CNIMA: A Universal Evaluation Framework and Automated Approach for Assessing Second Language Dialogues

Anonymous ACL submission

Abstract

We develop CNIMA (**C**hinese **N**on-Native **I**nteractivity **M**easurement and **A**utomation), a Chinese-as-a-second-language labelled dataset with 10K dialogues. We annotate CNIMA using an evaluation framework — originally introduced for English-as-a-second-language dialogues — that assesses micro-level features (e.g. backchannels) and macro-level interactivity labels (e.g. topic management) and test the framework's transferability from English to Chinese. We found the framework robust across languages and revealed universal and language-specific relationships between micro-level and macro-level features. Next, we propose an approach to automate the evaluation and find strong performance, creating a new tool for automated second language assessment. Our system can be adapted to other languages easily as it uses large language models and, as such, does not require large-scale annotated training data.

1 Introduction

In the context of second language (SL) assessment, speaking has always been considered an essential ability for SL speakers (Smith et al., 2022; Allwood, 2008; Huang et al., 2020), but prior studies have predominantly focused on written language proficiency (Paiva et al., 2022). Given the lack of datasets that capture the unique linguistic features of SL speakers in dialogues (especially in open-domain conversations), this leaves conversational interaction assessment under-explored and ultimately contribute to a limited understanding of conversational fluency and interactivity for SL speakers.

An exception is Gao et al. (2024), who propose a two-level framework for assessing the interactivity ability of English-as-a-second-language (ESL) speakers in open-domain conversations. They introduce micro-level word/utterance features (e.g. backchannels) and macro-level interactivity labels (e.g. topic management) and found that the micro-level features are highly predictive of the macro-level labels. Gao et al. (2024) reveal their intricate relationships, showing that certain micro-level features are highly correlated with a particular macro-level feature, e.g. frequent use of appropriate collocations and varied sentence structures (micro-level features) significantly correlate with topic management and tone appropriateness (macro-level labels). However, this research is conducted only in the context of ESL, raising questions about the transferability of the evaluation framework to other languages besides English. Furthermore, they did not propose an automated pipeline for assessing SL dialogue quality, as their predictive models rely on human-annotated micro-level features.

Our work aims to address these shortcomings by (1) developing and testing a Chinese-as-a-second-language (CSL) dialogue dataset using the evaluation framework of Gao et al. (2024); (2) introducing a fully automated pipeline to assess interactivity for CSL conversations. We experiment with different machine learning models, from classical machine learning models (e.g. logistic regression) to large language models (LLMs; e.g. GPT-4o (OpenAI, 2024)), to automate the prediction of micro/macro-level features and overall dialogue quality scores. Our best system shows strong predictive performance, paving the way for a new tool for assessing SL conversations. To summarise:

- We release CNIMA (**C**hinese **N**on-Native **I**nteractivity **M**easurement and **A**utomation), an annotated CSL dialogue dataset with 10K dialogues, based on the framework of Gao et al. (2024) originally designed for ESL.
- We test the cross-lingual transferability of the evaluation framework for CSL and find that it is robust across languages, and we further reveal universal and language-specific relation-

ships between the micro-level and macro-level features.

- We introduce a fully automated approach that predicts micro-level features, macro-level labels and an overall quality score given an input dialogue by exploring a spectrum of models from classical machine learning to LLMs. Our best system demonstrates strong performance, creating a new automatic tool for second language assessment. Importantly, our LLM-based system can be adapted to other languages without requiring annotated data.

2 Related work

2.1 Assessment of SL Spoken Conversation

Mainstream second language assessments in current industrial practice have mainly focused on grammatical accuracy, pronunciation standardization and vocabulary richness; for example, in TOEFL iBT, PTE Academic and Cambridge IELTS test (Xu, 2018; Paiva et al., 2022; Xu et al., 2021). However, few speaking assessments emphasised the importance of interaction (Khabbazbashi et al., 2021) and aspects such as how speakers manage the topics in communication (Shaxobiddin, 2024), how speakers perform social roles from speaking (Chen et al., 2023), and how speakers start and end a talk in an acceptable manner (Yap and Sahoo, 2024). There are, however, some exceptions. For example, Dai (2022) develop a test rubric for Chinese SL speakers. More recently, Gao and Wang (2024) introduce an interactivity scoring framework inspired by the IELTS speaking assessment. However, these studies only provide a theoretical framework for evaluation and lack an automated pipeline for large-scale assessments.

2.2 Automated Scoring System in SL

Automated scoring systems can improve the efficiency of processing and scoring dialogues compared to manual assessment methods, allowing for real-time feedback (Evanini et al., 2017). Some major second language assessment organisations have employed automated speaking scoring, including PTE Pearson (Jones and Liu, 2023), Duolingo English Test (Burstein et al., 2021), and TOEFL (Gong, 2023), but these automated systems fail to capture the interactive nature of SL conversations. That is, they struggle to offer detailed analyses and insights on the common errors and language usage patterns and where SL speakers struggle the most. This motivates a more explainable framework for SL assessment that can provide better feedback.

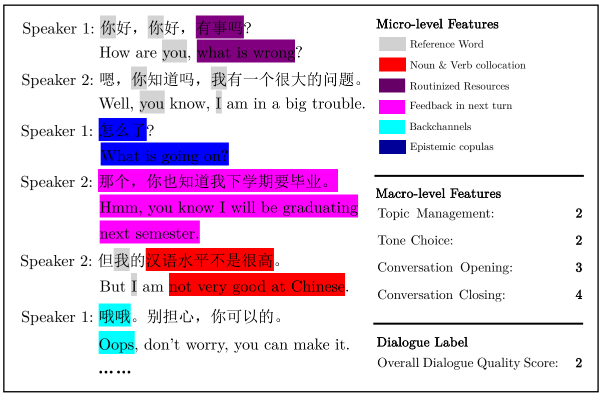
The main challenge is in designing an effective evaluation framework. Capturing the nuances of spoken language, such as interaction, attitudes, and cooperation in conversation, is complex to achieve (Gao and Wang, 2024). Automated scoring systems must handle diverse patterns, which can vary widely among different non-native SL speakers. Recognizing and assessing interactive features like turn-taking, interruptions, and response appropriateness further complicates the process. Ultimately, the dynamic nature of conversations makes it challenging for automated systems to accurately evaluate interactive speaking, necessitating sophisticated algorithms and continuous refinement (Cumbal, 2024; Engwall et al., 2022).

3 Evaluation Framework

We adapt the dialogue evaluation framework of Gao et al. (Gao et al., 2024) for assessing SL dialogues, which considers micro-level features and macro-level interactivity labels. One addition we have made is to introduce an overall dialogue quality score. Figure 1 shows an example of CSL dialogue annotation with the evaluation framework.

