on critiques, instructional questions and suggestions than the high-performance students did, which led to more detailed feedback for the low-performing learners. Therefore, we recommend that the feedback should be crafted to present detailed content such as a description about the successful goals, suggestions about the improvement on the

subsequent assignment, and questions about learners' understanding of the assignment requirement.

Contribution 10: Automatic evaluation of communicative patterns in assessment feedback.

1. Training on the engineered features (learner performance on Assignment I and features extracted from assessment feedback based on the learner-centred framework), the GTB model can achieve promising reliability in predicting learner performance on Assignment II. The trained GTB model can be used to estimate the extent to which the crafted feedback content will help learners improve their grades on the subsequent assignment.

2. The SHAP framework has the potential to deliver trustworthy analytic results. By applying the SHAP framework to the trained GTB model, feedback researchers can identify the rank of significant features which contributed to the prediction of the GTB model. Additionally, the SHAP framework can present the direction (i.e., positive or negative) of the correlation between the use of features and the learners' grade changes (i.e., increase or not) on the subsequent assignment. Thus, we deem that the SHAP framework can help researchers or educators understand how the feedback can be further improved to support learners.

Chapter 7

Conclusion

The importance of educational feedback has been widely acknowledged by educators and researchers. The current PhD thesis contributed to the feedback research based on the *Feedback Triangle* framework [1]. This feedback framework characterises feedback from the *structural dimension*, *cognitive dimension*, and *social-affective dimension*. The current PhD thesis investigated the communicative patterns from the *cognitive dimension* and *social-affective dimension*. By scrutinising the feedback literature, we identified three main research gaps as described in Section 1.3.

First, we argued that the provision of one-on-one online tutoring dialogue can be considered a form of dialogic feedback. To provide tutoring dialogue for large learner cohorts, researchers have developed dialogue-based Intelligent Tutoring Systems (ITSs) to address the high demand for human expert educators [54–57]. Despite many successful implementations of dialogue-based ITSs for providing tutoring dialogues, a recent systematic literature review [59] reports that the communicative patterns of existing ITSs still need to be improved to become comparably effective as human educators. By scrutinising the literature, we identified that one of the major improvements proposed in prior research is to communicate politely with a learner [56, 60]. However, the benefits and hindrances of guiding learners politely coexist in the tutoring dialogues [37] and it still remains unclear about the extent to which dialogue-based ITSs and human educators should express politeness to maintain the effectiveness of one-on-one online tutoring dialogues. Thus, further research efforts should be directed towards investigating the communicative patterns in tutoring dialogues to bridge this gap (Gap 1). *Secondly*, a fundamental element of tutoring dialogue is how educators communicate with learners [44, 70, 112]. Thus, it is important to identify what educators and learners do, which can be further used to understand effective communicative patterns in the feedback provision process. A widely-used approach to identify what educators and learners do is to encode the educator and learner's conversational sentences/utterances into dialogue acts [44, 70, 112]. The dialogue acts can

represent the intention in the communication process of dialogue (e.g., "*What will you do?*" can be encoded as an `Open Question` dialogue act). Though many prior studies employed machine learning models to automatically identify the dialogue acts from tutoring dialogues [90, 116, 117], we argued that the classification performance of the dialogue act classifier still needs to be improved to become comparably accurate as that of human annotation. Thus, further efforts are required for enhancing the identification of dialogue acts in tutoring dialogues (Gap 2). *Thirdly*, many prior studies also employed learning analytics approaches to analyse the communicative patterns (e.g., linguistic features analysis [32–35] and sentiment analysis [32, 33, 35]) from the assessment feedback. As discussed, the conceptualisation of feedback has gradually moved from feedback as information to feedback as a process [11]. We noted that limited studies investigated communicative patterns based on the theoretical framework of feedback as a process. Due to the importance of understanding effective feedback, more efforts should be devoted to the investigation of effective communicative patterns in assessment feedback based on the theoretical framework of feedback as a process (Gap 3). In the current PhD thesis, we aimed to fill the research gaps by using the learning analytic approaches to analyse the communicative patterns (from the *cognitive* and *social-affective* dimensions) in both tutoring dialogue and assessment feedback. In addition, we also explored the solutions for the improvements of identifying dialogue acts from educators and learners in tutoring dialogue.

In this chapter, we summarise the main findings and implications of the research conducted in this PhD thesis with respect to the research questions outlined in Section 1.3. Additionally, we provide an outline of future research directions.

7.1 Implications for Research and Practice

7.1.1 RQ 1: Displaying politeness in tutoring dialogues

We started the investigation of communicative patterns from the *social-affective dimension*. Particularly, we focused on the use of linguistic politeness [81] in tutoring dialogues. Showing politeness in communication is important to maintain the relationship between the interlocutors [81]. In Chapter 2, we presented the investigation of linguistic politeness in tutoring dialogues by using learning analytics approaches. Our study revealed the role of politeness by taking the proxy of learners' prior knowledge (i.e., learners' prior progress on a problem-based learning task) into account. Our work has brought several implications as discussed below:

Educational practice of using politeness in tutoring dialogues. We demonstrated the extent to which educators and learners display politeness in the tutoring dialogues. From a theoretical view, our findings in Chapter 2 contributed to the feedback literature in terms of the *social-affective dimension* [1]. The *social-affective dimension* suggested that educators should build

rapport in the feedback communication process, which could enhance learners' trust in educators' feedback to enhance their performance [1]. Since politeness is the human behaviour that aims to strengthen the relationship between the interlocutors [81], our study contributed to the *social-affective dimension* by investigating the correlation between educator politeness and learner problem-solving performance. From a practical viewpoint, our research results may inform educational practices for the use of politeness in the communication process of feedback, which aims to support the construction of feedback on the *social-affective dimension*. For example, in the effective tutoring dialogue sessions, the educators tended to use more the words "*We*" (a politeness strategy, "*We can work on this question together*") to guide the learners with prior progress than those without. The learners without prior progress might present low-level self-efficacy and the perception of an imbalanced power relationship between educators and learners might further lower learners' self-efficacy. As a remedy, educators can use the word "*we*" to indicate a sense of shared responsibility towards the joint goal, which might show certain emotional caring to the learners without prior progress [60]. The use of this polite strategy can mitigate the learners' perception of power from educators [118], which might be helpful to improve the learning experience in the tutoring process. We also suggest that educators might consider using positive emotional words ("*Good work!*") to learners with prior progress, which might encourage the learners to make a further effort. The polite strategy of expressing indirect requests (e.g., "*Should you calculate the value of X first?*") could help educators mitigate the sense of directness (e.g., "*You should calculate the value of X first*"). However, this polite strategy should be carefully used in delivering instructional hints, especially for learners without prior progress since the misuse of expressing indirect requests might make the instruction unclear [119], which might further hinder learners' understanding. Instead, we suggest educators consider using more direct expressions to explicitly guide learners without prior progress.

