Moderation Matters: Measuring Conversational Moderation Impact in English as a Second Language Group Discussion

Rena Gao*, **Ming-Bin Chen***, **Lea Frermann**, **Jey Han Lau**
The University of Melbourne, Parkville, 3052, Australia
{rena.gao,mingbin,lea.frermann}@unimelb.edu.au, jeyhan.lau@gmail.com

Abstract

English as a Second Language (ESL) speakers often struggle to engage in group discussions due to language barriers. While moderators can facilitate participation, few studies assess conversational engagement and evaluate moderation effectiveness. To address this gap, we develop a dataset comprising 17 sessions from an online ESL conversation club, which includes both moderated and non-moderated discussions. We then introduce an approach that integrates automatic ESL dialogue assessment and a framework that categorizes moderation strategies. Our findings indicate that moderators help improve the flow of topics and start/end a conversation. Interestingly, we find active acknowledgement and encouragement to be the most effective moderation strategy, while excessive information and opinion sharing by moderators has a negative impact. Ultimately, our study paves the way for analyzing ESL group discussions and the role of moderators in non-native conversation settings. Code and data are available at https://github.com/RenaGao/L2Moderator.

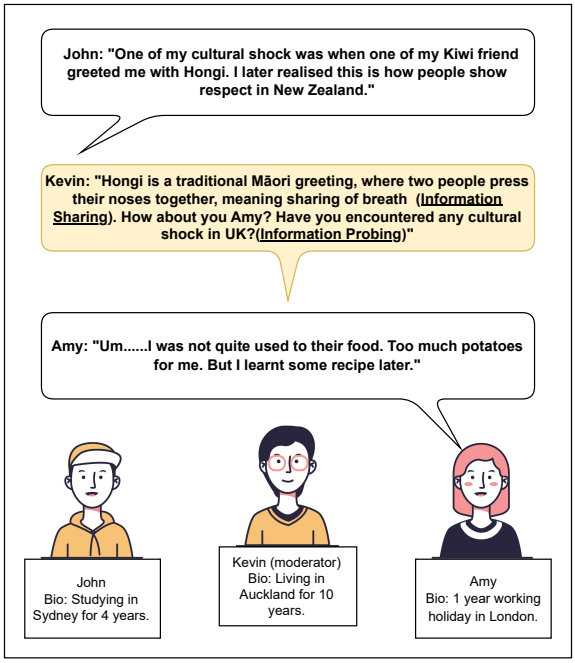

Figure 1: Example of a moderated conversation involving ESL Asian speakers. Highlighted dialogue bubbles represent moderator interventions, while underlined text indicates the tags of the moderation strategies.

1 Introduction

Participation in group discussions has been widely recognized as an effective means for language acquisition (Crisianita and Mandasari, 2022; Pica and Doughty, 1985; Hudgins and Edelman, 1986). However, numerous studies across diverse fields have highlighted the challenges encountered by English as a Second Language (ESL) learners—particularly those from Asian linguistic and cultural backgrounds—when interacting in English group discussions (King, 2013; Lee, 2009; Li and Jia, 2006; Yang, 2010). As illustrated in Figure 1, introducing a moderator into the discussion has been identified as an effective solution to enhance ESL speakers' participation (Hamzah and Asokan, 2016). Moderators assist through various interventions such as providing guidance (Hamzah and Asokan, 2016), which help to support engagement (Reddington, 2018) and communication (Vasodavan et al., 2020). Existing ESL assessment tools predominantly focus on evaluating writing skills and lack in support for dynamic spoken interactions, especially in group settings. While recent studies have begun exploring the measurement of interaction and engagement in two-party second language conversations (Gao et al., 2025, 2024), a notable gap remains: how to assess engagement and interaction in multi-party ESL discussions. Furthermore, it's unclear how moderator interventions influence ESL group dynamics.

In this study, we seek to quantify the impact

*Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2025, pages 2070–2095

of moderators on participants' dialogue quality directly from dialogue transcripts (without relying on external measures such as questionnaires). Specifically, we investigate whether moderator presence influences performance and which moderation strategies are most effective to stimulate discussion. To this end, we develop a dataset of 17 ESL group discussions, covering both moderated and non-moderated conditions. Building on recent advancements in ESL dialogue evaluation (Gao et al., 2024) and conversational moderation analysis (Chen et al., 2024), we develop an approach that automates the assessment of ESL group discussions based on the transcripts and examines moderation strategies by categorizing them into ten distinct types tailored for ESL contexts. We find that active acknowledgment and encouragement from the moderator is the most effective strategy to improve ESL discussions, while excessive information or opinion sharing has the opposite effect. In sum, our contributions are:

- We introduce the first moderated multi-party ESL group discussion conversation dataset, comprising 17 sessions (total 16.5 hours of recorded transcripts and 9,843 sentences), which includes parallel moderated and non-moderated sessions on the same set of topics.
- By integrating the dialogue quality assessment approach from Gao et al. (2024) with the WHoW moderation analysis framework from Chen et al. (2024), we introduce a novel method that offers an automatic, fine-grained evaluation of dialogue quality and moderator intervention in ESL group discussions.
- Comparative analysis reveals that the presence of moderator significantly improves topic management and the quality of conversation openings and closings. The most effective moderator intervention is active acknowledgement and encouragement; and the least is information/opinion sharing.

2 Related work

Barriers to participation in group discussions for ESL speakers ESL speakers often encounter significant barriers to effective communication and participation in group discussions, stemming from cross-language ability challenges (Higuchi et al., 2023). Unlike native speakers, ESL speakers may struggle to construct sentences that accurately convey their intended meaning and ideas (Turnbull, 2015; Hessel, 2017). They often require additional time to formulate responses in spoken interactions (Lam, 2006), which can be particularly problematic in informal discussions where quick exchanges are expected. Moreover, ESL speakers are less likely to initiate conversations (Tan et al., 2020), further limiting their participation. Above barriers disrupt the flow of conversations, increase the risk of misunderstandings, and significantly reduce ESL speakers' engagement in group discussions (Sampson, 2024; Gao and Wang, 2024).

Effectively participating in group interactions is essential for ESL speakers (Rao, 2019). Interactive conversations provide ESL speakers with exposure to the practical use of language, allowing them to learn how to naturally communicate (Gao and Wang, 2024; Rao, 2019). This learning cannot be fully replaced by passive methods such as reading or listening (Wu and Roever, 2025; Maarof et al., 2018). Consequently, ESL speakers face a dilemma: improving their language skills requires active participation, yet their limited proficiency and confidence often undermines their ability to engage effectively. To address this, educational strategies such as structured group discussions and explicit guidance are commonly employed to facilitate meaningful participation (Pino-Silva and Mayora, 2010; Webster, 2012).

Dialogue quality evaluation Most existing dialogue evaluation methods focus on assessing the quality of machine-generated dialogues, emphasizing features like fluency (Ou et al., 2024), grammar (Lin and Chen, 2023), and accuracy (Han et al., 2022; Chen et al., 2023), while overlooking nuanced aspects such as interactions. These approaches primarily assess machine responses in isolation and often lack interpretability (Smith et al., 2022). In contrast, traditional evaluations of human-to-human conversations rely heavily on manual coding and human interpretation (O. Nyumba et al., 2018; McKenzie and Murphy, 2000), which, although detailed, are limited in scalability.

Recent studies have begun utilizing large language models (LLMs) to evaluate human dialogues, such as classroom interactions (Long et al., 2024). In the context of ESL conversation, Gao et al. (2024) proposed a fine-grained automatic evaluation tool that assesses dialogue quality on both mi-

cro (e.g., reference word usage) and macro levels (e.g., tone of utterance) for ESL speakers. Building on this foundation, our study adapts their framework to evaluate the quality of ESL dialogues.

Moderation in ESL group discussions Previous studies have documented that moderators employ strategies such as linguistic scaffolding (Kayi-Aydar, 2013; Gagné and Parks, 2013), providing instructions (Hamzah and Asokan, 2016), seeking clarification, and offering acknowledgments (Braham and Piela, 2009) to mitigate linguistic disparities (Jones, 1999), bridge cultural gaps (Osman and Herring, 2007), and address knowledge deficiencies (Asterhan and Schwarz, 2010; Vasodavan et al., 2020). However, these findings rely on manual evaluation methods such as interviews, case studies, and surveys (Osman and Herring, 2007; Hew and Cheung, 2008; Hamzah and Asokan, 2016; Kayi-Aydar, 2013), limiting scalability and cross-context applicability.

Large-scale analyses of dialogue transcripts typically use dialogue acts to categorize speakers' intentions (D'Andrade and Wish, 1985). While existing moderation dialogue act schema (e.g., Park et al. (2012)) provide a structured approach, they may not fully address the specific needs of ESL moderation. At the same time, defining entirely new dialogue acts in isolation could hinder cross-domain comparability. To address this challenge, we develop a tailored set of dialogue acts by adapting the WHoW moderation analysis framework. This approach ensures that the dialogue acts capture the nuances of ESL moderation while remaining compatible for broader cross-domain comparisons of moderator behavior.