The framework has 17 micro-level features, and it can be further broken into token-level (such as 'Reference Word': *she* and 'Backchannel': *hmm*) and utterance-level features (such as 'Formulaic response': *How's going*). The micro-level features are annotated as spans (i.e. annotators mark text spans corresponding to a feature). For macro-level interactivity labels, there are four: Topic Management, Tone Choice Appropriateness, Conversation Opening and Conversation Closing, and they are annotated as (dialogue) labels. Briefly, topic management refers to the speaker's ability to control and navigate the flow of topics; tone choice the suitability of the tone; and conversation opening/-closing the naturalness in the initial exchange and conclusion of the discussion. A more elaborate definition can be found in Appendix Table 13. For each macro-level interactivity label, the score ranges from 1 to 5 (categorical), and higher scores indicate a more natural and authentic interactivity quality. Note that for tone choice appropriateness, higher scores denote a casual tone, and lower scores indicate a formal tone.[1]

[1] The rationale of assigning higher scores to casual tone is that in our experiments, the conversations are designed to



Speaker 1: 你好，你好，有事吗?
How are you, what is wrong?
Speaker 2: 嗯，你知道吗，我有一个很大的问题。
Well, you know, I am in a big trouble.
Speaker 1: 怎么了?
What is going on?
Speaker 2: 那个，你也知道我下学期要毕业。
Hmm, you know I will be graduating next semester.
Speaker 2: 但我的汉语水平不是很高。
But I am not very good at Chinese.
Speaker 1: 哦哦。别担心，你可以的。
Oops, don't worry, you can make it.
……

Micro-level Features

- ■ Reference Word
- ■ Noun & Verb collocation
- ■ Routinized Resources
- ■ Feedback in next turn
- ■ Backchannels
- ■ Epistemic copulas

Macro-level Features

Topic Management:	2
Tone Choice:	2
Conversation Opening:	3
Conversation Closing:	4

Dialogue Label

Overall Dialogue Quality Score: 2

Figure 1: An example of a CSL dialogue annotated with the micro-level features, macro-level interactivity labels and overall dialogue quality score.

In addition to micro- and macro-level features, we introduce an *overall score* for each dialogue to measure second language speakers' interaction abilities in open-domain conversation. Like the macro-level labels, it is scored from 1 to 5 (categorical). The overall score is designed to capture the holistic quality of a conversation, integrating elements like contextualisation, responsiveness, and communicative purposes across the whole conversation, and a high score reflects the speaker's ability to engage in fluid and meaningful interactions. The full description of each score is detailed in Appendix Table 12.

4 CNIMA Development

4.1 CSL Dialogue Collection

We first extend the CSL dialogue dataset developed by Wu and Roever (2021), which has approximately 8,000 dialogues. We followed a similar process, where we recruited 20 learners of Chinese Mandarin at four different proficiency levels.[2] Each participant was assigned to a dyadic group with a partner of a similar proficiency level. Each pair of participants did one elicited conversation task, in which they discussed an assigned topic, and two role-play activities (three tasks in total; dialogue collection instructions can be found in Appendix A.4).[3] After collecting the conversations, we recruited in-house workers to segment the conversations based on the discussion topics.[4] Including the original 8,000 dialogues from Wu and Roever (2021), our extended CSL dataset has 10,908 dialogues in total; table 1 presents some statistics.

Statistic	Count
#dialogues	10,908
#turns (average)	6
#turns (max)	13
#tokens /w token-level features	170,852
#tokens /w utterance-level features	94,516

Table 1: CNIMA Statistics.

Measure	Micro-level Features	Macro-level Labels	Overall Score
α	0.66	0.67	0.61
r	0.65	0.68	0.62

Table 2: Inter-annotator agreement for micro-level, macro-level features and overall scores.

4.2 Macro/Micro-level Feature Annotation

Given the dialogues, we annotate them for micro-level features, macro-level interactivity labels and overall quality scores based on the evaluation framework introduced in Section 3. To this end, we have recruited twelve postgraduate students who are native Chinese speakers. The authors of this paper first trained the annotators and the annotator training manual can be found in Appendix A.3. During the training, annotators will see one dialogue as an example to understand the requirements and learn how to use the annotation platform (Appendix A.1). After training, each annotator was assigned 950 dialogues, and each dialogue was annotated by two annotators.[5] The annotation was conducted over two months, where in the first few weeks we (first author) checked for initial agreement, discussed feedback, and fixed any annotation errors (e.g. missing values in the annotation results) before proceeding with the full annotation. The annotation process was guided by an annotation guide (Appendix A.3), which provided definitions and examples for each micro-level feature, macro-level

be informal discussions, and appropriate use of casual tone signifies a more natural communicative interaction.

[2] Proficiency level can be categorized into the following levels: beginner, lower, intermediate, and high (Wu and Roever, 2021).

[3] All participants have signed consent approved by an ethics board to agree to audio recording.

[4] The workers are undergraduate linguistics major students, and they are native Chinese speakers.

[5] The assignment is created in a way such that we have the same pair of annotators for every batch of 40 dialogues.

interactivity label and the overall dialogue quality score. For the overall score, in addition to providing the score, the annotators are also asked to write down their reasons for justifying their label.

As explained in Section 3, the micro-level features are annotated as spans, and the macro-level labels and overall quality scores are document (dialogue) labels. To aggregate the span annotations by the two annotators for the micro-level features, for each dialogue and each micro-level feature, we iterate through each turn and select the shorter (longer) span between the two annotators if it is a token-level (utterance-level) feature.[6] For the macro-level interactivity labels, in dialogues where we have disagreement between the two annotators, we use the *majority label from their larger unsegmented conversation* as the ground truth label.[7] For the overall dialogue quality scores, for each dialogue with disagreement, we (two authors of this paper) manually assess the justification provided by the annotators to determine the ground truth.[8]

To measure annotation quality, for the micro-level features, we measure agreement between the annotators at the token level for each micro-level feature, i.e., we compute agreement statistics based on the presence or absence of the feature as marked by the annotators for each word token.[9] We calculate Pearson correlation coefficient r (Cohen et al., 2009) and Krippendorff's α (Krippendorff, 2018) summarise the results in Table 2. The agreement is above 0.6 over all the features/labels, indicating a good consensus between annotators under this framework.