Insights of using politeness on the design of automated feedback system. Our study demonstrated politeness has the potential to predict learner problem-solving performance. To achieve better prediction performance, the politeness features should be incorporated with other features (e.g., sentiment level). In practice, the earlier a learner's problem-solving performance can be predicted, the more tailored support the feedback system can provide [120–122]. As politeness has the potential to enhance the model prediction performance and support learners' problem-solving performance, we recommend that researchers and software developers add politeness features to the prediction process. Additionally, our predictive model can be further embedded in the design of the feedback evaluation systems which are used to train novice educators. Notably, many educational platforms (e.g., PLUS [123]) employ computer-assisted technologies to support novice educators to provide effective feedback for improving learning achievement. As our predictive model could learn the correlation between the politeness features and learner problem-solving performance, we expect that the model can evaluate the educators' politeness and help novice educators craft feedback regarding politeness.

7.1.2 RQ 2: Identifying effective instructional strategies from tutoring dialogues

In Chapter 2, we demonstrated the use of politeness in tutoring dialogues from the perspective of *social-affective dimension*. We, then, aimed to investigate the communicative patterns of tutoring dialogues from the *cognitive dimension*. Motivated by the studies by Rus and colleagues [44, 70], our study explored the dialogue acts of educators and learners in tutoring dialogues, which can be used to inform how expert educators deliver effective instructional strategies to learners [44, 70]. In Chapter 3, we presented the commonly used dialogue acts taken by educators and learners in tutoring dialogues. Our study examined effective instructional strategies with the consideration of learners' prior knowledge (i.e., learners' prior progress on a problem-based learning task). The key implications of the results of our research are as follows.

Educational practice of instructional strategies in tutoring dialogues. We demonstrated how educators and learners communicate in tutoring dialogues. From a theoretical view, our findings in Chapter 3 contributed to the feedback literature in terms of the *cognitive dimension* [1]. The *cognitive dimension* suggests that educators should focus on feedback content and help learners solve the problems effectively [1]. Dialogue acts can reflect what educators and learners do in tutoring dialogues, which can demonstrate how educators use instructional strategies to guide learners. Understanding the use of dialogue acts can contribute to the *cognitive dimension* by investigating the educators' instructional strategies in tutoring dialogues. From a practical viewpoint, our results identified several educational practices for instructional strategies in tutoring dialogues. For example, in the effective tutoring dialogue sessions, the educators tended to have more probing questions (e.g., "*How many options can it be?*") to guide the learners without prior progress than those with prior progress. The use of probing questions can teach learners through questioning, instead of providing information hints to learners directly, which can better enhance learners' understanding of the learning task [124]. It should be noted that the learners' prior knowledge level is positively correlated with their engagement level [125]. Thus, for the learner without prior progress, we suggest educators to ask learners probing questions that can enhance learners' understanding of the task and prompt them to engage with the dialogue tutoring process. Overall, we expect that educators might consider referring to the instructional strategies proposed in Chapter 3 with the consideration of learner prior knowledge (e.g., prior progress), which can better guide learners toward the expected achievements.

Insights on the design of automated feedback system. As indicated in [126], the identification of dialogue acts is a significant part of developing dialogue-based ITSs, which can help the system adjust the instructional strategies to support learners. Most previous studies [44, 72, 116, 117, 127, 128] invested much effort in feature engineering and used the features to train many traditional machine learning methods. Motivated by the recent advancement in natural language processing, our study advanced the previous works by employing

the well-established pre-trained language model, BERT [129], for the dialogue act classification. Our study demonstrated promising classification performance in identifying dialogue acts. We open-sourced our dialogue act classifier for researchers and practitioners to reuse, which can be accessed via https://github.com/bertDA/BertDA. Additionally, our study demonstrated that dialogue acts from educators and learners have the potential to predict learner problem-solving performance. By incorporating other non-dialogue-act features (e.g., sentiment level) in the prediction, our predictive model achieved better prediction performance on predicting learner problem-solving performance. Thus, we recommend the researchers and software developers to add dialogue act features in the prediction process.

7.1.3 RQ 3: Enhancement on the identification of dialogue acts

We analysed the misclassified dialogue acts from the classifier described in Chapter 3 and summarised the potential improvements of the dialogue act classifier from the two main aspects: discourse context and sample informativeness. To validate our proposed improvements, in the study reported in Chapter 4, we investigated the impacts of both aspects on the dialogue act classification. The key implications for research and practice are summarised below.

The impact of discourse context. We investigated the impact of discourse context on the classification performance of the dialogue act classifier from two scheme levels where the first-level and second-level schemes have 12 dialogue acts and 31 dialogue acts, respectively. Intuitively, the classification of the second-level dialogue act scheme was more complex than the first-level dialogue act scheme. We found that the incorporation of discourse context could enhance the performance of the dialogue act classifier at both levels. Particularly, when incorporating one preceding sentence, the classification performance of the dialogue act classifier on both levels achieved significant improvements. These findings indicate the necessity of incorporating the discourse context in the dialogue act classification. Thus, we recommend the researchers and practitioners to add at least one preceding sentence to the dialogue act classifier training process.

The efficacy of statistical active learning. We also found that the statistical active learning methods (CoreMSE and CoreLOG [97]) generally tended to alleviate data demand for training the dialogue act classifier. Compared with the random sampling method, active learning methods demonstrated their potential for sampling the highly informative instances for training dialogue act classifiers, which could reduce the human annotation on the repetitive instances and further alleviate the demand for human labour. Our results also provided evidence that the CoreLOG and CoreMSE were effective active learning methods that can be of practical use to alleviate the costs of manual annotation on the task of educational dialogue act classification. As noted by many prior studies [38, 44, 90, 130, 131], the annotation process of dialogue acts is quite time-consuming and labour-intensive. Thus, we recommend that when developing a

dialogue act classifier, researchers might consider incorporating the statistical active learning methods (e.g., CoreMSE and CoreLOG) in their annotation process.

7.1.4 RQ 4: Showing the politeness of instructional strategies in tutoring dialogues

In Chapter 2, we found that the use of politeness might be correlated with learners' problem-solving performance in instructional communication. To understand the types of instructional communication in the tutoring dialogue, in Chapter 3, we presented the study in which we developed the dialogue act classifier, which can be used to identify certain types of instructional communication (e.g., `Probing Question`, `Open Question`, and `Information Hint`). With the use of the politeness levels in the study reported in Chapter 2 and the dialogue acts in Chapter 3, we explored the extent to which educators express politeness in instructional communication in the study reported in Chapter 5. The key implications are summarised below.

Educational practice of expressing politeness in instructional strategies. As noted in the *Feedback Triangle* model [1], three feedback dimensions are always interconnected in terms of mutual support. In Chapter 5, we demonstrated the interplay between the *cognitive dimension* and *social-affective dimension* by investigating the politeness levels of instructional strategies from two aspects. Firstly, the politeness level can be varied in identical instructional strategies. For example, the non-polite expression (or direct expression) of `Negative Feedback` could be "*Not exactly!*" whereas the polite expression could be "*Sorry, I saw the wrong numbers*". Secondly, the politeness level can be varied in different instructional strategies. For example, the average politeness level of `Evaluation Question` (e.g., "Do you understand what I mean?") was generally lower than that of the `Information Hints` (e.g., "*It is P = Force/Area*").