In summary, there exists three major gaps: the absence of an automated method for measuring dialogue quality among ESL speakers in group discussions, the need for dialogue act schema specifically tailored for ESL group discussions moderation analysis, and the lack of quantification of the impact of conversation moderation in ESL settings.

3 Dataset Development

Data sets of moderated, multi-party ESL discussions are scarce. To address this gap, we created the first corpus of ESL group discussions with paired moderated and non-moderated conditions. The structure and materials for the discussions come from an existing online ESL Conversation Club, designed to foster reflective and thought-provoking conversations for Asian ESL speakers that go beyond everyday small talk.

Discussion Topic	mod ESL club	mod student	non-mod student
Laugh: What makes you laugh? What does it convey and how is it perceived?		✓	✓
Romance: Romantic relationships in modern society.	✓	✓	✓
Stress: How to deal with stress?	✓	✓	✓
Boss: How to cope with bosses with different cultural backgrounds.	✓	✓	✓
Time: Perspectives on how we perceive, manage, and value time.	✓		✓
AI: How AI impacts our daily life now and the future?		✓	✓
Ghost: Are you superstitious? Why do people believe in 'weird' things?		✓	✓

Table 1: The seven discussion topics and their inclusion in different session settings: moderated (mod) online discussion club sessions, moderated student volunteer discussion sessions, and non-moderated student volunteer discussion sessions.

3.1 ESL Conversation Session

Each session focuses on a central theme (e.g., "How AI impacts our daily life now and in the future", Table 1), accompanied by a background paragraph and three to five discussion questions (e.g., "Compared to AI, what are the things humans possess that are irreplaceable?"). All discussion materials are provided in Appendix F. Sessions typically last 45 to 75 minutes and involve four to eight speakers online, and are moderated by experienced ESL moderators who have lived in English-speaking countries and have prior teaching experience. The moderator's responsibilities include (1) guiding participants in addressing the discussion questions, (2) fostering a friendly and engaging discussion environment, and (3) monitoring the discussion flow and timing.

3.2 Data Collection

The collected data were sourced from two formats: (1) participants of the online ESL conversation club and (2) international volunteer students. Data collection began with obtaining transcripts from the conversation club, hosted and recorded via Zoom, resulting in four sessions.[1] This primary dataset

[1]To preserve the original content and linguistic characteristics of ESL speakers, we do not edit the transcripts for grammar errors. Although common grammar errors (e.g.,

Source	Online club	Student volunteer		Total
Moderated	☑	☑	⊗	—
Sessions	4	6	7	17
Speakers Avg	6.5	6	5.3	44*
Segments Avg	9.3	9.7	7.9	150
Sentences Avg	836	591	421	9843
Mod Sent Avg	374	298	0	3284
Tokens Avg	9450	7364	4595	114152

Table 2: Descriptive statistics of the collected ESL group discussion transcripts, including the average number of speakers, segments, sentences, moderator sentences, and tokens per session type, and grand totals (right). *The total value for speakers represents the number of unique individual participants.

served as an initial field study to explore the general patterns and structure of the discussions.

We collected additional conversational data in a more controlled setting. In this setting: (a) the demographics of participants are more consistent; and (b) for each topic we ran two sessions, one with moderator and the other without (there's no overlap in terms of participants for the two sessions).[2] All participants were native speakers of Chinese, most were postgraduate students or recent graduates who had been living in an English-speaking country for less than three years (IELTS scores of 6.5 to 7.5). Recruitment advertisements were distributed via social media platforms and the school's email channels. Participation was entirely voluntary and primarily motivated by participants' desire to practice English or engage in topic discussions. Consent form for data collection is provided in Appendix G.

Our final data set (Table 1) comprises four club sessions, and thirteen controlled discussion sessions (six moderated, seven non-moderated sessions).[3] Seven conversation topics were selected, all of which had been recently used in the conversation club. Four of these topics overlapped with the four earlier recorded sessions from the club.

We segmented the transcripts into shorter sections to facilitate our analysis. The nuanced linguistic structures and contextual dependencies in second-language conversations necessitated a manual approach to ensure contextually meaningful segmentation (Gao et al., 2025). Two authors manually segmented each session based on sub-topic transitions within the main discussion theme, resulting in 6–11 segments per session. These segments served as the basis for dialogue interactivity quality evaluations and subsequent moderation dialogue act analyses. In total, we collected 16.5 hours of transcripts spanning 17 sessions. Descriptive statistics for the dataset are presented in Table 2.

4 Method

To systematically assess the impact of moderation in ESL discussions, we first adapted the WHoW moderation framework (Chen et al., 2024) and applied topic modeling to our conversation transcripts to identify ten ESL-specific moderation strategies. Next, we incorporated an automated evaluation method for ESL group discussion quality based on Gao et al. (2024). By comparing moderated and non-moderated sessions and analyzing the relationship between moderation strategies and dialogue quality, our approach quantifies the moderator's impact in ESL discussions, providing a data-driven method to evaluate moderation effectiveness.

4.1 Discovery of ESL Moderation Strategies

Conversational strategy analysis typically involves using dialogue acts—labels that represent speakers' intent—to identify sequential patterns within dialogue (Chawla et al., 2022). In this study, rather than relying on existing dialogue acts developed for other contexts, we derived a set of ten ESL-specific moderation strategies (ESLMOD) by adapting the domain-agnostic WHoW moderation analytic framework (Chen et al., 2024). Definitions for the ESLMOD strategies are in Table 3 and examples of each strategy in Appendix Table 17.

The WHoW framework proposes a generic set of moderator-specific facilitation strategies, validated on moderated debates and panel discussions. It characterizes a moderator's role along three dimensions: Motives ("Why"), Dialogue Acts ("How"), and targeted speakers ("Who"). It defines three motive categories—informational, social, and coordinative—and six dialogue acts: probing, confronting, instruction, interpretation, supplement, and utility (see Appendix C for detailed definitions and examples). Following the WHoW framework,

"He don't like it.") and incomplete sentences were present, manual review confirmed that the majority of the content was intelligible.

[2]Some students attended only a single session, while others participated in multiple sessions. To ensure balanced exposure, students who joined multiple sessions were assigned to both moderated and non-moderated discussions. Note, however, that for the same topic, a student is only allowed to participate in either the moderated or non-moderated session.

[3]We lost one moderated student volunteer session for the topic "Time" due to technical issue during recording.

Strategy	Source	Definition
Information Probing	I & Probing	Prompting participants to share thoughts, opinions, knowledge or experiences.
Opinion Sharing	I & Supplement	Express personal views, beliefs, or subjective opinions related to the topic.
Information Sharing	I & Supplement	Provide factual, contextual content or knowledge to inform or orient others.
Experience Sharing	S & Supplement	Share a personal experience or anecdote.
Echoing	I/S & supplement	Reinforce or support a prior statement by sharing similar views and thoughts.
Informational Interpretation	I & Interpretation	Interpret, clarify, reframe, summarize, paraphrase, or make connections to earlier conversation content.
Acknowledgement	S & Supplement	Recognize, validate, or show appreciation for another participant's contribution, insight, or effort.
Backchanneling	S & Utility	Brief verbal or non-verbal responses for indicating active listening, understanding, or agreement.
Social Utility	S & utility	Use polite or respectful phrases to show courtesy.
Coordinative Instruction	C & Instruction	Explicitly command, influence, or halt the immediate behavior of the recipients for coordinating the process of the session.

Table 3: ESLMOD strategies classes, along with their corresponding source WHoW motive labels (I for informational, S for social, and C for coordinative) and dialogue act labels and definitions.

Motive/Dialogue Act	Probing	Confronting	Instruction	Interpretation	Supplement	Utility	Total
Informational	**0.26(838)**	0.01(17)	0.00(11)	0.09(290)	**0.37(1223)**	0.01(22)	0.73(2401)
Coordinative	0.01(37)	0.00(2)	0.04(119)	0.00(6)	0.01(35)	0.01(19)	0.07(218)
Social	0.02(68)	0.00(3)	0.00(10)	0.02(80)	**0.12(401)**	**0.10(315)**	0.27(877)
Total	0.26(856)	0.01(19)	0.04(124)	0.10(342)	0.46(1527)	0.13(416)	1(3284)

Table 4: Probabilities (frequencies) of motives (rows) and dialogue acts (DA; columns), and conditional probabilities of DA given motive (cells), as identified using the WHoW framework. Bold values highlight intersected categories that surpass the threshold (0.1) for further domain adaption.

we used GPT-4o (OpenAI, 2024) to automatically annotate all moderator sentences in our data for these dimensions. We also prompted GPT-4o for a reason/justification of the predicted labels.[4]

Table 4 presents the distribution of WHoW motives and dialogue acts, highlighting that the moderator's primary motivation is informational (73%), with a secondary focus on fostering a social atmosphere and building relationships (27%), and minimal emphasis on coordinating procedural rules or program-related aspects. In terms of functional roles (columns), moderators are heavily involved in supplementing information (46%), driven by both informational (37%) and social motives (12%). Additionally, moderators frequently probe participants for input (26%), interpret responses (10%), and demonstrate active listening through various utility acts (13%), such as back-channeling.