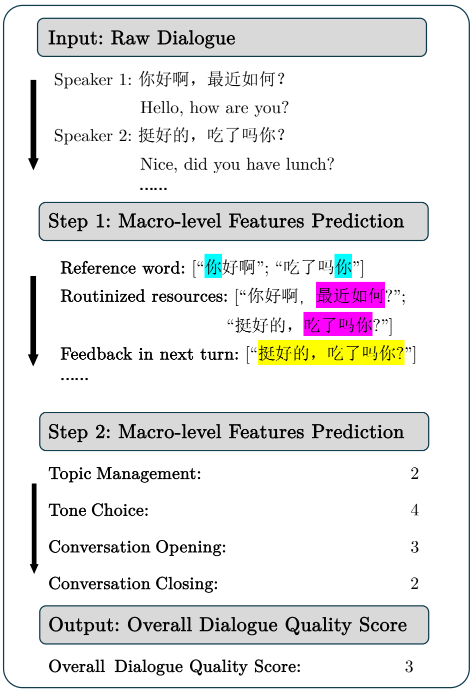

Input: Raw Dialogue
Speaker 1: 你好啊，最近如何？
Hello, how are you?
Speaker 2: 挺好的，吃了吗你？
Nice, did you have lunch?
......
Step 1: Macro-level Features Prediction
Reference word: [“你好啊”; “吃了吗你”]
Routinized resources: [“你好啊，最近如何?”;
“挺好的，吃了吗你?”]
Feedback in next turn: [“挺好的，吃了吗你?”]
......
Step 2: Macro-level Features Prediction
Topic Management: 2
Tone Choice: 4
Conversation Opening: 3
Conversation Closing: 2
Output: Overall Dialogue Quality Score
Overall Dialogue Quality Score: 3

Figure 2: Pipeline for automated scoring of the CSL dialogue on three steps

5 Automation Pipeline

We now propose automating the prediction of the micro-level features (step 1), macro-level interactivity labels (step 2), and overall score (step 3); see Figure 2 for an illustration. We experiment with classical machine learning models (Logistic Regression (LR), Random Forest (RF), and Naïve Bayes (NB)), fine-tuned Chinese BERT (Cui et al., 2021), and GPT-4o (OpenAI, 2024) with prompts to automate these steps. Note that these predictions are done in sequence, where the output from the previous step is used as input for the next step (e.g. when predicting the macro-level labels, we use the predicted micro-level features as input). We partition CNIMA into train, development and test sets (ratio = 7:1:2) for these experiments.

Step 1 This step predicts the spans of the 17 micro-level features given an input dialogue, and we experiment with BERT and GPT-4o here. For BERT, we fine-tune 17 span classifiers (one for each feature) to classify the presence or absence of a feature for each word in the dialogue. For GPT-4o, we do one-shot prompting (i.e. 1 dialogue with the expected output as a demonstration) with instructions to generate spans for the token-level and utterance-level features separately.[10]

[6] We do this because token-level features (e.g., ‘Reference word’ (She, her, he) or ‘Tense Choice’ (is doing, done, did) tends to be very short and limited to 1 or 2 words and so the shorter spans are likely to be more accurate (Greer, 2023). The same reasoning applies to utterance-level spans.

[7] Recall that these dialogues are segmented from a larger, longer conversation. The majority label is the most frequent label when we pool together all the labels from the segmented, smaller dialogues that belong to the same original conversation. Our rationale for deferring to this majority label is that these are cases where it is difficult to determine the correct dialogue level label using the smaller segmented dialogue (this happens more with topic management than other interactivity labels), and so it is best that we look at all the labels across all the segmented dialogues collection that belong to the same conversation.

[8] Note that when resolving the disagreement, we only consider the justification *without* looking at their macro-level interactivity labels. This is so we create an unbiased ground truth for the overall score that is independent of the macro-level labels.

[9] In other words, the unit of analysis here is a word token, and the output is a binary value for each annotator indicating whether it has been marked for the feature.

[10] In other words, we have 2 prompts for each dialogue.

Models	Topic	Tone	Opening	Closing
LR	0.829	0.859	0.821	0.810
RF	0.831	0.816	0.858	0.852
NB	0.846	0.783	0.811	0.846

Table 3: F1 performance of predicting the macro-level features using human-annotated micro-level features.

Step 2 This step predicts the macro-level interactivity labels for topic management, tone choice appropriateness, and conversation opening and closing. Here, we experiment with classical models (LR, RF and NB), BERT and GPT-4o, and for each of them, we train four (one for each interactivity label) 5-class classifiers (each interactivity label has 5 classes). For the classical models, we follow Gao et al. (2024) and convert the micro-level feature spans into normalised counts and use them as input features to train LR/RF/NB to predict the interactivity labels. For BERT and GPT-4o, the input is the dialogue concatenated with the micro-level feature spans. BERT is fine-tuned as a document classifier, and GPT-4o is one-shot prompted with instructions.

Step 3 This step predicts the overall dialogue quality score. As this is a document classification like step 2, we follow largely the same process for building LR, RF, NB, BERT and GPT-4o. Note that the classical models (LR, RF, NB) use only the normalised macro-level interactivity score as input features (in other words, they have only 4 input features), while BERT and GPT-4o uses a concatenation of the dialogue, micro-level feature spans, and macro-level interactivity scores as input.

For more details on configurations and prompts, see Appendix A.5 for BERT and A.6 for GPT-4o.

6 Experiments

We first replicate the ESL experiments of Gao et al. (2024) for CSL and compare their findings in Section 6.1. We then assess the performance of our proposed automated approach for SL assessment in Section 6.2.

6.1 Transferability of the Evaluation Framework from ESL to CSL

Based on the *human-annotated* micro-level features, Gao et al. (2024) convert them into normalised counts and use them as input features to train an LR, NB and RF classifier to predict macro-level interactivity labels (i.e. topic management, tone appropriateness, conversation opening and closing). They found: (1) strong prediction performance; (2) high-impact micro-level features that are *common across all* interactivity labels (by interpreting feature importance given by the trained classifiers); and (3) high-impact micro-level features that are *specifically* predictive for an interactivity label. We follow Gao et al. (2024) to compute the common and specific features for (2) and (3); details in Appendix A.10.

We replicate their experiments here, using our annotated CSL data (CNIMA), to see whether their findings transfer across languages.

For (1), we present the F1 performance of predicting macro-level interactivity labels for CNIMA in Table 3. We see a strong performance, where F1 is over 0.8 in most models over the 4 interactivity labels. These results echo the ESL results in Gao et al. (2024),[11] providing evidence that the evaluation framework is robust across languages.

For (2), we present the results in Table 4. We found that micro-level features such as ‘Feedback in Next Turn’ and ‘Reference Word’ are high-impact features for both ESL and CSL — this underlines their fundamental role in impacting dialogue interactive dynamics, and it is language-universal. Interestingly, however, we also found some differences. For example, ‘Noun & Verb Collocation’ and ‘Routinized Resources’ are strong features only for CSL, and this might be because, in Chinese, these fixed terms of expressions are often used to show social closeness in open-domain conversations (Roever and Dai, 2021). ‘Code Switching’ and ‘Tense Choice’, on the other hand, are two strong features only for ESL. This is intuitive, as Chinese has no tenses while English tenses correlate with the social expression in communications (Lam, 2018). For the full comparison results, see Appendix Table 14.

For (3), Table 5 and 6 present the CSL and ESL results, respectively. We found that for topic management, ‘Negotiation of Meaning’ and ‘Question-Based Responses’ are high-impact micro-level features for both ESL and CSL, demonstrating that these are language-universal features important to drive the flow of topics in conversation. For tone

[11] Gao et al. (2024) found marginally higher performance for conversation opening/closing (F1 scores are over 0.9) and lower performance for topic management and tone choice appropriateness (F1 scores are a little below 0.8).