From a practical viewpoint, our results offer several educational practices for expressing politeness in instructional strategies in tutoring dialogues. For example, in the effective tutoring dialogue sessions, the educators tended to ask more polite open-ended questions (e.g., "*What do you think we could try first*") for the learners without prior progress than they did for the learners with prior progress. The open-ended questions often require learners to demonstrate substantive elaboration on the learner's understanding. It should be noted that learners with low prior knowledge are not confident to demonstrate self-explanation activities [132] and they commonly engage more with polite educators [68, 133]. Thus, it is recommended to ask more polite open questions for the learners without prior progress on the learning tasks. Overall, we expect that educators might consider referring to the way of expressing politeness in instructional strategies proposed in Chapter 5 with the consideration of learners' prior knowledge (e.g., prior progress), which can better guide learners toward the expected achievements.

Insights of using instructional strategies politeness on the design of feedback system. Our results indicated that the politeness levels of instructional strategies can support the model to predict learner problem-solving performance, which confirmed our findings in Chapter 2, i.e., the use of politeness might correlate with learner performance in instructional communication. Thus, we suggest that researchers can add the politeness of instructional strategies to the predictive models which could enhance the prediction accuracy. Besides, our results shed light on the design of dialogue-based intelligent tutoring systems regarding the presence of instructional strategies. For example, the hints can be presented into different politeness levels such as **Direct** (e.g., "*Remember the service charge is 0.51 per therm!*"), **Neutral** (e.g., "*The service charge is 0.51 per therm*"), and **Polite** (e.g., "*We need to keep in mind that the service charge is 0.51 per therm* "). The learners with different prior knowledge levels (i.e., with/without prior progress) might have different preferences for polite expression. Driven by this, dialogue-based intelligent tutoring systems can provide various responses regarding the politeness level of the learners so as to deliver more personalised feedback.

7.1.5 RQ 5: Revisit the communicative patterns in assessment feedback

Given the need of building effective feedback, in the study reported in Chapter 6, we further investigated the use of communicative patterns in the assessment written feedback. In line with the tutoring dialogues, we also investigated feedback content based on the *cognitive dimension* and *social-affective dimension* of *Feedback Triangle* framework [1]. We referred to a recent assessment feedback framework (i.e., learner-centred feedback framework [8]), which was built upon the *Feedback Triangle* model [1], for analysing the feedback. The learner-centred feedback framework [8] can further categorise the *cognitive dimension* and *social-affective dimension* into many feedback attributes. We employed well-established tools (e.g., LIWC and Coh-Metrix) to extract textual features from the feedback and used these features to analyse the feedback content. Specifically, we investigated the correlation between the identified features from the educators' feedback and learner subsequent performance. We demonstrated a set of feedback practices for educators to consider. Additionally, beyond the prediction, we proposed to incorporate the SHAP framework to explain the prediction results, which can help educators better understand the adjustment (e.g., provide more detailed suggestions on the learners' works or use more supportive words to motivate learners) to their feedback comments. The key implications are summarised below.

Educational practice of the communicative patterns in assessment feedback. The results in Chapter 6 demonstrated several recommendations for feedback practice for educators. When crafting the feedback content, educators need to take the learner's knowledge level on the initial assignment (i.e., report proposal or Assignment I in our study) into account. For example, high-performing learners on Assignment I often received more positive emotional words

(e.g., comments about encouraging positive affect) than low-performing learners. The occurrence of positive emotional words often positively correlated with the occurrence of highlighting the strengths of learners' performance (e.g., "*Your report is well structured*"). However, we found that the comments involved highlighting the strengths of learners' performance were almost equally distributed among the low- and high-performing learners. Thus, when crafting the feedback content for the initial assignment (i.e., Assignment I), educators may consider identifying learners' strengths in their report writing (e.g., "*The report structure is unclear but the research topic is quite novel*") even though the learners did not perform well on the initial assignment. Positive feedback can affirm learners' understanding of success criteria, enhance learners' self-efficacy, maintain the learner-educator relationship, and mitigate the potential harshness of criticism [8]. Then, informing the insufficient aspects of learners' work as part of feedback is unavoidable. To mitigate the sense of criticism, we also recommend that educators might consider rephrasing the negative feedback in a polite manner (e.g., "*You have made a good effort on the report but the structure is not clear enough.*"). To support low-performing learners improve their work regarding the insufficient aspects (e.g., "*...the structure is not clear enough.*"), educators may consider providing more detailed instructional information to guide learners towards the expected learning performance compared with the high-performing learners. Notably, the expected performance (e.g., the rubric about high achievements in our study) should need to be frequently presented in the feedback. By knowing the gap between current and expected performance, learners could take the advantage of the detailed suggestions to improve their work on the subsequent assignment.

The insights on the design of automated feedback evaluation systems. Our study employed the machine learning models (i.e., Logistic Regression, Random Forest, and Gradient Tree Boosting), which were widely used in assessment written feedback studies [32–35], to predict learner grade changes (i.e., `Increase` or `Not Increase`) on the subsequent assignment; and further interpreted the models' prediction by using the SHAP framework [104]. Informed by [32], building a model for predicting learners' grade changes can be used to evaluate educators' feedback. Specifically, when receiving feedback crafted by novice educators, we recommend that feedback researchers can first use the tools (e.g., LIWC, CohMetrix, and Politeness tools) introduced in Chapter 6 to extract the textual features (e.g., politeness strategies and writing metrics) for analysis from their feedback. Then, feedback researchers may consider using our pre-trained GTB model to predict whether learners can improve their grades on the subsequent assignment based on the extracted features. When the GTB model predicts the `Not Increase` results, we can use the SHAP framework to locate which features contributed most to the predicted results and to inform novice educators how to make the feedback more effective. For example, when novice educators overly use positive emotional words in the feedback, our pre-trained GTB model is likely to predict learners into `Not Increase` on the subsequent

assignment since the high frequency of using positive emotional words in feedback positively correlated with the prediction of `Not Increase` as interpreted by the SHAP framework.