While WHoW could capture general moderation characteristics, its dialogue act categories are too broad to provide practical insights for ESL discussions. To refine these categories, we used K-means clustering to identify domain-specific dialogue acts (Rus et al., 2012). Based on initial WHoW predictions (Table 4), we applied frequency thresholds to identify the most dominant categories (12 categories with an intersecting probability $< 2.5\%$ were removed). Of the remaining categories, four were selected for adaptation, each accounting for $> 10\%$ of instances. Two categories falling between 2.5% and 10% were included in the final list without further refinement.

For each selected prominent category, we extracted all sentences with the corresponding WhoW label, and the reasoning generated during the WHoW annotation which reflects the moderator's intent. We then applied a separate BERTopic model (Grootendorst, 2022) with k-means cluster-

[4]We manually validated the GPT-4o predictions of dialogue acts in 3 out of the 10 moderated sessions, achieving a macro-accuracy of 0.71 and a macro-F1 score of 0.6, both of which surpass the performance reported in the original study for debate and panel scenarios.

ing to the generated reasonings (separately for each prominent category). The topic model identified fine-grained sub-categories for each original category. We validated our result using topic coherence, obtaining a refined set of domain-specific moderation strategies for ESL discussions.[5]

After this process, we arrive at 10 ESL-specific moderation strategies (Table 3). Specifically, the "Informational & Supplement" category was subdivided based on the type of information shared ("Opinion Sharing", "Information Sharing", and "Echoing"). Similarly, the "Social & Supplement" category was further divided into three subcategories: "Experience Sharing", "Acknowledgment", and "Echoing". These subcategories are distinguished by the presence of agreement and the extent to which the discussion is expanded.[6] We merged instances of "Echoing"—previously identified separately from the two intersecting categories—into a single strategy because it encompasses both social motives (relating to or agreeing with the participant) and informational motives (expanding the discussion). Additionally, the "Social Utility" category was divided into "Backchanneling" (e.g., "Hmm") and "Social Utility" (e.g., "Thank you!").

With this refined set of strategies, we updated the WHoW prompt and re-annotated the moderator sentences (see Appendix Table 18 for the revised prompt). To validate the new prompt, we randomly sampled 10 instances per refined category from the annotations generated by GPT-4o. Three PhD students, who were briefed on the strategy definitions and provided with examples, independently reviewed the annotations, either confirming the assigned label or selecting an alternative. This validation process yielded a macro-accuracy of 0.74 and a Krippendorff's α of 0.41, slightly exceeding WHoW's annotation performance (macro-accuracy = 0.71) and demonstrating moderate reliability. To conclude, our adaptation process has produced 10 conversation moderation strategies tailored to ESL group discussion moderation analysis.

[5]Further details on the domain adaptation process and validation criteria are provided in Appendix E.

[6]While "Acknowledgment" shows agreement or appreciation without expanding the discussion, "Experience Sharing" does not necessarily include explicit agreement, and "Echoing" both affirms the participant's contribution and elaborates on it.

4.2 ESL Multi-party Dialogue Evaluation

We use the dialogue evaluation framework of Gao et al. (2025), which provides: (1) micro-level; (2) macro-level; and (3) overall dialogue quality assessment. At the micro-level, the framework evaluates 17 fundamental linguistic properties, such as "Code Switching", "Feedback in Next Turn", and "Negotiation of Meaning". At the macro-level, it assesses discourse quality through 4 dimensions: Topic Management, Tone Appropriateness, Conversation Openings, and Conversation Closings. They are rated on a standardized scale from 1 to 5, where 1 represents poor and and 5 high quality (details in Appendix B). The overall dialogue quality in terms of conversation naturalness, the achievement of communication purposes and interactive participation engagements, was defined and evaluated by Gao et al. (2024) who automated and extended this framework using large language models (LLMs), showing that LLMs like GPT-4o can accurately predict overall dialogue quality, based on four macro-level quality scores (topic management, tone appropriateness, conversation opening/closing) and 17 micro-level features.

As the dialogues in Gao et al. (2024) only have two speakers with similar proficiency, they computed one overall dialogue and 4 macro-level quality scores for each conversation. Given that our ESL conversations have *multiple* speakers, we adapted their framework to compute an overall dialogue score and the macro-level quality scores for *each speaker* in a conversation. The full prompt is given in Appendix B.

5 Analysis and Results

In Section 5.1, we characterize the ESL moderator behavior and role by examining the frequency distribution of the ESLMOD strategies. In Section 5.2, we assess the dialogue quality of group discussions and compare the outcomes of moderated versus non-moderated sessions. Finally in Section 5.3, we identify most effective strategies for enhancing dialogue quality by analyzing the correlation between moderation strategies and dialogue quality.

5.1 Characterizing Conversational Moderation in ESL Group Discussion

Table 5 presents the frequency distribution of ESLMOD strategies across all moderator sentences from the moderated sessions, including both online club and student discussion sessions. The data

Moderation Strategy	Frequency	%
Information Probing	849	25.8%
Information Interpretation	548	16.7%
Information Sharing	430	13.1%
Backchanneling	318	9.7%
Opinion Sharing	316	9.6%
Experience Sharing	247	7.5%
Acknowledgment	193	5.9%
Echoing	179	5.5%
Coordination Instruction	146	4.4%
Social Utility	58	1.8%

Table 5: Frequency distribution of ESLMOD strategies in moderator sentences across online club and student discussion sessions.

Moderated	Dialogue Quality	Topic Managmt	Tone Choice	Conv Open	Conv Close
☑	**4.14**	**4.03**	3.48	**4.25**	**4.62**
✪	3.41	2.42	2.68	2.19	2.07

Table 6: Dialogue quality scores for moderated (☑) and non-moderated sessions (✪). Significant differences under one-tailed t-test ($p<0.05$) are bolded.

reveals that the moderator primarily employs strategies that encourage in-depth exploration and clarification of information, with the highest frequencies observed in "Information Probing" (25.8%) and "Information Interpretation" (16.7%). This pattern suggests a strong focus on eliciting detailed responses and ensuring that participants articulate their thoughts clearly. This is particularly beneficial in language learning contexts, as it helps reinforce comprehension, critical thinking, and language practice. Regarding the type of content shared by the moderator, "Information" (13.1%) and "Opinion" (9.6%) prevail over their own "Experience" (7.5%), indicating a greater emphasis on intellectual exchange rather than personal interaction in discussions. Strategies involving social support, such as "Acknowledgment" (5.9%) and "Echoing" (5.5%), are comparatively rare.

5.2 Comparative Dialogue Quality Evaluation

To evaluate moderation's impact on dialogue quality and interactivity, we analyze *only* the controlled and paired student volunteering sessions (N=6). As shown in Table 6, the moderated sessions achieve a significantly higher overall dialogue quality score (4.14) than non-moderated ones (3.41). Conversation Opening and Closing significantly improve

Topic	Overall Quality	Topic Managmt	Tone choice	Conv Open	Conv Close
AI	$1.09^{\uparrow}$	$0.72^{\uparrow}$	$0.28^{\uparrow}$	$2.07^{\uparrow}$	$1.74^{\uparrow}$
Boss	$2.01^{\uparrow}$	$1.88^{\uparrow}$	$0.90^{\uparrow}$	$2.01^{\uparrow}$	$1.72^{\uparrow}$
Ghost	$1.92^{\uparrow}$	$1.72^{\uparrow}$	$0.83^{\uparrow}$	$2.71^{\uparrow}$	$2.19^{\uparrow}$
Laugh	$1.21^{\uparrow}$	$2.01^{\uparrow}$	$0.07^{\uparrow}$	$2.19^{\uparrow}$	$2.01^{\uparrow}$
Stress	$1.09^{\uparrow}$	$1.02^{\uparrow}$	$0.28^{\uparrow}$	$1.08^{\uparrow}$	$0.87^{\uparrow}$
Romance	$0.87^{\uparrow}$	$0.93^{\uparrow}$	$0.48^{\uparrow}$	$1.29^{\uparrow}$	$1.28^{\uparrow}$

Table 7: The difference in dialogue quality scores comparing moderated and non-moderated sessions within the same topic. ↑ denote improvement.

Moderated	Dialogue Quality	Topic Managmt	Tone Choice	Conv Open	Conv Close
☑	**3.60**	**4.30**	3.19	**4.02**	**4.31**
✪	2.73	2.86	3.17	2.80	2.91

Table 8: Dialogue quality scores for moderated (☑) and non-moderated (✪) sessions, when controlling for speakers. Significant differences under one-tailed t-test ($p<0.05$) are bolded.

from 2.19 and 2.07 to 4.25 and 4.62, while Tone Choice shows only slight improvement. These results indicate that moderation enhances overall dialogue quality, particularly in structuring ESL multiparty discussions across all sessions with moderator interventions. To ensure the results in Table 6 are not affected by topic or speaker differences, we conducted two additional analyses: (1) controlling for session topics and (2) controlling for speakers. When we controlled for topic (Table 7), we saw consistent improvements for all scores across topics, even though the magnitude of the gain changes depending on the topic ("Romance" e.g., seem to benefit least from the moderator's presence). In line with Table 6, tone choice had the smallest improvement.