Language	LR	RF	NB
ESL	**Code Switching** **Reference Word*** **Feedback in Next Turn*** **Formulaic Responses** **Tense Choice**	**Code Switching** **Feedback in Next Turn*** Question-based responses Non-factive Verb **Reference Word***	**Feedback in Next Turn*** **Formulaic Responses** **Reference Word*** Negotiation of Meaning **Tense Choice**
CSL	**Feedback in the Next Turn** **Noun & Verb Collocation** Tense Choice **Reference Word*** **Subordinate Clauses***	**Subordinate Clauses*** **Routinized Resources** **Reference Word*** Code-Switching **Feedback in Next Turn**	**Negotiation of Meaning** **Noun & Verb Collocation** **Routinized Resources** **Subordinate Clauses*** **Reference Word***

Table 4: High impact common micro-level features over the three classifiers for predicting macro-level features with overlapping features in two/three classifiers by Bold/asterisk. ESL results are reproduced from Gao et al. (2024).

appropriateness, we generally see less commonality (exceptions: 'Feedback in Next Turn', and 'Routinized Resources' which appear in both languages), suggesting that languages tend to use different features for managing tones (Zilio et al., 2017). For conversation opening, we see some similarities (e.g. 'Question-Based Responses') and also divergences (e.g. 'Subordinate Clauses' and 'Adj./Adv. Expressing' for ESL and 'Epistemic Copulas' and 'Epistemic Modals' for CSL). One explanation is that English tends to use adjectives and adverbs to extend a topic (e.g. *Totally, I think...*), while Chinese prefers modals and copulas "应该吧，或许呢" (translation: *possibly yeah*) to control topic change (Alduais et al., 2022). For conversation closing, we see a more similar trend, where 'Collaborative Finishes' and 'Backchannels' are strong features across both ESL and CSL. Despite some language-specific variations, the strategies of ending a conversation are largely similar in human communication (Lam, 2021).

6.2 Evaluation of Automated Pipeline

We now evaluate our automated 3-step approach to predicting the overall dialogue quality score and present the results in Table 7. In terms of model names, each component denotes the model used in a particular step, e.g. "BERT+LR+LR" means we use BERT for step 1 and LR for steps 2 and 3. Also, "GPT4" refers to GPT-4o. For brevity, we only include LR results for the classical models as they all have similar performances. In addition to our 3-step approach, we also include 2 baselines that predict the overall dialogue quality score *directly* based on the input dialogue: (1) fine-tuned BERT ("BERT (One-step)"); and (2) one-shot GPT-4o with instructions ("GPT4 (One-step)"). We also include a variation where the first step uses human-annotated micro-level features (e.g. "Human+LR+LR") to understand how much performance degrades when substituting them with predicted features.

Interestingly, the baselines ("BERT (One-step)" and "GPT4 (One-step)") perform quite poorly, achieving F1 scores of 0.379 and 0.585, respectively. This indicates using the raw dialogue directly for predicting the overall quality is difficult (most studies in NLP, however, follow this setup for dialogue evaluation (Finch et al., 2023; Zhao et al., 2022; Yang et al., 2024)). For the variation where we use human-annotated micro-level features ("Human+LR+LR", "Human+BERT+BERT", and "Human+GPT4+GPT4"), we see that BERT is generally the best model and LR the worst, which is no surprise, given that BERT is pre-trained. GPT4, however, is not far from BERT, even though it is not fine-tuned. Overall, the performance is encouraging, and we see that the best models achieve over 0.80 F1.

When we look at the fully automated pipeline (bottom 4 rows in Table 7), "'BERT+BERT+BERT" and "'GPT4+GPT4+GPT4" perform very strongly (0.807 and 0.791), demonstrating that we have a fully automated system that can reliably assess the quality score of a dialogue. We also notice an interesting trend: When we use either BERT or GPT4 for predicting the micro-level features, LR (i.e. "BERT+LR+LR" and "GPT4+LR+LR") performs very poorly for predicting the overall score, even though the span prediction performance of

Topic	Tone	Opening	Closing
Logistic Regression			
Negotiation of Meaning	**Feedback in Next Turn**	**Epistemic Copulas**	**Backchannels**
Epistemic Copulas	**Collaborative finishes**	Formulaic Responses	**Collaborative Finishes***
Collaborative Finishes	**Routinized Resources**	**Question-Based Responses***	**Epistemic Copulas**
Question-Based Responses	Formulaic Responses	**Epistemic Modals**	Subordinate Clauses
Backchannels*	**Question-Based Responses**	Collaborative finishes	Formulaic Responses
Naïve Bayes			
Non-factive Verb Phrase	Epistemic copulas	Code-switching	Code-switching
Collaborative Finishes	**Collaborative finishes**	**Feedback in Next Turn**	Epistemic Modals
Code-switching	Impersonal subject + non-factive verb	**Epistemic Modals**	**Epistemic Copulas**
Backchannels*	**Backchannels**	**Epistemic Copulas**	**Collaborative Finishes***
Epistemic Copulas	**Question-Based Responses**	**Question-Based Responses***	**Question-Based Responses**
Random Forest			
Noun & verb collocation	Tense choice	**Feedback in Next Turn**	**Backchannels**
Negotiation of Meaning	**Backchannels**	Noun & Verb Collocation	**Subordinate Clauses**
Backchannels*	Negotiation of Meaning	Collaborative Finishes	Formulaic Responses
Question-Based Responses	**Feedback in Next Turn**	**Question-Based Responses***	**Collaborative Finishes***
Collaborative Finishes	**Routinized Resources**	Formulaic Responses	**Question-based Responses**

Table 5: CSL: High impact interactivity-specific micro-level features. For each interactivity label, bold/asterisk indicates overlapping features in two/three classifiers.

Topic	Tone	Opening	Closing
Logistic Regression			
Negotiation of Meaning*	**Routinized Resources***	Epistemic Modals	**Backchannels***
Subordinate Clauses*	Adj./Adv. Expressing	**Formulaic Responses**	**Adj./Adv. Expressing**
Noun&Verb Collocation	**Feedback in Next Turn***	**Question-Based Responses***	Formulaic Responses
Question-Based Responses	**Formulaic Responses***	**Subordinate Clauses***	**Collaborative Finishes***
Subordinate clauses*	**Reference Word**	**Adj./Adv. Expressing***	Epistemic Copulas
Naïve Bayes			
Non-factive Verb Phrase	**Routinized Resources***	**Adj./Adv. Expressing***	**Adj./Adv. Expressing**
Question-Based Responses	**Feedback in Next Turn***	**Routinized Resources**	Epistemic Modals
Adj./Adv. Expressing	**Epistemic Copulas**	**Subordinate Clauses***	**Backchannels***
Negotiation of Meaning*	Question-Based Responses	Epistemic Copulas	**Collaborative Finishes***
Subordinate clauses*	**Subordinate Clauses***	**Question-Based Responses***	Question-Based Responses
Random Forest			
Negotiation of Meaning*	**Epistemic Copulas**	Feedback in Next Turn	Feedback in Next Turn
Formulaic Responses	Backchannels	**Subordinate Clauses***	Subordinate clauses
Subordinate Clauses*	**Feedback in Next Turn***	**Adj./Adv. Expressing***	**Collaborative Finishes***
Epistemic Copulas	**Negotiation of Meaning**	**Question-Based Responses***	**Formulaic Responses**
Question-Based Responses	**Routinized Resources***	**Formulaic Responses**	**Backchannels***

Table 6: ESL: High impact interactivity-specific micro-level features (reproduced from Gao et al. (2024)). For each interactivity label, bold/asterisk indicates overlapping features in two/three classifiers.