7.2 Limitations and Future works

1. Human evaluation on the effectiveness of communicative patterns. Despite identifying a set of communicative patterns (e.g., In Chapter 2, we found that educators were inclined to use more the word "*we*" to guide the learners without prior progress.) from the feedback in the dialogue form, we should further validate the effectiveness of these communicative patterns on learners' achievements in the real-world teaching and learning environment. Given that educators may use the identified communicative patterns to carry out various teaching practices, we plan to investigate the extent to which learners benefit from the adoption of the identified communicative patterns. As indicated in Chapters 2, 3, and 5, it would be important to conduct a randomised experiment to examine the effectiveness of identified communicative patterns in tutoring dialogues. As we found that the factor of learners' prior progress in tutoring dialogues is a significant factor that is associated with how educators deliver their feedback [37–39], it is also important to include this learner factor in the randomised experiment. As an initial step, it would be promising to allocate educators into a control group and an experimental group. Educators in the experimental group would receive suggestions for communicative patterns on the basis of learners' prior progress. For example, when educators intend to use the non-polite expression to guide learners without prior progress (e.g., "*Solve the first problem*"), they should receive suggestions on elaborating the expressions in a polite manner (e.g., "*We can work on the first problem together*") for educators to consider. In contrast, educators in the control group would not receive any support. Then, the disparities in learning outcomes (i.e., grades) between the control and experimental groups should be measured to validate the effectiveness of our identified communicative patterns. Additionally, it would be useful to conduct follow-up interviews with educators about their viewpoints on the suggestions of communicative patterns and with learners about their learning experience (e.g., whether educators can explicitly and politely guide them towards solving the learning tasks).

2. Exploring the potentials of large language models on feedback generation. As indicated in Chapters 2–6, we aimed to investigate the communicative patterns of feedback from real-world teaching environments and further contribute to the design of Automated Feedback Systems (e.g., dialogue-based intelligent tutoring systems and automated assessment feedback). To date, many automated feedback systems are built upon rule-based systems. For example, a well-established automated feedback system, *OnTask* [51], can provide feedback based on a set of rules (e.g., learner grades and residential status). The OnTask system has demonstrated its effectiveness in supporting learners in many learning subjects (e.g., biological science courses

[134] and health sciences [5]). However, rule-based automated feedback systems inherently encounter the issue of being less flexible as they often rely on a fixed set of rules to generate textual information [135], which might limit the potential of automated feedback systems. Inspired by the recent advancement of natural language generation, we posit that large language models might overcome the issue of rule-based systems and have the potential value to provide automated feedback to learners. For example, ChatGPT is a recently released chat-based generative model, which can generate human-like textual responses [136–138]. Though ChatGPT was originally designed for general communication, we expect that ChatGPT might have the capability to facilitate communication in the educational feedback situation. Driven by this, it is worthwhile to further investigate the extent to which ChatGPT can generate effective communicative patterns in their feedback. By doing so, researchers should identify the communicative patterns from the feedback generated by ChatGPT and further compare the differences in communicative patterns between human educators and ChatGPT. Specifically, researchers might consider using a similar dataset on the assessment written feedback that we presented in Chapter 6. Then, researchers could employ the ChatGPT to assess learner report proposals based on some prompts (e.g., "*Please give feedback on the text in terms of the clarity. <INSERT THE TEXT>*") and provide assessment feedback correspondingly. We recommend that the researchers can follow the approaches presented in Chapter 6 to identify the communicative patterns from the feedback generated by ChatGPT. Finally, researchers can compare the communicative patterns from the feedback generated by ChatGPT and human educators. We deem that the results of the comparison would reveal insights of the feasibility and practicality of deploying ChatGPT as an automated feedback system for the feedback provision.

3. Establish trustworthy automated feedback systems. In Chapters 2-6, we have investigated the communicative patterns from the real-world feedback provision process. The insights summarised in our study would shed light on the design of automated feedback systems. For example, the insights from the *social-affective dimension* would guide the automated feedback systems to deliver personalised feedback content to maintain the relationship between learner and educator, which can help learners engage with the feedback. Though our study obtained many promising findings for the design of the automated feedback systems, to successfully deploy the automated feedback systems in a real-world teaching environment, a necessary step is to establish trust between the automated feedback systems and human stakeholders (e.g., learners, educators, and institutional administrators) [139]. Our recent study [139] raised concerns regarding trustworthiness issues in the development of automated feedback systems from five dimensions, i.e., acceptance, explainability, accountability, fairness, and privacy. These issues could potentially lower human trust in automated feedback systems, which might shape the barrier between automated feedback systems and educational stakeholders [139]. For example, acceptance of artificial intelligence (AI) applications is a significant requirement to establish trustworthy AI [140]. Users, such as educational stakeholders, may have different expectations

of AI applications, and if these expectations are not met, users might be less inclined to use the AI applications [140]. To enhance the acceptance, researchers need to investigate factors (e.g., culture, gender, and age) that might affect users' expectations [140]. Taking the culture variance in feedback as an example, students from various cultural backgrounds might have different perceptions based on feedback expressions [141]. In Chapter 2, we acknowledged that linguistic politeness is a cultural dependant factor and our study found that human educators from the USA cultural background commonly used *please* in tutoring dialogues to guide learners in a polite manner [36, 37]. However, using the word *please* to make a request could be considered awkward in Arab culture [142]. Therefore, it is crucial to incorporate the student's cultural background into the feedback generation process to decide on the use of politeness in the feedback content during the development of automated feedback systems. Driven by this, future research should investigate the trustworthy elements (described in [140, 143]) of automated feedback systems to facilitate their deployment in real-world learning and teaching environments.

4. Enhancing the model robustness on the identification of educational dialogue acts. In Chapters 3 and 4, we developed a dialogue act classifier based on our annotated sentences from educators and learners, which aimed to automate the identification of dialogue acts from the tutoring dialogues. Despite being effective when evaluating the dialogue act classifier, our study encountered the issue of imbalanced label distribution where certain types of dialogue act may be the minority class in the dataset, e.g., `Ready Answer` dialogue act only had 0.07% in the annotated instances. The issues of imbalanced classes might negatively impact the dialogue act classification, especially for identifying crucial but underrepresented classes. It should be noted that the issue of imbalanced class distribution has been widely reported in many prior studies [72, 127, 128]. To obtain more reliable classified dialogue acts, it is necessary to investigate the approach to enhance the capability of the classifier on learning the patterns from different class distributions, which is about the model robustness [144–146]. A robust classifier is able to maintain the performance when the distribution of input data is imbalanced [145]. Driven by this, future research should investigate approaches from the model robustness literature to enhance the robustness of the dialogue act classifier. A possible approach would be to optimise the dialogue act classifier by maximising the *Area Under ROC Curve* (AUC) score which is a metric that can measure the capability of the classifier in distinguishing different classes [144, 145, 147]. Maximizing the AUC score in the model training process has been shown to benefit the performance of classifiers on the highly-imbalanced scenarios in many domains (e.g., medical and biological domains) [144, 145]. However, the effectiveness of applying AUC maximisation approaches in the educational domain is still under-explored. Thus, it is worthwhile to examine the potential values of AUC maximisation in enhancing the robustness of the dialogue act classifier to the imbalanced data distribution.