Table 8 compares speakers who participated in both moderated and non-moderated sessions to ensure the variance of +moderator sessions and -moderator sessions are not due to participants individual differences, highlighting differences in dialogue quality across four interactivity aspects. Moderated sessions yield higher and more consistent scores, demonstrating the moderator's role in enhancing dialogue quality and reducing variability. In contrast, non-moderated sessions show greater score variability and lower averages, indicating that ESL speakers face significant challenges in managing discussions without external support.

Moderation Strategy	Mean	Difference	p-value
Echoing	3.26	$0.51^{\uparrow *}$	0.04
Backchanneling	3.17	$0.77^{\uparrow}$	0.09
Experience Sharing	3.11	$0.12^{\uparrow}$	0.61
Coordination Instruction	3.09	$0.17^{\uparrow}$	0.46
Social Utility	3.09	$0.07^{\uparrow}$	0.79
Acknowledgement	3.06	$0.09^{\uparrow}$	0.77
Information Probing	3.02	$-1.23^{\downarrow}$	0.16
Information Interpretation	3.01	$-0.33^{\downarrow}$	0.54
Information Sharing	2.95	$-0.54^{\downarrow}$	0.06
Opinion Sharing	2.89	$-0.68^{\downarrow *}$	0.03

Table 9: Mean dialogue quality scores for segments involving the specified ESLMOD strategy, alongside the score differences compared to segments without the strategy. Statistically significant differences (determined by a two-sample t-test, $p < 0.05$) are marked with an asterisk (*).

5.3 ESLMOD Strategies Comparison

To evaluate the effectiveness of ESLMOD strategies, we first computed an overall dialogue quality score Q for each dialogue segment D. For each D, each non-moderator speaker s_i is assigned a dialogue quality score q_i (ranging from 1 to 5), computed from Subsection 5.2. Let $\mathcal{S}$ denote the set of non-moderator speakers in D and let t_i denote the number of tokens contributed by s_i; we compute Q by weighting the speaker scores q_i:

$$Q = \sum_{s_i \in \mathcal{S}} \frac{t_i}{\sum_{s_j \in \mathcal{S}} t_j} \cdot q_i$$

This approach ensures that speakers with greater contributions have a proportionally larger influence on the segment's quality score. Next, to assess the impact of a specific moderation strategy m, we computed the difference of Q between segments with the strategy ($\mathcal{D}^{\mathrm{m}}$) and those without ($\mathcal{D}^{\not\mathrm{m}}$):

$$\Delta_m = \frac{\sum_{D_k \in \mathcal{D}^{\mathrm{m}}} Q_k}{|\mathcal{D}^{\mathrm{m}}|} - \frac{\sum_{D_k \in \mathcal{D}^{\not\mathrm{m}}} Q_k}{|\mathcal{D}^{\not\mathrm{m}}|}$$

We evaluate the statistical significance of using a two-sample t-test. Table 9 presents the mean overall score (Q) for segments incorporating specific moderation strategies, along with difference (Δ_m) and p-value. Generally speaking, strategies involving the moderator's sharing of content—such as information sharing, opinion sharing, and personal interpretations—tend to result in lower dialogue quality scores compared to segments where these strategies are absent. In contrast, strategies rooted in social motivation, including acknowledgment, social utility, experience sharing, echoing, and backchanneling, are with higher scores.

When we focus on the most significant strategies (low p-value), the most effective moderator interventions are "Echoing" (e.g. "Yeah, absolutely, I feel like even just showing the willingness to to do or share housework can be applause.") and "Backchanneling" (e.g. "Okay"). On the other hand, the least effective ones are "Opinion Sharing" and "Information Sharing". This suggests that positive feedback is most helpful for ESL speakers as it helps build their confidence (and excessive information or opinion sharing might have the opposite effect). Moreover, comparing "Echoing" with simple "Acknowledgment" reveals that superficial positive feedback (e.g., "That is interesting.") yields only minimal improvement, underscoring the need for engaging responses that extend the discussion. These findings align with prior educational studies (Neusiedler, 2024; McClure and Vasconcelos, 2011; Gao and Wang, 2024), which emphasize the importance of recognition and positive engagement in enhancing ESL learners' participation.

6 Conclusions

In this study, We present the first ESL group discussion corpus, comprising 17 sessions (both moderated and non-moderated). Building on previous research in ESL dialogue evaluation (Gao et al., 2025) and moderation analysis (Chen et al., 2024), we identify and integrate ten strategy classes with automated dialogue quality evaluation, allowing for the analysis of moderator influence without relying on external measures such as questionnaires. Our analysis shows that moderated sessions consistently exhibit a higher dialogue quality with less variability. Analysis of moderator behavior reveals a focus on informational exchange, but our comparative analysis suggests that socially motivated strategies improve dialogue quality. Notably, "Echoing" and "Backchanneling" are the most effective strategies, while "Opinion Sharing" and "Information Sharing" the least. These results show the value of social engagement and active listenership in ESL discussion moderation.

All participants in this study were ESL speakers from similar cultural backgrounds. Previous research suggests that discussions involving a mix of ESL and native speakers, or those led by native-speaking moderators (Zhu, 2001; Freiermuth, 2001), may produce significantly different dynamics. Future research could extend our approach to more diverse ESL dialogue settings, in-

cluding interactions with native English speakers, to explore how cultural and linguistic diversity influences moderation effectiveness.

Limitations

This study has several limitations. First, the dataset is relatively small, which may limit the generalizability of the findings to broader contexts or diverse dialogue scenarios. Second, the analysis is based on sessions moderated by only two individuals, potentially introducing bias due to their specific moderation styles. And, it includes only language learners with an Asian background, and Chinese as native language. As a result, the findings may not fully capture the variability in moderation practices, nor their effects on dialogues involving participants from diverse or multiple cultures.

Ethics Statement

This study was approved by the The University of Melbourne ethics board (Human Ethics Committee LNR 1D), Reference Number 2022-24988-32929-3, and data acquisition and analysis has been taken out to the according ethical standards. Personal identifiable information of all speakers as well as potentially offensive content was manually removed from the conversation transcripts.

Acknowledgements

The project research is supported by the Australian Research Council Linkage Project (ID: LP210200917). We extend our sincere gratitude to the University of Melbourne master's students and external participants whose generous contributions of time and insight through recorded discussions were vital to the completion of this research. Their active participation and thoughtful input significantly enriched the study. Special appreciation is also owed to Mrs. Elda Chang, organizer of the SoulReMe English Conversation Club, and Prof. Carsten Roever, who supported the participants advertisement for this research.

References

Christa SC Asterhan and Baruch B Schwarz. 2010. Online moderation of synchronous e-argumentation. *International Journal of Computer-Supported Collaborative Learning*, 5:259–282.

Julia Braham and Anna Piela. 2009. Acknowledgement of others' contributions as a peer facilitation skill in online discussions. In *Collected Conference Papers and Abstracts September 2009*, page 103.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2022. Social influence dialogue systems: A survey of datasets and models for social influence tasks. *arXiv preprint arXiv:2210.05664*.

Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. 2023. Automatic evaluate dialogue appropriateness by using dialogue act. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7361–7372.

Ming-Bin Chen, Lea Frermann, and Jey Han Lau. 2024. Whow: A cross-domain approach for analysing conversation moderation. *arXiv preprint arXiv:2410.15551*.

Sintya Crisianita and Berlinda Mandasari. 2022. The use of small-group discussion to imrpove students'speaking skill. *Journal of English Language Teaching and Learning*, 3(1):61–66.

Roy G D'Andrade and Myron Wish. 1985. Speech act theory in quantitative research on interpersonal behavior. *Discourse Processes*, 8(2):229–259.

Mark R Freiermuth. 2001. Native speakers or non-native speakers: Who has the floor? online and face-to-face interaction in culturally mixed small groups. *Computer assisted language learning*, 14(2):169–199.

Nathalie Gagné and Susan Parks. 2013. Cooperative learning tasks in a grade 6 intensive esl class: Role of scaffolding. *Language teaching research*, 17(2):188–209.

Rena Gao, Carsten Roever, and Jey Han Lau. 2025. Interaction matters: An evaluation framework for interactive dialogue assessment on English second language conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10977–11012, Abu Dhabi, UAE. Association for Computational Linguistics.

Rena Gao and Menghan Wang. 2024. Listenership always matters: active listening ability in l2 business english paired speaking tasks. *International Review of Applied Linguistics in Language Teaching*, (0).

Rena Gao, Jingxuan Wu, Carsten Roever, Xuetong Wu, Jing Wu, Long Lv, and Jey Han Lau. 2024. Cnima: A universal evaluation framework and automated approach for assessing second language dialogues. *arXiv preprint arXiv:2408.16518*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Mohd Hilmi Hamzah and Thivya Asokan. 2016. The effect of participation instruction on esl students'

speaking skills and language anxiety. In *International Conference of Higher Order Thinking Skills 2016 in Conjunction with 2nd International Seminar on Science and Mathematics Education*, pages 1–11.

Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. *arXiv preprint arXiv:2210.05725*.

Gianna Hessel. 2017. A new take on individual differences in l2 proficiency gain during study abroad. *System*, 66:39–55.

Khe Foon Hew and Wing Sum Cheung. 2008. Attracting student participation in asynchronous online discussions: A case study of peer facilitation. *Computers & Education*, 51(3):1111–1124.

Yuki Higuchi, Makiko Nakamuro, Carsten Roever, Miyuki Sasaki, and Tomoko Yashima. 2023. Impact of studying abroad on language skill development: Regression discontinuity evidence from japanese university students. *Journal of the Japanese and International Economies*, 70:101284.

Bryce B Hudgins and Sybil Edelman. 1986. Teaching critical thinking skills to fourth and fifth graders through teacher-led small-group discussions. *The journal of educational research*, pages 333–342.

Jeremy F Jones. 1999. From silence to talk: Cross-cultural ideas onstudents participation in academic group discussion. *English for specific Purposes*, 18(3):243–259.

Hayriye Kayi-Aydar. 2013. Scaffolding language learning in an academic esl classroom. *ELT journal*, 67(3):324–335.

Jim King. 2013. Silence in the second language classrooms of japanese universities. *Applied linguistics*, 34(3):325–343.

Yuen Kwan Wendy Lam. 2006. Gauging the effects of esl oral communication strategy teaching: A multi-method approach. *Electronic Journal of Foreign Language Teaching*, 3(2):142–157.

Given Lee. 2009. Speaking up: Six korean students' oral participation in class discussions in us graduate seminars. *English for Specific Purposes*, 28(3):142–156.

Xiaoshi Li and Xuerui Jia. 2006. Why don't you speak up?: East asian students' participation patterns in american and chinese esl classrooms. *Intercultural Communication Studies*, 15(1):192.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Yun Long, Haifeng Luo, and Yu Zhang. 2024. Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, 9(1):60.

Nooreiny Maarof et al. 2018. The effect of role-play and simulation approach on enhancing esl oral communication skills. *International Journal of Research in English Education*, 3(3):63–71.

Greg McClure and Erika França de S Vasconcelos. 2011. From "i am" to "we could be": Creating dialogic learning communities in esol teacher education. *Pedagogies: An International Journal*, 6(2):104–122.

Wendy McKenzie and David Murphy. 2000. " i hope this goes somewhere": Evaluation of an online discussion group. *Australasian Journal of Educational Technology*, 16(3).

Alice Neusiedler. 2024. Engaging with the voice of the other through echoing: insights from participatory art. *Culture and Organization*, pages 1–18.

Tobias O. Nyumba, Kerrie Wilson, Christina J Derrick, and Nibedita Mukherjee. 2018. The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and evolution*, 9(1):20–32.

OpenAI. 2024. Openai api. OpenAI, https://openai.com/index/hello-gpt-4o/. Accessed: 2024-11-20.

Gihan Osman and Susan C Herring. 2007. Interaction, facilitation, and deep learning in cross-cultural chat: A case study. *The Internet and Higher Education*, 10(2):125–141.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. DialogBench: Evaluating LLMs as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170, Mexico City, Mexico. Association for Computational Linguistics.

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, pages 173–182.

Teresa Pica and Catherine Doughty. 1985. The role of group work in classroom second language acquisition. *Studies in second language acquisition*, 7(2):233–248.

Juan Pino-Silva and Carlos A Mayora. 2010. English teachers' moderating and participating in ocps. *System*, 38(2):262–271.

Paruralli Sprinivas Rao. 2019. Enhancing effective oral communication skills among the efl/esl learners. *Alford Council Of International English & Literature Journal*, 2:62–74.

Elizabeth Reddington. 2018. Managing participation in the adult esl classroom: Engagement and exit practices. *Classroom Discourse*, 9(2):132–149.

Vasile Rus, Cristian Moldovan, Nobal Niraula, and Arthur C Graesser. 2012. Automated discovery of speech act categories in educational games. *International Educational Data Mining Society*.

Richard J Sampson. 2024. The emergence of gratitude in l2 group discussion: A small-lens study. *System*, 123:103300.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*.

Richard K Tan, Ronald B Polong, Leila M Collates, and Joel M Torres. 2020. Influence of small group discussion on the english oral communication self-efficacy of filipino esl learners in central luzon. *TESOL International Journal*, 15(1):100–106.

Blake Alexander Turnbull. 2015. *The effects of L1 and L2 group discussions on L2 reading comprehension*. Ph.D. thesis, University of Otago.

Vinothini Vasodavan, Dorothy DeWitt, Norlidah Alias, and Mariani Md Noh. 2020. E-moderation skills in discussion forums: Patterns of online interactions for knowledge construction. *Pertanika Journal of Social Sciences and Humanities*, 28(4):3025–3045.

Andrew R Webster. 2012. *Teaching EFL Online: An e-moderator's report*. BoD–Books on Demand.

Jingxuan Wu and Carsten Roever. 2025. Data from role plays and elicited conversations: What do they show about l2 interactional competence? *Research Methods in Applied Linguistics*, 4(1):100165.

Luxin Yang. 2010. Doing a group presentation: Negotiations and challenges experienced by five chinese esl students of commerce at a canadian university. *Language Teaching Research*, 14(2):141–160.

Wei Zhu. 2001. Interaction and feedback in mixed peer response groups. *Journal of second language writing*, 10(4):251–276.

A Score definition of dialogue interactivity quality evaluation

Interactivity Macro-level Features	**Definition**
Topic Management	the strategies and techniques used to control and navigate the flow of topics
Tone Choice Appropriateness	the suitability of the tone used in communication, ensuring it aligns with the context, audience, and purpose to convey the intended message
Conversation Opening	the initial interaction or exchange that begins a dialogue, often setting the tone and context for the dialogue
Conversation Closing	the process of ending a dialogue or interaction, which involves signaling the conclusion of the discussion, summarizing key points, and often expressing a farewell

Table 10: Definitions of macro-level interactivity features, with higher score emphasising on natural, authentic interaction and active engagement in the dialogue

Interactivity Labels	**Scores**	**Description of Scores**
Topic Management	[5]	topic extension with clear new context
	[4]	topic extension under the previous direction
	[3]	topic extension with the same content
	[2]	repeat and no topic extension
	[1]	no topic extension and stop the topic at this point
Tone Appropriateness	[5]	very informal
	[4]	quite informal, but some expressions are still formal
	[3]	relatively not formal, and most expressions are quite informal
	[2]	quite formal, and some expressions are not that formal
	[1]	very formal
Conversation Opening	[5]	nice greeting and showing a good understanding of the opening of conversation in social interactions.
	[4]	sounded greeting and showed a basic understanding of the social role.
	[3]	general greeting but not understanding the social role well.
	[2]	basic greeting.
	[1]	no opening, start the discussion immediately.
Conversation Closing	[5]	detailed summarization and smooth transition to the closing of the conversation.
	[4]	transit to the closing naturally, but without summarising the discussion.
	[3]	transit to the discussion.
	[2]	demonstrate a translation to the end of the conversation.
	[1]	no closing, directly stop the conversation.

Table 11: Description of scores for dialogue-level interactivity labels. Higher score indicates better interactivity ability, for example, *Tone Appropriateness* scores higher with more informality shows that the speakers are able to employ more active linguistics resources in dialogue communication to perform more causal and natural interactions compared with the formal tone, which has limited linguistics resources and not naturally occurred in real-life conversations.

B Overall Dialogue Quality Score Description and Definitions

The following Table 12 shows the descriptions and definitions for dialogue's overall quality score.

Scores	Descriptions
5	Smooth and fluent daily communication, easy and pleasant through the whole chat
4	Somewhat less fluent communication, but the communication purpose is achieved
3	Slightly awkward communication in some places, such as not being able to understand the other person's question
2	Overall communication is not fluent and mostly awkward, but some parts can be mutually understood
1	Unable to accurately achieve the communication purpose, awkward conversation, and failed to talk throughout the conversation.

Table 12: Score description for overall dialogue quality

Prompts for GPT-4o Dialogue Overall Evaluation The following Table 11 shows the prompts for dialogue overall quality score with GPT-4o.

Field	Description
Conversation	A dialogue of second language Chinese conversation.
Output Fields	**score**: The score of the interactivity of the English second language dialogue (1 to 5). **rationale**: The reason why and how the score is made based on each participant's utterance.
Evaluation Criteria	**5**: Smooth and fluent daily communication, easy and pleasant. **4**: Somewhat less fluent communication, but the communication purpose is achieved. **3**: Slightly awkward communication, such as not being able to immediately understand the other person's question with hesitation. **2**: Overall communication is not fluent and awkward, but some parts can be mutually understood. **1**: Unable to accurately achieve the communication purpose, awkward conversation, failed to talk throughout the conversation.
Human Validation	Authors from this study verified the LLM results manually and pass the 75% judgement compared with human evaluation.