BERT and GPT4 for the micro-level features is no+t poor (we will revisit this in Section 6.3). This suggests that the classical machine learning models are less tolerant of noise.[12]

Taking all these results together, we show that when it comes to assessing SL dialogue quality, it is important to take a pipeline approach to predict important intermediate features (i.e. micro- and macro-level features) before predicting the overall quality score. This design also has another advantage: it is more interpretable because, given an overall score, we can also look at the intermediate results to understand what or where went wrong, providing more nuanced feedback that can benefit both teachers and second language students. Lastly, the strong performance of GPT4 (which is only marginally behind fine-tuned BERT) also has

[12] Note that LR does not use the dialogue as input, so any noise in the intermediate features (micro-level or macro-level) will have a bigger impact.

Models	F1
BERT (One-step)	0.379
GPT4 (One-step)	0.585
Human+LR+LR	0.772
Human+BERT+BERT	0.860
Human+GPT4+GPT4	0.842
BERT+LR+LR	0.329
BERT+BERT+BERT	0.807
GPT-4+LR+LR	0.667
GPT4+GPT4+GPT4	0.791

Table 7: F1 performance of the dialogue overall score for the CSL dialogue in three steps in different models

Models	Topic	Tone	Opening	Closing
LR	0.643	0.357	0.081	0.079
RF	0.463	0.235	0.066	0.090
NB	0.560	0.290	0.106	0.117

Table 8: Feature Importance of the four interactivity labels across three machine learning models

another important implication: we can adapt our SL assessment system to another language *without requiring* large-scale manually-annotated data by prompting LLMs.

6.3 Additional Analyses

To understand how macro-level interactivity labels contribute to the overall quality score, we present their feature importance as learned by LR, RF and NB in the "Human+LR+LR", "Human+RF+RF" and "Human+NB+NB" models in Table 8. Topic management appears to be the most important predictor, followed by tone appropriateness. Conversation opening and closing, on the other hand, has relatively small weights. These observations are supported by Roever and Ikeda (2023), who similarly found that topic flow and tone are dominant factors that decide dialogue quality.

Next, we assess how well BERT and GPT4 perform for the first (micro-level feature span prediction) and the second step predictions (macro-level label classification); full results in AppendixA.5 Table14 and Table**??** for BERT and GPT4 respectively. To summarise, for most of the 17 micro-level features (step 1), both BERT and GPT4 perform very well (average F1 for BERT and GPT4 is 0.89 and 0.81, respectively). That said, BERT and GPT4 struggle on two features related to non-factive verbs, and this can be attributed to the rarity of these features in the Chinese (Roever and Ikeda, 2023). Overall, these results show that the proposed micro-features by Gao et al. (2024) transfer very well from English to Chinese.

For the second step, we look at the performance of "BERT+BERT" and "GPT4+GPT4" for predicting the macro-level labels (Appendix Table 16). Across all labels, both BERT and GPT4 perform well, averaging about 0.83 F1 for BERT and 0.78 F1 for GPT4, respectively. Together, these encouraging intermediate evaluation results explain why our pipeline approach can consistently score the overall dialogue quality.

7 Conclusion

We develop CNIMA, a CSL dataset with 10K dialogues annotated based on an evaluation framework that assesses micro-level features, macro-level interactivity labels, and overall quality scores. As the evaluation framework is originally designed for ESL dialogues (Gao et al., 2024), our first contribution tests how well the framework transfers to CSL dialogues. We found that the evaluation framework is robust across languages and further revealed universal and language-specific insights about the relationships between micro-level and macro-level features. Our second contribution is we propose an automated approach that predicts the micro-level features, macro-level labels and overall quality scores in sequence. We experimented with classical models, BERT and GPT-4o, and found that the BERT and GPT4-o-based systems perform very well in predicting the overall dialogue quality score. Our approach is interpretable, and the LLM-based variant can be easily adapted to another language for second language assessment without requiring annotated training data.

8 Limitations

The scope of the current paper is limited to CSL, but it would be equally interesting to see how the evaluation framework would work for other SL dialogues. Admittedly the models we experiment with are limited in terms of novelty, but we contend that is not the focus of our contribution in this paper.

Ethics Statement

This study is conducted under the guidance of the ACL Code of Ethics. We manually filtered out potentially offensive content and removed all information related to the identification of annotators. The annotation protocol is approved under the university's Human Ethics Application (reference number ANONYMISED).

References

Ahmed Alduais, Issa Al-Qaderi, and Hind Alfadda. 2022. Pragmatic language development: Analysis of mapping knowledge domains on how infants and children become pragmatically competent. *Children*, 9(9):1407.

Jens Allwood. 2008. Dimensions of embodied communication-towards a typology of embodied communication. *Embodied communication in humans and machines*, pages 257–284.

Jill Burstein, Geoffrey T LaFlair, Antony John Kunnan, and Alina A von Davier. 2021. A theoretical assessment ecosystem for a digital-first assessment—the duolingo english test. *DRR-21-04*.

Ming-Bin Chen, Jey Han Lau, and Lea Frermann. 2023. The uncivil empathy: Investigating the relation between empathy and toxicity in online mental health support forums. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association*, pages 136–147.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Ronald Cumbal. 2024. *Robots Beyond Borders: The Role of Social Robots in Spoken Second Language Practice*. Ph.D. thesis, KTH Royal Institute of Technology.

David Wei Dai. 2022. *Design and validation of an L2-Chinese interactional competence test*. Ph.D. thesis, University of Melbourne (Australia).

Olov Engwall, Ronald Cumbal, José Lopes, Mikael Ljung, and Linnea Månsson. 2022. Identification of low-engaged learners in robot-led second language conversations with adults. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(2):1–33.

Keelan Evanini, Maurice Cogan Hauck, and Kenji Hakuta. 2017. Approaches to automated scoring of speaking for k–12 english language proficiency assessments. *ETS Research Report Series*, 2017(1):1–11.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada. Association for Computational Linguistics.

Rena Gao, Carsten Roever, and Jey Han Lau. 2024. Interaction matters: An evaluation framework for interactive dialogue assessment on english second language conversations.

Wei 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*.

Kaixuan Gong. 2023. Challenges and opportunities for spoken english learning and instruction brought by automated speech scoring in large-scale speaking tests: a mixed-method investigation into the washback of speechrater in toefl ibt. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):25.

Tim Greer. 2023. Grammar-in-interaction and its place in assessing interactional competence. *Applied Pragmatics*, 5(2).