5. Explore the communicative patterns in multi-modal feedback. Throughout this thesis,

we mainly focused on the analytics of communicative patterns in various types of textual feedback (i.e., assessment written feedback) in Chapters 2–6. The existing literature indicates that learners might benefit from the feedback in different modalities (e.g., video feedback [148–150] and audio feedback [151, 152]) during the learning process [148, 151]. For example, the video feedback was created by educators using a webcam to record the educators' facial expressions, gestures, and verbal communication (e.g., comments on learners' works) [148]. Most learners in Henderson and Phillips study [148] expressed that the video feedback was more personalised, supportive, constructive and clear than the textual feedback. In addition, Henderson and Phillips [148] also proposed a feedback framework to analyse the structure of video feedback. The framework categorised the feedback into different components including *Salutation* (greeting, e.g., "*Hi Tom*") and *Evaluative Summary* (a general evaluative statement of learners' works) [148]. Additionally, the nonverbal communicative patterns (e.g., tone of voice and facial expressions) in the feedback provision process are also important to the design of video feedback. Given the importance of video feedback, for future work, we aim to investigate the communicative patterns in video feedback based on the feedback framework [148], which can further facilitate the practice of delivering video feedback. Specifically, we can analyse the communicative patterns by comparing the learners' grade changes on consecutive tasks. Inspired by the study in Chapter 6, we can measure the effects of feedback by observing the learners' grade changes on consecutive assignments. Thus, future research should also focus on connective assignments where educators craft feedback in multi-modalities. For example, learners will only receive video feedback from their educators throughout the semester. For each video feedback, we suggest annotating the educators' communicative patterns based on the video feedback framework [148]. Then, researchers can investigate a correlation between the features of the communicative patterns in the video feedback and learner grade changes.

Bibliography

[1] Min Yang and David Carless. The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education*, 18(3):285–297, 2013.

[2] Yi-Shan Tsai, Rafael Ferreira Mello, Jelena Jovanović, and Dragan Gašević. Student appreciation of data-driven feedback: A pilot study on ontask. In *LAK21: 11th International Learning Analytics and Knowledge*, pages 511–517, 2021.

[3] Phillip Dawson, Michael Henderson, Paige Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, 44(1):25–36, 2019.

[4] Cathrine Derham, Kieran Balloo, and Naomi Winstone. The focus, function and framing of feedback information: Linguistic and content analysis of in-text feedback comments. *Assessment & Evaluation in Higher Education*, pages 1–14, 2021.

[5] Lisa-Angelique Lim, Shane Dawson, Dragan Gašević, Srecko Joksimović, Abelardo Pardo, Anthea Fudge, and Sheridan Gentili. Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses. *Assessment & Evaluation in Higher Education*, 46(3):339–359, 2021.

[6] Berry M O'Donovan, Birgit den Outer, Margaret Price, and Andy Lloyd. What makes good feedback good? *Studies in Higher Education*, 46(2):318–329, 2021.

[7] Michael Henderson, Rola Ajjawi, David Boud, and Elizabeth Molloy. Identifying feedback that has impact. In *The Impact of Feedback in Higher Education*, pages 15–34. Springer, 2019.

[8] Tracii Ryan, Michael Henderson, Kris Ryan, and Gregor Kennedy. Designing learner-centred text-based feedback: a rapid review and qualitative synthesis. *Assessment & Evaluation in Higher Education*, 46(6):894–912, 2021.

[9] David Boud and Elizabeth Molloy. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, 38(6):698–712, 2013.

[10] John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007.

[11] Naomi Winstone, David Boud, Phillip Dawson, and Marion Heron. From feedback-as-information to feedback-as-process: a linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education*, 47(2):213–230, 2022.

[12] Rola Ajjawi, David Boud, Michael Henderson, and Elizabeth Molloy. Improving feedback research in naturalistic settings. In *The Impact of Feedback in Higher Education*, pages 245–265. Springer, 2019.

[13] Phillip Dawson, Michael Henderson, Tracii Ryan, Paige Mahoney, David Boud, Michael Phillips, and Elizabeth Molloy. Technology and feedback design. *Learning, Design, and Technology*, 2018.

[14] David Carless. Trust and its role in facilitating dialogic feedback. In *Feedback in higher and professional education*, pages 90–103. Routledge, 2012.

[15] David Carless. Feedback as dialogue. 2016.

[16] Edd Pitt and Naomi Winstone. Towards technology enhanced dialogic feedback. *Reimagining university assessment in a digital world*, pages 79–94, 2020.

[17] Hans Schmidt. Communication patterns that define the role of the university-level tutor. *Journal of College Reading and Learning*, 42(1):45–60, 2011.

[18] Rola Ajjawi and David Boud. Examining the nature and effects of feedback dialogue. *Assessment & Evaluation in Higher Education*, 43(7):1106–1119, 2018.

[19] Cristina Mercader, Georgeta Ion, and Anna Díaz-Vicario. Factors influencing students' peer feedback uptake: instructional design matters. *Assessment & Evaluation in Higher Education*, 45(8):1169–1180, 2020.

[20] Jiming Zhou, Yongyan Zheng, and Joanna Hong-Meng Tai. Grudges and gratitude: the social-affective impacts of peer assessment. *Assessment & Evaluation in Higher Education*, 45(3):345–358, 2020.

[21] Yueting Xu and David Carless. 'only true friends could be cruelly honest': cognitive scaffolding and social-affective support in teacher feedback literacy. *Assessment & Evaluation in Higher Education*, 42(7):1082–1094, 2017.

[22] A Espasa, T Guasch, RM Mayordomo, Montserrat Martínez-Melo, and D Carless. A dialogic feedback index measuring key aspects of feedback processes in online learning environments. *Higher Education Research & Development*, 37(3):499–513, 2018.

[23] Anna Steen-Utheim and Anne Line Wittek. Dialogic feedback and potentialities for student learning. *Learning, Culture and Social Interaction*, 15:18–30, 2017.

[24] Lenore Adie, Fabienne van der Kleij, and Joy Cumming. The development and application of coding frameworks to explore dialogic feedback interactions and self-regulated learning. *British Educational Research Journal*, 44(4):704–723, 2018.

[25] Marrigje E Duitsman, Marije van Braak, Wyke Stommel, Marianne ten Kate-Booij, Jacqueline de Graaf, Cornelia RMG Fluit, and Debbie ADC Jaarsma. Using conversation analysis to explore feedback on resident performance. *Advances in Health Sciences Education*, 24:577–594, 2019.

[26] Marion Heron, Emma Medland, Naomi Winstone, and Edd Pitt. Developing the relational in teacher feedback literacy: exploring feedback talk. *Assessment & Evaluation in Higher Education*, pages 1–14, 2021.

[27] Susanne Voelkel, Tunde Varga-Atkins, and Luciane V Mello. Students tell us what good written feedback looks like. *FEBS Open bio*, 10(5):692–706, 2020.

[28] Margaret Price, Karen Handley, and Jill Millar. Feedback: Focusing attention on engagement. *Studies in Higher Education*, 36(8):879–896, 2011.

[29] Richard Higgins, Peter Hartley, and Alan Skelton. Getting the message across: the problem of communicating assessment feedback. *Teaching in Higher Education*, 6(2):269–274, 2001.

[30] Edd Pitt and Lin Norton. 'now that's the feedback i want!'students' reactions to feedback on graded work and what they do with it. *Assessment & Evaluation in Higher Education*, 42(4):499–516, 2017.

[31] David Carless and David Boud. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8):1315–1325, 2018.