Table 13: LLM Dialogue Overall Dialogue Quality Evaluation Prompts and Human Validation Process

C WHoW Analytic Framework

Dimension	Label	Definition
Motives	Informational (IM)	Provide or acquire relevant information to constructively advance the topic or goal of the conversation.
	Coordinative (CM)	Ensure adherence to rules, plans, and broader contextual constraints, such as time and environment.
	Social (SM)	Enhance the social atmosphere and connections among participants by addressing feelings, emotions, and interpersonal dynamics within the group.
Dialogue acts	Probing (prob)	Prompt speaker for responses.
	Confronting (conf)	Prompt one speaker to response or engage with another speaker's statement, question or opinion.
	Instruction (inst)	Explicitly command, influence, halt, or shape the immediate behavior of the recipients.
	Interpretation (inte)	Clarify, reframe, summarize, paraphrase, or make connection to earlier conversation content.
	Supplement (supp)	Enrich the conversation by supplementing details or information without immediately changing the target speaker's behavior.
	Utility (util)	All other unspecified acts.
Target speaker	Target speaker (TS)	The group or person addressed by the moderator.

Table 14: Definitions and acronyms for the labels across the three dimensions: motives (Why), dialogue acts (How), and target speakers (Who). Target Speaker is a categorical variable with values corresponding to each participant in the dialogue, plus "audience", "self", "everyone", "support side", "against side", "all speakers", and "unknkown".

DAs	IM	CM	SM
Prob	Can you take that on? (prompting) As long as the political spectrum is covered overall, what's wrong with that? (follow up question) Siva? (name calling prompt)	Which of you would like to go first? (preference inquiry) Did this gentleman come down yet? (coordinative question) It's working, right? (question managing environment)	Is that a relief to you or– (asking feeling) Could you tell us your name, please? (social question) Do you have eyeglasses? (humour question)
Conf	That landed pretty well I think, so can you respond to that? (counter confronting) On this side, do you want to respond, or do you agree? (consensus confronting) You actually asked a perfect question, and so Mark Zandi, do you want to take that on? (confronting question)	The other side care to respond, if not I'll move on.(coordinative consensus) Response from the other side, or do you want to pass? (coordinative confronting) Marc Thiessen, do you want to join your partner on this one, because I think– (coordinative consensus)	Bryan Caplan, I think he just described your fantasy, come true.(social confronting) I'd love to hear your answer to that question, so go for it. (confronting with affective appeal) Jared Bernstein, the guy you called "nuts" just said you're unfair. (humour confronting)
Inst	Can you frame your question as a question? (articulate instruction) Relate that point to this motion. (back to topic) I want to stay on the merits of the Obama plan. (manage topic)	Remember, about 30 seconds is what you'll get. (time control) Can you go up three steps, please, and turn right? (coordinating instruction) I'll be right back after this message. (program management)	Do not be afraid. (emotion instruction) Those who agree, just a round of applause to that. (pro-social instruction) –because it's turning into a personal attack. (stop anti-social)
Inte	So, Matt, you're saying that it's not true that it's inevitable that Amazon will control everything. (summarization) Their point is that it would be a bad thing. (simplification) But that would be the question of mobility. (reframe)	That was an ambiguous signal. (situation interpretation) You're pointing to Lawrence Korb.(preference interpretation) And you want the side arguing for the motion to address that (preference interpretation)	I think it was a rhetorical question, and it got a good laugh. (humour interpretation) And it's a little bit insulting almost to say (toxicity interpretation) —honestly, I don't think that was an—a personal attack— (toxicity interpretation)
Supp	I agree that it is.(agreement) The fact is that one of the US manufacturers, with 1 percent of its yearly production, would run us out of the whole market.(add information) They had never paid any attention whatsoever to Africa. (share opinion)	Fifty-one of you voted against the motion. (vote reporting) And the mic's coming down to you. (describe situation) Round two is where the debaters address each other directly (rule explanation)	You have a colorful sleeve. (social chit-chat) I hate to reward it but I'm going to. (encouragement) And I think all of us probably share a sense that we want things to improve. (state common feeling)
Util	Fair question. (acknowledgement) Right (acknowledgement) So the– (floor grabbing)	All right. (backchanneling) Actually, I– (floor grabbing) Well—(floor grabbing)	Thank you Evgeny Morozov. (thanks) I'm sorry. (apology) Hi. (greeting)

Table 15: This table presents a collection of exemplar sentences from the original paper at the intersection of the motives and dialogue acts dimensions.

section	prompt part
Role & topic	Your role is an annotator, annotating the moderation behavior and speech of a debate TV show. The debate topic is “When It Comes To Politics, The Internet Is Closing Our Minds"
Task instruction	given the definition and the examples, the context of prior and posterior dialogue, please label if the target utterance carries informational motive?
Dimension instruction	Motives: Motives are the high level motivation that the moderator aim to achieve. The definitions and examples of the informational motive are below:
Label definition	informational motive: Provide or acquire relevant information to constructively advance the topic or goal of the conversation.
Label examples	examples: “Why do you think minimum wage is unfair?” (Relevant information seeking.) “The legal system has many loopholes.” (Expressing opinion.) “Yea! I agree with your point!” (Agreement relevant to the topic.) “The law was established in 1998.” (Providing topic relevant information.)
Dialogue prior context	Dialogue context before the target sentence: (including dialogue up to 5 utterance prior) Eli Pariser (for): Right, and the question is, can you trust them? John Donvan (mod): Let me– Jacob, I think Eli left a pretty good image hanging out there, of these folks truly not knowing how much they don’t know and believing what they’re getting and not understanding how slanted it is.
Target sentence	Target sentence: John Donvan (mod): That landed pretty well I think, so can you respond to that?
Dialogue post context	Dialogue context after the target sentence: Jacob Weisberg (against): But a guy who called into a radio show? I know the plural of anecdote is data.....(more) John Donvan (mod): Siva. (including dialogue up to 2 utterance after the target.)
Formatting instruction	Please answer only for the target sentence with the JSON format:{"motives": List(None or more from "informational motive", "social motive", "coordinative motive"),"dialogue act": String(one option from "Probing", "Confronting", "Supplement", "Interpretation", "Instruction", "All Utility"),"target speaker(s)": String(one option from "0 (Unknown)", "1 (Self)", "2 (Everyone)", "3 (Audience)", "4 (Eli Pariser- for)", "5 (Siva Vaidhyanathan- for)", "6 (Evgeny Morozov- against)", "7 (Jacob Weisberg- against)", "8 (Support team)", "9 (Against team)", "10 (All speakers)"),"reason": String} For example: answer: {"motive": ["informational motive"], "dialogue act": "Probing", "target speaker(s)": "7 (Joe Smith- for)", "reason": "The moderator asks a question to Joe Smith aimed at eliciting his viewpoint or reaction to a statement from the recent policy change for combatting climate change....."}

Table 16: A simplified WHoW framework prompt for annotating a target sentence regarding the motives, dialogue acts, and the target speaker. The ’Dimension Instruction’ section is repeated for the ’Dialogue Act’ dimension, while ’Label Definition’ and ’Label Examples’ are repeated for each label under the motives and dialogue acts dimensions.

D ESLMOD moderation strategies

Category	Examples
Information Probing	"Anyone else who wants to share their thoughts or opinions about what the purpose of a relationship is for them?" "Related to stress. And how do you manage in such situations?" "Yeah. How about you, Chantelle?" "Do you agree with this statement?"
Opinion Sharing	"For me, managing stress is all about maintaining a good work-life balance." "To me, sharing housework equally is a sign of respect and partnership in a relationship." "I believe that having diverse perspectives in a team leads to more creative solutions."
Information Sharing	"Today's topic is related to the recent trends in the job market." "Research shows that group discussions can improve second language acquisition by increasing practice opportunities." "You can find free online courses on platforms like Coursera or edX to learn new skills."
Echoing	"Yeah, my friend had a similar experience when he was in the US, as he struggled to find a job." "I can relate to that feeling of being overwhelmed when learning a new language. I have been through it too." "I completely agree—my friend also struggled with finding a balance between work and family responsibilities."
Experience Sharing	"There was a time when I had to make a tough decision about changing my career path—it was such a challenging moment for me." "Once, during my university days, I stayed up all night preparing for a group project because I wanted everything to be perfect."
Acknowledgement	"That is a very interesting insight." "Great point, and I think it really ties back to what we were discussing earlier." "I appreciate you bringing this up—it's a really valuable perspective." "Thanks for sharing that example—it really helped clarify the idea."
Backchanneling	"Yeah." "Hmm." "Okay." "Uh-huh." "Right." "I see." "Mhm."
Social Utility	"Goodbye!" "Thank you!" "Please, go ahead!" "Excuse me." "I appreciate your time."
Informational Interpretation	"If I understand correctly, you're suggesting that online courses are beneficial because they provide flexibility." "To summarize, the main takeaway here is that building relationships in the workplace helps reduce stress." "In other words, you're arguing that peer feedback plays a critical role in language learning success."
Coordinative Instruction	"Can we wrap up this discussion and move on to the next point?" "I'd like everyone to think about this question and share your thoughts one by one." "Now everyone is here, let's start the session." "Please turn off your microphone when you are not speaking."