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Jeremy F Jones and Quanling Liu. 2023. Analyzing test-takers' experiences of high-stakes automated language testing. *English as a Foreign Language International Journal*, 3(1):1–41.

Nahal Khabbazbashi, Jing Xu, and Evelina D Galaczi. 2021. Opening the black box: Exploring automated speaking evaluation. *Challenges in Language Testing Around the World: Insights for language test users*, pages 333–343.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Daniel MK Lam. 2018. What counts as "responding"? contingency on previous speaker contribution as a feature of interactional competence. *Language Testing*, 35(3):377–401.

Daniel MK Lam. 2021. Don't turn a deaf ear: A case for assessing interactive listening. *Applied Linguistics*, 42(4):740–764.

OpenAI. 2024. Hello gpt-4o.

José Carlos Paiva, José Paulo Leal, and Álvaro Figueira. 2022. Automated assessment in computer science education: A state-of-the-art review. *ACM Transactions on Computing Education (TOCE)*, 22(3):1–40.

Carsten Roever and David W Dai. 2021. Reconceptualizing interactional competence for language testing. *Assessing speaking in context: Expanding the construct and its applications*, pages 23–49.

Carsten Roever and Naoki Ikeda. 2023. The relationship between l2 interactional competence and proficiency. *Applied Linguistics*, page amad053.

Abdullayev Shaxobiddin. 2024. A discourse analysis of modal verbs in modern english: Patterns and functions. *Journal of new century innovations*, 50(2):145–147.

Eric 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. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Jingxuan Wu and Carsten Roever. 2021. Proficiency and preference organization in second language mandarin chinese refusals. *The Modern Language Journal*, 105(4):897–918.

Jing Xu. 2018. Measuring "spoken collocational competence" in communicative speaking assessment. *Language Assessment Quarterly*, 15(3):255–272.

Jing Xu, Edmund Jones, Victoria Laxton, and Evelina Galaczi. 2021. Assessing l2 english speaking using automated scoring technology: examining automarker reliability. *Assessment in Education: Principles, Policy & Practice*, 28(4):411–436.

Bohao Yang, Kun Zhao, Chen Tang, Liang Zhan, and Chenghua Lin. 2024. Structured information matters: Incorporating abstract meaning representation into llms for improved open-domain dialogue evaluation. *arXiv preprint arXiv:2404.01129*.

Foong Ha Yap and Anindita Sahoo. 2024. Versatile copulas and their stance-marking uses in conversational odia, an indo-aryan language. *Lingua*, 297:103641.

Jianqiao Zhao, Yanyang Li, Wanyu Du, Yangfeng Ji, Dong Yu, Michael Lyu, and Liwei Wang. 2022. FlowEval: A consensus-based dialogue evaluation framework using segment act flows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10469–10483, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2017. Using nlp for enhancing second language acquisition. In *RANLP*, pages 839–846.

A Appendix

A.1 Software Availability

To contribute to the research community and facilitate further development and collaboration, we have made the source codes of our innovative annotation tool publicly available. The tool, designed with a focus on enhancing the efficiency and accuracy of data annotation processes, has been developed through meticulous research and development efforts. It incorporates a range of features tailored to meet the needs of researchers and practitioners working in fields that require precise and reliable annotation of datasets.

Accessing the Source Code

The source codes are hosted on GitHub, a platform widely recognized for its robust version control and collaborative features. Interested parties can access the repository at the following link: https://anonymous.4open.science/r/AnnotationTool2023-CFE1/README.md. This repository is intended for research usage, underlining our commitment to supporting academic and scientific endeavours.

Key Features and Capabilities

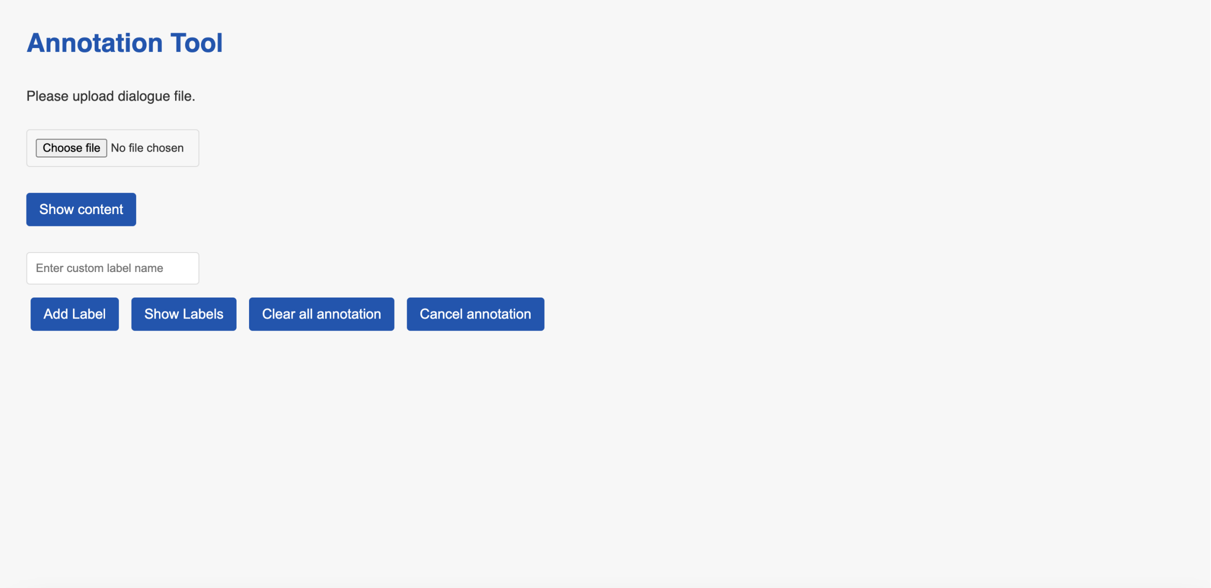
Our annotation tool stands out for its user-friendly interface, which simplifies the annotation process and allows users to work more efficiently. Among its key features are:

- **Customizable Annotation Labels:** Users can add their own set of labels to cater to the specific requirements of their projects.

- **Collaborative Annotation Support:** Facilitating teamwork, the tool allows multiple annotators to work on the same dataset simultaneously, ensuring consistency and reducing the time required for project completion.