[32] Serena Nicoll, Kerrie Douglas, and Christopher Brinton. Giving feedback on feedback: An assessment of grader feedback construction on student performance. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 239–249, 2022.

[33] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. How good is my feedback? a content analysis of written feedback. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 428–437, 2020.

[34] Anderson Pinheiro Cavalcanti, Rafael Ferreira Leite de Mello, Vitor Rolim, Máverick André, Fred Freitas, and Dragan Gaševic. An analysis of the use of good feedback practices in online learning courses. In *2019 IEEE 19th ICALT*, volume 2161, pages 153–157. IEEE, 2019.

[35] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gašević, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, 2022.

[36] Jionghao Lin, David Lang, Haoran Xie, Dragan Gašević, and Guanliang Chen. Investigating the role of politeness in human-human online tutoring. In *International Conference on Artificial Intelligence in Education*, pages 174–179. Springer, 2020.

[37] Jionghao Lin, Mladen Rakovic, David Lang, Dragan Gasevic, and Guanliang Chen. Exploring the politeness of instructional strategies from human-human online tutoring dialogues. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 282–293, 2022.

[38] Jionghao Lin, Shaveen Singh, Lele Sha, Wei Tan, David Lang, Dragan Gašević, and Guanliang Chen. Is it a good move? Mining effective tutoring strategies from human–human tutorial dialogues. *Future Generation Computer Systems*, 127:194–207, 2022.

[39] Jionghao Lin, Wei Dai, Lisa-Angelique Lim, Yi-Shan Tsai, Rafael Ferreira Mello, Hassan Khosravi, Dragan Gasevic, and Guanliang Chen. Learner-centred analytics of feedback content in higher education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, page 100–110, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398657. doi: 10.1145/3576050.3576064. URL https://doi.org/10.1145/3576050.3576064.

[40] Phillip Long, George Siemens, Gráinne Conole, and Dragam Gašević. Proceedings of the 1st international conference on learning analytics and knowledge. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 2011.

[41] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[42] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, 2014.

[43] David J Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218, 2006.

[44] Vasile Rus, Nabin Maharjan, Lasang Jimba Tamang, Michael Yudelson, Susan Berman, Stephen E Fancsali, and Steve Ritter. An analysis of human tutors' actions in tutorial dialogues. In *The Thirtieth International Flairs Conference*, 2017.

[45] Nabin Maharjan, Vasile Rus, and Dipesh Gautam. Discovering effective tutorial strategies in human tutorial sessions. In *The Thirty-First International Flairs Conference*, 2018.

[46] Aysu Ezen-Can and Kristy Elizabeth Boyer. A tutorial dialogue system for real-time evaluation of unsupervised dialogue act classifiers: Exploring system outcomes. In *AIED*, pages 105–114. Springer, 2015.

[47] Danielle R Chine, Pallavi Chhabra, Adetunji Adeniran, Joseph Kopko, Cindy Tipper, Shivang Gupta, and Kenneth R Koedinger. Scenario-based training and on-the-job support for equitable mentoring. In *The Learning Ideas Conference*, pages 581–592. Springer, 2023.

[48] Danielle R Chine, Pallavi Chhabra, Adetunji Adeniran, Shivang Gupta, and Kenneth R Koedinger. Development of scenario-based mentor lessons: An iterative design process for training at scale. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pages 469–471, 2022.

[49] Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469, 2014.

[50] Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *IJAIED*, 24 (3):284–332, 2014.

[51] Abelardo Pardo, Kathryn Bartimote, Simon Buckingham Shum, Shane Dawson, Jing Gao, Dragan Gašević, Steve Leichtweis, Danny Liu, Roberto Martínez-Maldonado, Negin Mirriahi, et al. OnTask: Delivering data-informed, personalized learning support actions. *Journal of Learning Analytics*, 5(3):235–249, 2018.

[52] Danny Yen-Ting Liu, Kathryn Bartimote-Aufflick, Abelardo Pardo, and Adam J Bridgeman. Data-driven personalization of student learning support in higher education. In *Learning analytics: Fundaments, applications, and trends*, pages 143–169. Springer, 2017.

[53] Galina Deeva, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162:104094, 2021.

[54] Ali Alkhatlan and Jugal Kalita. Intelligent tutoring systems: A comprehensive historical survey with recent developments. *International Journal of Computer Applications*, 181(43):1–20, Mar 2019. ISSN 0975-8887. doi: 10.5120/ijca2019918451. URL http://www.ijcaonline.org/archives/volume181/number43/30402-2019918451.

[55] Robert A Sottilare, Arthur Graesser, Xiangen Hu, and Heather Holden. *Design recommendations for intelligent tutoring systems: Volume 1-learner modeling*, volume 1. US Army Research Laboratory, 2013.

[56] Robert A Sottilare, Arthur Graesser, Xiangen Hu, and Benjamin Goldberg. *Design Recommendations for Intelligent Tutoring Systems: Volume 2-Instructional Management*, volume 2. US Army Research Laboratory, 2014.

[57] Robert A Sottilare, Arthur C Graesser, Xiangen Hu, Andrew Olney, Benjamin Nye, and Anna M Sinatra. *Design recommendations for intelligent tutoring systems: Volume 4-domain modeling*, volume 4. US Army Research Laboratory, 2016.

[58] Patricia Albacete, Pamela Jordan, Dennis Lusetich, Irene Angelica Chounta, Sandra Katz, and Bruce M McLaren. Providing proactive scaffolding during tutorial dialogue using guidance from student model predictions. In *International Conference on Artificial Intelligence in Education*, pages 20–25. Springer, 2018.

[59] José Paladines and Jaime Ramirez. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*, 8:164246–164267, 2020. doi: 10.1109/ACCESS.2020.3021383.

[60] Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-computer Studies*, 66(2):98–112, 2008.

[61] Ning Wang, W Lewis Johnson, Paola Rizzo, Erin Shaw, and Richard E Mayer. Experimental evaluation of polite interaction tactics for pedagogical agents. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 12–19, 2005.

[62] Ning Wang and W Lewis Johnson. The politeness effect in an intelligent foreign language tutoring system. In *International Conference on Intelligent Tutoring Systems*, pages 270–280. Springer, 2008.

[63] Swati Gupta, Marilyn A Walker, and Daniela M Romano. How rude are you?: Evaluating politeness and affect in interaction. In *International Conference on Affective Computing and Intelligent Interaction*, pages 203–217. Springer, 2007.

[64] Maria Mikheeva, Sascha Schneider, Maik Beege, and Günter Daniel Rey. Boundary conditions of the politeness effect in online mathematical learning. *Computers in Human Behavior*, 92:419–427, 2019.

[65] Natalie K Person, Roger J Kreuz, Rolf A Zwaan, and Arthur C Graesser. Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction*, 13(2):161–188, 1995.

[66] Benjamin Brummernhenrich and Regina Jucks. Managing face threats and instructions in online tutoring. *Journal of Educational Psychology*, 105(2):341, 2013.