Table 17: ESLMOD strategies categories and examples.

section	prompt part
Role & topic	Your role is an annotator, annotating the moderation behavior of a second language speakers" English conversation session. The topic is "Are you superstitious? Why do people believe in 'weird' things?"
Task instruction	given the definition and the examples, the context of prior and posterior dialogue, please label which dialogue act the target sentence belong to? And who is the moderator talking to?
Dimension instruction	Dialogue act: Dialogue acts is referring to the function of a piece of a speech/sentence. The definitions and examples of the dialogue acts are below:
Label definition	Information Probing: Prompting participants to share their thoughts, opinions, or experiences by posing questions or directly inviting input from individuals or the group.
Label examples	examples: "Anyone else who want to share their thought or opinion about why they what is the purpose of relationship for them?" "Related to stress. And how do you manage in such situation?" "Yeah. How about you, Chantelle?" "Anyone else who want to share their thought or opinion about why they what is the purpose of relationship for them?" "Do you agree with this statement?"
Dialogue prior context	Dialogue context before the target sentence: (including dialogue up to 5 utterance prior) Andrew (participant): And why should I contribute a huge amount of effort in my life to these gods, to me they don't exist, or maybe they do. Andrew (participant): It's just that they gotta find a way to manifest themselves at least. Andrew (participant): But I do believe in luck in and also, I believe in all the coincidences in life, like, Yeah. Bryan(host) (moderator): But I think it's not probably not just about religions.
Target sentence	Target sentence: Bryan (moderator): But sometimes they people just have this kind of like superstitution, like, you know if today, by wearing this color, or I have my watch, which is, I don't know, and it might probably bring luck to me.
Dialogue post context	Dialogue context after the target sentence: Christine (participant): Yeah, I think, like, actually, I think, like, everyone, have some sort of religious or cultural specifications like at least influenced by one.....(more) (including dialogue up to 2 utterance after the target.)
Formatting instruction	Please answer only for the target sentence with the JSON format:"dialogue act": String(one option from 0 (Information Probing), 1 (Opinion Sharing), 2 (Information Sharing), 3 (Echoing), 4 (Experience Sharing), 5 (Acknowledgement), 6 (Backchanneling), 7 (Social Utility), 8 (Information Interpretation), 9 (Coordination Instruction)),"target speaker(s)": String(one option from "0 (Unknown)", "1 (Everyone)", "2 (Emma)", "3 (Andrew)", "4 (Christine)", "5 (Jodie)", "6 (Leo)", "7 (Yuki)", "8 (Yale)"),"reason": String For example: answer: {"dialogue act": "1 (Information Probing)", "target speaker(s)": "3 (Joe Smith)", "reason": "The moderator asks a question to Joe Smith aimed at eliciting his viewpoint or reaction to a statement from the recent policy change for combatting climate change....."}

Table 18: An example of ESLMOD prompt updated from the WHoW prompt incorporating the identified moderation strategies for ESL group discussions. The 'Label Definition' and 'Label Examples' subsections are repeated for each newly identified moderation strategy.

E Domain Adaption Using Topic Modeling with WHoW

motive / dialogue act	hyper-parameters	coherence score	topic models words	cluster size	decision	refined lable
Informational probing	kmean_n_clusters=3 umap_n_neighbor=5, umap_min_dists=0.0	0.729	question, asking, ask, aim, elicit	574	merge	Informational probing
			ask, address, question, attempting, information	221		
			asking, information, perspective, question, seek	70		
Informational supplement	kmean_n_clusters=3, umap_n_neighbor=10, umap_min_dists=0.5	0.453	additional, opinion, personal, perspective, insight	823	expand	Opinion sharing
			information, additional, provides, response, supplement	292	expand	Information sharing
			add, build, reflect, reinforce, agree	82	expand	Echoing
Social supplement	kmean_n_clusters=4, umap_n_neighbor=30, umap_min_dists=0.1	0.474	personal, shares, experience, atmosphere, social	213	expand	Experience Sharing
			expresses, positive, appreciating, agreement, statement	52	merge	Acknowledgement
			acknowledges, point, gratitude, previous, statement	74		
			respond, compliment, reflect, serve, statement	50	merge to Echoing	Echoing
social utility	kmean_n_clusters=2, umap_n_neighbor=10, umap_min_dists=0.0	0.453	expresses, acknowledges, gratitude, uses, bye	73	expand	Social utility
			serves, acts, serving, backchannelling, respond	249	extend	Backchanneling

Table 19: This table presents the optimized hyper-parameters for BERTopic applied to the four prominent WHoW intersected labels. It details the k value used for KMeans clustering, as well as the number of neighbors and the minimum distance parameters for UMAP. Additionally, the table reports the coherence score for each cluster and lists the top five keywords for the sub-topic clusters. The final two columns indicate whether each cluster was merged or expanded, along with the manually refined names for the new sub-topics.

E.1 Motivation and Advantage of Using WHoW as a Foundational Framework

The WHoW analysis provides a high-level characterization of the moderator's role in specific conversational scenarios. For example, a debate moderator tends to prioritize information dissemination and coordination, often adopting a stronger functional role in managing conflicts and providing instructions. While these general insights allow for comparisons across scenarios, they may lack the granularity necessary to guide specific strategies or actions. For instance, identifying a moderator as high in information supplement indicates that information is being shared but does not specify the type of information (e.g., opinions or experiences).

Nevertheless, the WHoW analysis serves as a foundational framework—a "skeleton"—that can be refined into a more detailed set of strategies tailored to specific domains. Our literature review indicates that dialogue act schemas developed for various domains often include fine-grained categories for domain-specific acts, while relying on coarser categories for more general or less frequent acts. For example, in e-rulemaking discussions (Park et al., 2012), probing interventions are subdivided into three types: prompting users to provide additional information, encouraging them to propose or consider solutions, and posing open-ended questions. Similarly, information supplements are categorized into providing details about proposed rules, pointing to relevant resources, and identifying characteristics of effective commenting. However, only a single type of intervention for information interpretation is included: correcting misstatements. Conversely, in e-learning scenarios (Vasodavan et al., 2020), the focus shifts, with three distinct types of interpretive interventions: summarizing discussions, highlighting contributions, and archiving information. The WHoW analysis identifies broad, high-frequency categories that can serve as a starting point for further refinement into fine-grained, domain-specific dialogue acts.

A key advantage of using WHoW as the foundational framework for domain adaptation is that it allows for easier comparison of moderator behavior across different domains. For example, the schemas developed in the two earlier studies are not directly comparable because they use different label sets. However, if the labels were based on the WHoW framework, it would be possible to measure the similarity between moderators' functions and motivations in various scenarios. This kind of cross-domain comparison can help reveal patterns that are consistent across contexts, offering useful insights into effective moderation practices and potentially guiding improvements in moderation strategies across different settings.

E.2 Dialogue Act Domain Adaption For ESL Discussion Moderation

E.2.1 Labels Selection

Building on the WHoW analysis applied to the ESL moderated conversation datasets, we excluded 12 of the 18 intersected motive/dialogue categories that accounted for less than 2.5% of all instances. Among the remaining combinations, four prominent categories emerged—informational probing, informational supplement, social supplement, and social utility—each representing more than 10% of instances. These labels then served as the basis for further exploration and potential domain-specific expansion. Two categories that fell between 2.5% and 10% were directly included in the final list without additional exploration.

It is important to note that the selected thresholds were tailored for the current study. Categories representing less than 2.5% (approximately 75 samples) were deemed too insignificant, while those with fewer than 10% (around 300 samples) were considered insufficient for robust topic modeling. These thresholds can be adjusted based on the specific use case and overall sample size.

E.2.2 Pre-processing annotation reasons

For each prominent label, we extracted the associated moderator sentences and applied topic modeling to explore potential specifications. To ensure that the topic modeling results reflected the moderator's intent and actions rather than merely the discussion topics, we leveraged the "reasons" generated by GPT-4o during the WHoW analysis. For example, GPT-4o generated the following reason:

> "The moderator is providing a contextual or explanatory statement intended to set the scene for the discussion about societal and cultural beliefs in things beyond rational or logical understanding, which is shared information intended for all participants."

Typically, the first sentence of a reason summarizes the core intent and action of the modera-

tion, while subsequent sentences elaborate on motives, dialogue acts, and target speakers. To make the topic modeling results more sensitive to the moderator's intent, we applied the following preprocessing steps:

- Extract the first sentence from each reason generated by GPT-4o during the WHoW analysis.
- Process the sentence with SpaCy's dependency parser to identify the root verb and its direct object subtree. For instance, the sentence "The moderator shares a personal anecdote about his experience working in an AI company to contribute work-related insights to the topic 'The impact of AI'" is reduced to "shares a personal anecdote about his experience working."
- Curate an additional list of stop words, including speaker names (e.g., "Amy") and keywords specific to the discussion topic, such as "AI" and "stress."