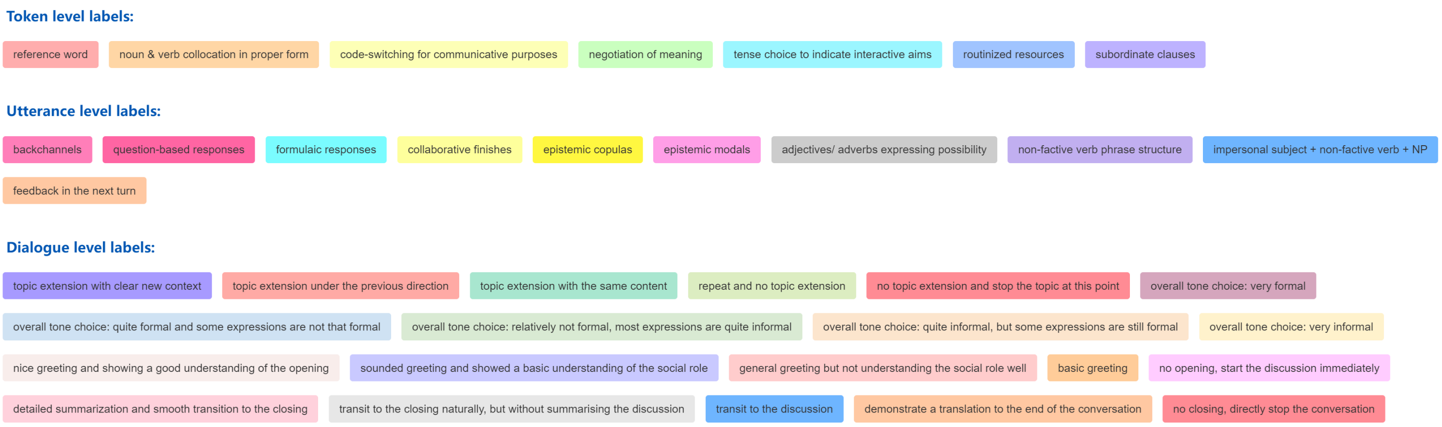
- **Annotation History Tracking:** All the annotation history, such as changes made, can be tracked, and any further modifications can be done at any time for the user's convenience. Export Functionality: Annotated data can be exported in several formats, accommodating further analysis or use in machine learning models.

A.2 Pages View For Span Annotation Website Interface



Please upload dialogue file.

Figure 3: Annotation tool Demo



Token level labels:

reference word | noun & verb collocation in proper form | code-switching for communicative purposes | negotiation of meaning | tense choice to indicate interactive aims | routinized resources | subordinate clauses

Utterance level labels:

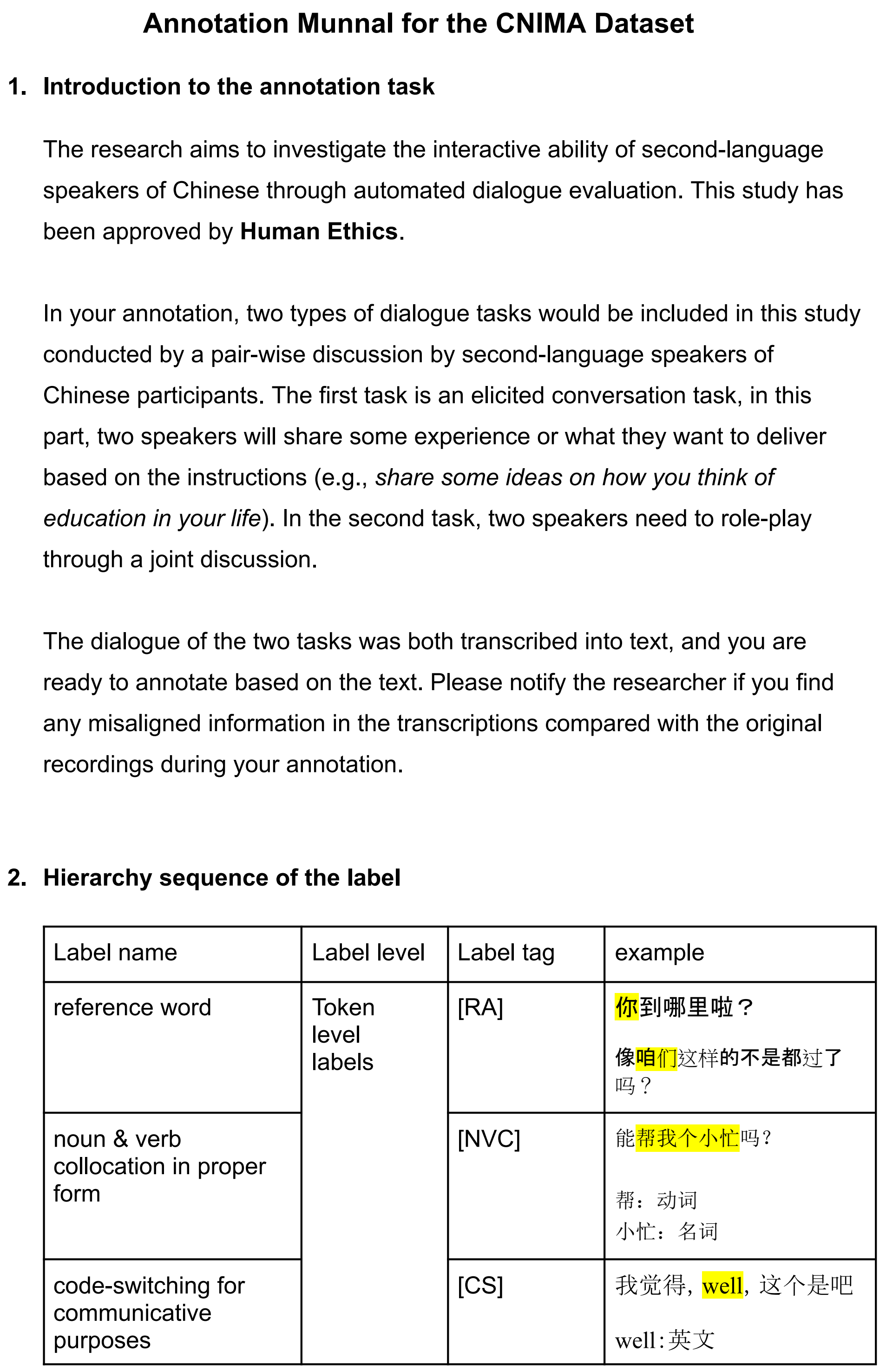
backchannels | question-based responses | formulaic responses | collaborative finishes | epistemic copulas | epistemic modals | adjectives/ adverbs expressing possibility | non-factive verb phrase structure | impersonal subject + non-factive verb + NP | feedback in the next turn

Dialogue level labels:

topic extension with clear new context | topic extension under the previous direction | topic extension with the same content | repeat and no topic extension | no topic extension and stop the topic at this point | overall tone choice: very formal | overall tone choice: quite formal and some expressions are not that formal | overall tone choice: relatively not formal, most expressions are quite informal | overall tone choice: quite informal, but some expressions are still formal | overall tone choice: very informal | nice greeting and showing a good understanding of the opening | sounded greeting and showed a basic understanding of the social role | general greeting but not understanding the social role well | basic greeting | no opening, start the discussion immediately | detailed summarization and smooth transition to the closing | transit to the closing naturally, but without summarising the discussion | transit to the discussion | demonstrate a translation to the end of the conversation | no closing, directly stop the conversation

Figure 4: Hierarchical Label Assignment Demo

A.3 Manual



Annotation Munnal for the CNIMA Dataset

1. Introduction to the annotation task

The research aims to investigate the interactive ability of second-language speakers of Chinese through automated dialogue evaluation. This study has been approved by **Human Ethics**.