[67] Benjamin Brummernhenrich and Regina Jucks. "He shouldn't have put it that way!" How face threats and mitigation strategies affect person perception in online tutoring. *Communication Education*, 65(3):290–306, 2016.

[68] Bruce M McLaren, Krista E DeLeeuw, and Richard E Mayer. A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies*, 69(1-2):70–79, 2011.

[69] Alexandria Katarina Vail and Kristy Elizabeth Boyer. Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes. In *Intelligent Tutoring Systems*, 2014.

[70] Vasile Rus, Nabin Maharjan, and Rajendra Banjade. Dialogue act classification in human-to-human tutorial dialogues. In *Innovations in smart learning*, pages 185–188. Springer, 2017.

[71] Donald Morrison, Benjamin Nye, Borhan Samei, Vivek Varma Datla, Craig Kelly, and Vasile Rus. Building an intelligent pal from the tutor. com session database phase 1: Data mining. In *Educational Data Mining 2014*. Citeseer, 2014.

[72] Sidney D'Mello, Andrew Olney, and Natalie Person. Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining*, 2(1):1–37, 2010.

[73] Diane Litman and Kate Forbes-Riley. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, 12(2):161, 2006.

[74] Christopher M Mitchell, Eun Young Ha, Kristy Elizabeth Boyer, and James C Lester. Learner characteristics and dialogue: recognising effective and student-adaptive tutorial strategies. *International Journal of Learning Technology 25*, 8(4):382–403, 2013.

[75] Kristy Boyer, Eun Young Ha, Robert Phillips, Michael Wallis, Mladen Vouk, and James Lester. Dialogue act modeling in a complex task-oriented domain. In *Proceedings of the SIGDIAL 2010 Conference*, pages 297–305, 2010.

[76] Vasile Rus, Rajendra Banjade, Nabin Maharjan, Donald Morrison, Steve Ritter, and Michael Yudelson. Preliminary results on dialogue act classification in chatbased online tutorial dialogues. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 630–631, 2016.

[77] Travis Rasor, Andrew Olney, and Sidney D'Mello. Student speech act classification using machine learning. In *Twenty-Fourth International FLAIRS Conference*, 2011.

[78] Aysu Ezen-Can, Joseph F Grafsgaard, James C Lester, and Kristy Elizabeth Boyer. Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pages 280–289, 2015.

[79] Fan Yang and Frederick WB Li. Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education*, 123:97–108, 2018.

[80] Judi Simmons Estes. Teacher preparation programs and learner-centered, technology-integrated instruction. In *Handbook of Research on Learner-centered Pedagogy in Teacher Education and Professional Development*, pages 85–103. IGI Global, 2017.

[81] Penelope Brown and Stephen Levinson. Politeness: Some universals in language usage. Cambridge, UK, 1987. Cambridge University Press.

[82] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1025.

[83] Tong Niu and Mohit Bansal. Polite dialogue generation without parallel data. *Transactions of the ACL*, 6:373–389, 2018.

[84] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 2016 Conference on KDD*, pages 785–794, 2016.

[85] Donald M Morrison, Benjamin Nye, Vasile Rus, Sarah Snyder, Jennifer Boller, and Kenneth Miller. Tutorial dialogue modes in a large corpus of online tutoring transcripts. In *International Conference on Artificial Intelligence in Education*, pages 722–725. Springer, 2015.

[86] Emanuel A. Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 8(4):289–327, 1973. doi: doi:10.1515/semi.1973.8.4.289. URL https://doi.org/10.1515/semi.1973.8.4.289.

[87] Alexandria Katarina Vail and Kristy Elizabeth Boyer. Identifying effective moves in tutoring: On the refinement of dialogue act annotation schemes. In *International Conference on Intelligent Tutoring Systems*, pages 199–209. Springer, 2014.

[88] Nabin Maharjan and Vasile Rus. A tutorial Markov analysis of effective human tutorial sessions. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 30–34, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3704. URL https://www.aclweb.org/anthology/W18-3704.

[89] Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, Wei Liu, Jialie Shen, and Dacheng Tao. Exploring representativeness and informativeness for active learning. *IEEE transactions on cybernetics*, 47(1):14–26, 2015.

[90] Benjamin D Nye, Donald M Morrison, and Borhan Samei. Automated session-quality assessment for human tutoring based on expert ratings of tutoring success. *International Educational Data Mining Society*, 2015.

[91] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, pages 3606–3611, 2019.

[92] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[93] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839*, 2015.

[94] Yang Liu, Kun Han, Zhao Tan, and Yun Lei. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, 2017.

[95] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, 66:861–899, 2019.

[96] Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prevot. The ISO standard for dialogue act annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, 2020.

[97] Wei Tan, Lan Du, and Wray Buntine. Diversity enhanced active learning with strictly proper scoring rules. *Advances in Neural Information Processing Systems*, 34, 2021.

[98] Guang Zhao, Edward Dougherty, Byung-Jun Yoon, Francis Alexander, and Xiaoning Qian. Uncertainty-aware active learning for optimal Bayesian classifier. In *International Conference on Learning Representations, ICLR 2021*, 2021.

[99] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

[100] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, 2020.

[101] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. The politeness package: Detecting politeness in natural language. *R Journal*, 10(2), 2018.

[102] Filipe Dwan Pereira, Samuel C Fonseca, Elaine HT Oliveira, Alexandra I Cristea, Henrik Bellhäuser, Luiz Rodrigues, David BF Oliveira, Seiji Isotani, and Leandro SG Carvalho. Explaining individual and collective programming students' behavior by interpreting a black-box predictive model. *IEEE Access*, 9:117097–117119, 2021.

[103] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[104] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[105] Abdelbaset Almasri, Adel Ahmed, Naser Almasri, Yousef S Abu Sultan, Ahmed Y Mahmoud, Ihab S Zaqout, Alaa N Akkila, and Samy S Abu-Naser. Intelligent tutoring systems survey for the period 2000-2018. 2019.

[106] Nabin Maharjan and Vasile Rus. A tutorial markov analysis of effective human tutorial sessions. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 30–34, 2018.

[107] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-computer Studies*, 70(5):377–398, 2012.

[108] Ying Tang, Joleen Liang, Ryan Hare, and Fei-Yue Wang. A personalized learning system for parallel intelligent education. *IEEE Transactions on Computational Social Systems*, 7(2):352–361, 2020.

[109] Santiago Schez-Sobrino, Cristian Gmez-Portes, David Vallejo, Carlos Glez-Morcillo, and Miguel A Redondo. An intelligent tutoring system to facilitate the learning of programming through the usage of dynamic graphic visualizations. *Applied Sciences*, 10(4):1518, 2020.

[110] Renu Balyan, Tracy Arner, Karen Taylor, Jinnie Shin, Michelle Banawan, Walter L Leite, and Danielle S McNamara. Modeling one-on-one online tutoring discourse using an accountable talk framework. *International Educational Data Mining Society*, 2022.