E.2.3 Topic modeling for identifying domain specific dialogue acts

We applied BERTopic (Grootendorst, 2022), a widely used neural topic modeling approach that leverages pre-trained BERT embeddings, combined with K-means clustering. We optimized the number of clusters (k) within a range of 2 to 5 to identify potential sub-categories. We selected K-means over DBScan to maintain control over the number of clusters. Additionally, we employed U-MAP for dimensionality reduction of the sentence representation vectors, reducing the impact of small sample sizes and sparse vectors on the clustering process. For each prominent intersecting label, we ultimately identified the optimized hyper-parameter sets, as summarized in Table 19.

Decisions on whether to expand or merge subtopics were based on the following criteria:

- Do the sub-topic's top five representative keywords distinctly differ from those of other acts or WHoW dialogue acts?
- Are the original reasons associated with the sub-topic's samples clearly distinguishable from those of other acts or WHoW dialogue acts?
- Do the sub-topic's keywords form a coherent theme?
- Are the moderator sentences in the sub-topic notably different from those in other acts or WHoW dialogue acts?

Finally, for each finalized sub-topic cluster, we manually assigned a label (e.g., "Echoing") and a definition, iteratively refining them with GPT-4o using the cluster samples to accurately capture the content. When the output topics largely aligned with the original WHoW motive and dialogue act intersected labels, those original labels were retained. Additionally, we incorporated two intersected labels that appeared in more than 2.5% but less than 10% of instances, ensuring their relevance was not overlooked. Ultimately, this process yielded 10 classes of moderation strategies: four classes derived from the original WHoW intersected labels and six newly defined, domain-specific strategies.

F Conversation club topics and material

Components	Content
Topic question	How AI impacts our daily life now and the future? 人工智能将会如何影响我们现在与未来的生活?
Description	In this discussion, we will explore the profound impacts of artificial intelligence on our daily lives, examining both current applications and future possibilities. As AI technologies advance, they integrate more seamlessly into various sectors such as healthcare, finance, education, and personal productivity, altering how we work, learn, and interact. This session aims to dissect the benefits and challenges of AI integration, and predict how it might shape our society in the coming decades. 在这次讨论中，我们将探讨人工智能对我们日常生活的深远影响，审视其当前的应用与未来的可能性。随着 AI 技术的进步，它正日益无缝融入医疗、金融、教育和个人生产力等各个领域，改变我们的工作、学习和互动方式。本次会议旨在剖析 AI 带来的机遇与挑战，并预测其在未来几十年内将如何塑造我们的社会和个人生活。
Questions	1. What are some of the most significant changes AI has brought to our personal and professional lives today? How do these changes enhance or complicate our daily activities? 人工智慧今天为我们的个人和职业生活带来了哪些重大变化？这些变化是如何影响我们的日常生活? 2. AI is poised to automate many jobs that currently require human labor. Do you feel being threaten? What strategies should us individuals and the society implements to adapt? 人工智慧即将自动化许多目前需要人力的工作。你感到受到威胁了吗？我们个人和社会应该实施哪些策略来适应? 3. How do you think AI might affect human relationships and social interactions in the future? Will it bring people closer or create more distance? 将来人工智慧会如何影响人际关和社交互动？它会让人们更亲近还是造成更多距离? 4. In comparison to AI, what are the things human possess that are irreplaceable? 与人工智慧相比，人类拥有哪些不可替代的特质?

Table 20: An exemplar discussion material featuring the topic "How AI impacts our daily life now and the future?" ncluding the main topic question, background description, and discussion prompts.

Session stage	Instruction
Introduction(3 ms)	During the discussion stage, the moderator will briefly introduce the topic using the provided material and give a short self-introduction.
Discussion(45 ms)	During the discussion stage, the moderator will guide the conversation using the provided questions, encouraging participants to share their thoughts and insights. If the discussion slows down, the moderator may contribute or introduce new prompts to stimulate engagement.
Conclusion(3 ms)	At the conclusion stage, the moderator will summarize the key points raised by participants for each question and provide a final thought. Additionally, the moderator will invite participants to share any final remarks. Finally, the session will be wrapped up with a closing statement and a farewell.

Table 21: The instruction for the moderator at different stages of the discussion session.

G Participation consent form

Plain Language Statement

School of XXXXXXXXXXXXX/Faculty of XXX

Project: Development of Interactional Competence: Prediction and Quantification on L2 English Dialogue

Responsible Researcher:
Prof XXXXXXX XXXXXX Tel: (XX) XXXX XXXX Email: XXXXXX@XXXXXXX;

Additional Researchers:
Prof XXXXXXX XXXXXX Tel: (XX) XXXX XXXX Email: XXXXXX@XXXXXXX;

Student Researcher:
XXXXXXX XXXXXX Tel: (XX) XXXX XXXX Email: XXXXXX@XXXXXXX;

XXXXXXX XXXXXX Tel: (XX) XXXX XXXX Email: XXXXXX@XXXXXXX;

Introduction

Thank you for your interest in participating in this research project. The following few pages will provide you with further information about the project, so that you can decide if you would like to take part in this research. Please take the time to read this information carefully. You may ask questions about anything you don't understand or want to know more about.Your participation is voluntary. If you don't wish to take part, you don't have to. If you begin participating, you can also stop at any time.

What is this research about?

The research is to investigate the discussion and participation patterns of ESL (English as Second Language) speaker in English group discussion. In particulat, we are interested in the benefit and effect of the presence of a moderator during the conversation.

What will I be asked to do?

If you agree to participate, you will be asked to provide language samples through group discussion sessions. These sessions may be conducted in a free-style discussion format or be moderated. If applicable, your voice may be recorded. You will be assigned to a group with participants who share similar non-native language backgrounds.

You will take part in one or more discussion sessions, depending on your availability and interest. Each session will include 3 to 5 other participants and will consist of the following:

1. **A themed topic** (e.g., *Modern Romance*).
2. **A set of 3–5 guiding questions** (e.g., *How do dating practices and expectations differ between generations, and what are the main factors driving these differences?*).
3. **A discussion format** that is either **free-flowing** or **moderated**.
4. **A duration of 40–60 minutes** per session.

Additional Considerations

- The discussion will be **recorded**, and all data will be **anonymized** before being used for research purposes.
- Some topics may involve **personal or emotional content**. Please consider whether the information you share contains **private details** or could impact your personal life.
- Our research focuses on **language expression and interaction patterns** rather than the authenticity of shared content. You are welcome to remain **anonymous** or modify details when discussing sensitive personal experiences.
- This study primarily involves **text and audio content**, so **turning on your camera is not required**.

What are the possible benefits?

This study on conversational moderation in ESL group discussions offers several potential benefits. It can enhance participants' English proficiency by improving their speaking, listening, and interaction skills in structured discussions. Effective moderation techniques may foster more inclusive and engaging conversations, reducing anxiety and encouraging balanced participation. The findings can contribute to language learning research by identifying best practices for facilitating discussions, which may inform ESL teaching methods and AI-driven moderation tools. Insights from this research may also benefit online learning platforms and multilingual workplaces by optimizing discussion facilitation strategies.

What are the possible risks?

There are few risks associated with this project and these risks are minimal. Potential risks might include research participants finding responding to research tasks difficult or tiring, or feeling shy or anxious about their performance, or worrying about being identified in later reports of the research. Thesis researchers will ensure that participant performance will not affect any of their records beyond the thesis project, such as school records, professional records, etc. Participant data will be used for research purposes only and raw data will only be available to researchers. Participants are assured that their identities will be protected by using pseudonyms (fake names) and removing identifying information.

Do I have to take part?

No. Participation is completely voluntary. You are able to withdraw at any time. You can also withdraw any data that has not been analysed. Participants who withdraw without completing data collection may not receive compensation

Will I hear about the results of this project?

The researcher will provide a summary of finding, or a copy of the completed thesis/paper to participants upon request.

What will happen to information about me?

Data are stored as digital files, e.g., spreadsheets, documents, or audio/video files. It is stored on the student researcher's computer and the University's cloud storage service in password protected files. When analysing the data, participants' names are changed to pseudonyms (fake names) and identifying information is removed, making participants not to be identified. Only the student researcher and the supervisor researcher have access to potentially identifiable data. Data reported in theses and possible later publications is not identifiable. Data is kept by the student researcher for five years after thesis submission or the last publication resulting from the thesis.Where can I get further information?

If you would like more information about the project, please contact the researchers; xxx xxxxx xxxxx@xxxxxx; xxxxx xxxxx xxxxx@xxxxx or the student researcher in this study: xxx xxx, xxxxxx@xxxxx.

Who can I contact if I have any concerns about the project?

This project has human research ethics approval from The University of XXXXX. If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the XXXXXXXXXX, Office of Research Ethics and Integrity, University of XXXXX, XXXXX. Tel: +XX XXX XXX XXX or Email: XXXXXX@XXXXX All complaints will be treated confidentially. In any correspondence please provide the name of the research team and/or the name or ethics ID number of the research project.Project ID Number: XXXXXX