[111] Arthur C Graesser, Xiangen Hu, Vasile Rus, and Zhiqiang Cai. Conversation-based learning and assessment environments. In *Handbook of Automated Scoring*, pages 383–402. Chapman and Hall/CRC, 2020.

[112] Vasile Rus and Dan Stefanescu. Toward non-intrusive assessment in dialogue-based intelligent tutoring systems. In *State-of-the-art and Future Directions of Smart Learning*, pages 231–241. Springer, 2016.

[113] Sheng Xu, Frank Andrasik, Zhiqiang Cai, and Xiangen Hu. Integrating deep learning to improve text understanding in conversation-based its. *International Journal of Smart Technology and Learning*, 2(4):304–324, 2021.

[114] Richard M Felder and Rebecca Brent. Active learning: An introduction. *ASQ Higher Education Brief*, 2(4):1–5, 2009.

[115] Tracii Ryan, Michael Henderson, Kris Ryan, and Gregor Kennedy. Identifying the components of effective learner-centred feedback information. *Teaching in Higher Education*, pages 1–18, 2021.

[116] Borhan Samei, Vasile Rus, Benjamin Nye, and Donald M Morrison. Hierarchical dialogue act classification in online tutoring sessions. In *International Conference on Educational Data Mining*, pages 600–601, 2015.

[117] Borhan Samei, Haiying Li, Fazel Keshtkar, Vasile Rus, and Arthur C Graesser. Context-based speech act classification in intelligent tutoring systems. In *International Conference on Intelligent Tutoring Systems*, pages 236–241. Springer, 2014.

[118] Maree Keating, Andrew Rixon, and Aron Perenyi. Deepening a sense of belonging. *Journal of Academic Language and Learning*, 14(2):40–56, 2020.

[119] Regina Jucks, Lena Päuler, and Benjamin Brummernhenrich. "i need to be explicit: You're wrong": Impact of face threats on social evaluations in online instructional communication. *Interacting with Computers*, 28(1):73–84, 2016.

[120] Pedro Manuel Moreno-Marcos, Ting-Chuen Pong, Pedro J Munoz-Merino, and Carlos Delgado Kloos. Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access*, 8:5264–5282, 2020.

[121] Georg Gutjahr, Kirthy Menon, and Prema Nedungadi. Using an intelligent tutoring system to predict mathematics and english assessments. In *2017 5th IEEE International Conference on MOOCs, Innovation and Technology in Education (MITE)*, pages 135–140. IEEE, 2017.

[122] Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[123] Danielle R Chine, Shivang Gupta, and Kenneth R Koedinger. Personalized learning2: A human mentoring and ai tutoring platform ensuring equity.

[124] Toyin Tofade, Jamie Elsner, and Stuart T Haines. Best practice strategies for effective use of questions as a teaching tool. *American Journal of Pharmaceutical Education*, 77(7), 2013.

[125] Anmei Dong, Morris Siu-Yung Jong, and Ronnel B King. How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontiers in Psychology*, 11:591203, 2020.

[126] Sidney D'Mello and Art Graesser. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2(4), jan 2013. ISSN 2160-6455. doi: 10.1145/2395123.2395128. URL https://doi.org/10.1145/2395123.2395128.

[127] Aysu Ezen-Can and Kristy Elizabeth Boyer. Understanding student language: An unsupervised dialogue act classification approach. *Journal of Educational Data Mining*, 7(1): 51–78, 2015.

[128] Wookhee Min, Joseph B Wiggins, Lydia G Pezzullo, Alexandria K Vail, Kristy Elizabeth Boyer, Bradford W Mott, Megan H Frankosky, Eric N Wiebe, and James C Lester. Predicting dialogue acts for intelligent virtual agents with multimodal student interaction data. *International Educational Data Mining Society*, 2016.

[129] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[130] Björn Gambäck, Fredrik Olsson, and Oscar Täckström. Active learning for dialogue act classification. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, 2011.

[131] Fabrizio Ghigi, Vicent Tamarit, Carlos-D Martínez-Hinarejos, and José-Miguel Benedí. Active learning for dialogue act labelling. In *Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas de Gran Canaria, Spain, June 8-10, 2011. Proceedings 5*, pages 652–659. Springer, 2011.

[132] Kirsten Berthold, Tessa HS Eysink, and Alexander Renkl. Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, 37(4):345–363, 2009.

[133] Kaśka Porayska-Pomsta, Manolis Mavrikis, and Helen Pain. Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction*, 18 (1-2):125–173, 2008.

[134] Lisa-Angelique Lim, Sheridan Gentili, Abelardo Pardo, Vitomir Kovanović, Alexander Whitelock-Wainwright, Dragan Gašević, and Shane Dawson. What changes, and for whom? a study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, 72:101202, 2021.

[135] Kees Van Deemter, Mariët Theune, and Emiel Krahmer. Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31(1):15–24, 2005.

[136] Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[137] David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Available at SSRN 4337484*, 2023.

[138] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[139] Jionghao Lin, Lele Sha, Yuheng Li, Dragan Gasevic, and Guanliang Chen. Establishing trustworthy artificial intelligence in automated feedback. 2022. URL https://edarxiv.org/5efxn/.

[140] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022.

[141] Carol Evans and Michael Waring. Exploring students' perceptions of feedback in relation to cognitive styles and culture. *Research Papers in Education*, 26(2):171–190, 2011.

[142] Dániel Z Kádár and Sara Mills. *Politeness in East Asia*. Cambridge University Press, 2011.

[143] Stéphan Vincent-Lancrin and Reyer Van der Vlies. Trustworthy artificial intelligence (AI) in education: Promises and challenges. 2020.

[144] Ngoc Dang Nguyen, Wei Tan, Wray Buntine, Richard Beare, Changyou Chen, and Lan Du. Auc maximization for low-resource named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/pdf/2212.04800.pdf.

[145] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[146] Wentao Wang. Obtaining robust models from imbalanced data. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1555–1556, 2022.

[147] Tianbao Yang and Yiming Ying. Auc maximization in the era of big data and ai: A survey. *ACM Computing Surveys*, 55(8), dec 2022. ISSN 0360-0300. doi: 10.1145/3554729. URL https://doi.org/10.1145/3554729.

[148] Michael Henderson and Michael Phillips. Video-based feedback on student assessment: Scarily personal. *Australasian Journal of Educational Technology*, 31(1), 2015.

[149] Paige Mahoney, Susie Macfarlane, and Rola Ajjawi. A qualitative synthesis of video feedback in higher education. *Teaching in Higher Education*, 24(2):157–179, 2019.

[150] T Bahula and R Kay. Exploring student perceptions of video feedback: A review of the literature. *ICERI2020 Proceedings*, pages 6535–6544, 2020.

[151] Tom Lunt and John Curran. 'Are you listening please?' The advantages of electronic audio feedback compared to written feedback. *Assessment & Evaluation in Higher Education*, 35(7):759–769, 2010.

[152] Jill Gould and Pat Day. Hearing you loud and clear: Student perspectives of audio feedback in higher education. *Assessment & Evaluation in Higher Education*, 38(5):554–566, 2013.