In your annotation, two types of dialogue tasks would be included in this study conducted by a pair-wise discussion by second-language speakers of Chinese participants. The first task is an elicited conversation task, in this part, two speakers will share some experience or what they want to deliver based on the instructions (e.g., *share some ideas on how you think of education in your life*). In the second task, two speakers need to role-play through a joint discussion.

The dialogue of the two tasks was both transcribed into text, and you are ready to annotate based on the text. Please notify the researcher if you find any misaligned information in the transcriptions compared with the original recordings during your annotation.

2. Hierarchy sequence of the label

Label name	Label level	Label tag	example
reference word	Token level labels	[RA]	你到哪里啦？ 像咱们这样的不是都过了吗？
noun & verb collocation in proper form		[NVC]	能帮我个小忙吗? 帮：动词 小忙：名词
code-switching for communicative purposes		[CS]	我觉得, well, 这个是吧 well:英文

negotiation of meaning (appropriate tense to show meaning)		[NM]	SPK_1: 苏州,苏州 SPK_2:哦,苏州是南方的城市就比较热一点 SPK_1: 那么看 你还是没有做 SPK_2:的确是的毕竟我一点也没学.
tense choice to indicate interactive aims (politeness in talking/ social distance/ context variance)		[TT]	SPK_1:一两门, 啊, 那我们 现在 还 能 学 啥 呀？
routinized resources (projector construction)		[RR]	哎, 你说你
subordinate clauses		[RC]	你是数学好 但是我还是觉得这种东西需要自己用很长的时间弄了, 才是 真的会了.
backchannels	Utterance level labels	[BC]	会的， 会有点累
question-based responses		[QR]	SPK_1: 我不是把她教的那个知识给反驳了吗？ SPK_2: 你既然有反驳的能力, 那你还是自学吧.
formulaic responses (固定用词)		[FR]	但是不管怎么样 我 也学会了。差不了多少。
collaborative finishes		[CF]	SPK_1: 好的 再见

			SPK_2： 再见
epistemic copulas		[H1]	一个人去还是觉得有点变扭.
epistemic modals		[H2]	我觉得像苏州 好像有个这种辅导
adjectives/ adverbs expressing possibility		[H3]	我估计可能有些这样的情况
non-factive verb phrase structure		[H4]	你可以认识他们 我姑且能跟上吧
impersonal subject + non-factive verb + NP		[H5]	我认为可能他会迟到 Impersonal Subject (可能): This subject is impersonal because it does not refer to a specific individual but rather expresses a possibility. Non-Factive Verb (认为): This verb is non-factive because it introduces a belief or opinion rather than a fact. Noun Phrase (他会迟到): This is the noun phrase that completes the sentence, stating what is believed or thought.
feedback in the next turn		[FB]	-你感觉如何？ -会的, 会 有点 累

topic extension with clear new context (change to utterance level, but more information context depends)	Dialogue level labels	[T1]	我只有你这个朋友，你又不肯帮我。别人跟我关系都很一般的啊。你知道我什么意思吧？
topic extension under the previous direction		[T2]	说到朋友，我只有你这个朋友。
topic extension with the same content		[T3]	你说朋友啊，我觉得吧，很难说。
repeat and no topic extension		[T4]	关于朋友的事吗？
no topic extension and stop the topic at this point		[T5]	朋友?
conversation opening		[CO1] [CO2] [CO3] [CO4] [CO5]	CO1: nice greeting and show a good understanding of conversation opening in social interactions. CO2: sounded greeting and show a basic understanding of the social role. CO3: general greeting and didn’t demonstrate a good understanding of the social role. CO4: basic greeting. CO5: no opening just start the discussion immediately.
conversation closing		[CC1] [CC2] [CC3] [CC4] [CC5]	CC1: detailed summarization and smooth transition to the closing of the conversation.

			CC2: transit to the closing naturally, but without any summarization of the discussion. CC3: demonstrate a translation to the end of the conversation. CC4: transit to the end of the discussion. CC5: no closing, just stop the conversation.
overall tone choice: very formal		[OT1]	很荣幸与您见面，幸会。
overall tone choice: quite formal and some expressions are not that formal		[OT2]	见到你真好啊，最近如何？
overall tone choice: relatively not formal, most expressions are quite informal		[OT3]	好久不见哎，真是有段日子了啊。
overall tone choice: quite informal, but some expressions are still formal		[OT4]	真是有阵子不见了，别来无恙啊哥们。
overall tone choice: very informal		[OT5]	我天，真是老久没见了铁子，抱一个！

3. Label detailed definitions

Label Category	Aspect	Definition
Reference word	Word choice	A reference word, also known as a referential word or referent, is a linguistic term used to describe a word or expression in a sentence that refers to or stands in place of something else in the text. Reference

		words are used to avoid repetition and to link different parts of a text together by indicating what a subsequent word or phrase relates to. Reference words can take various forms, including pronouns, demonstratives, and other words that replace or point to nouns or noun phrases.
Noun & verb collocation in proper form		Collocations are words or phrases that habitually occur together, forming a strong and natural linguistic association. In the case of noun-verb collocations, a particular noun is often paired with a particular verb due to convention, tradition, or linguistic patterns. These collocations contribute to the fluency, idiomaticity, and naturalness of language. Examples of noun-verb collocations: Make a decision: "I need to make a decision." Take a shower: "I usually take a shower in the morning." Catch a cold: "I hope I don't catch a cold." Give a speech: "She gave an inspiring speech."
Code-switching for communicative purposes		Code-switching for communicative purposes refers to the deliberate or subconscious alternation between two or more languages or dialects within a single conversation or utterance by bilingual or multilingual speakers. This linguistic phenomenon is employed to fulfill specific communicative needs or functions, such as clarifying a point, expressing identity, signaling solidarity or

		distinction, accommodating to the listener's language preference, or conveying concepts and emotions more effectively in one language over another. Code-switching is not merely a random mixing of languages but a sophisticated communicative strategy that reflects the speaker's linguistic competence and cultural awareness, often used to navigate and negotiate the social and contextual dynamics of interaction.
Negotiation of meaning (appropriate tense to show meaning)	Contextual tense usage	Negotiation of meaning refers to the interactive process through which speakers of different linguistic backgrounds or competencies collaboratively work to understand each other's intentions, messages, and linguistic expressions when communication breakdowns occur. This involves the use of clarification requests, confirmation checks, comprehension checks, and paraphrasing, among other communicative strategies, to ensure mutual understanding is achieved. The negotiation of meaning is a fundamental aspect of second language acquisition and communicative language teaching, highlighting the dynamic nature of language use and the active role learners play in constructing meaning through interaction.
Tense choice to indicate interactive aims (politeness / social distance/ context)		Tense choice to indicate interactive aims involves the strategic use of verb tenses by speakers to fulfill specific communicative goals or intentions within an interaction. This linguistic strategy encompasses the selection of

		present, past, future, or perfect tenses to convey nuances of time, mood, or aspect, directly influencing the interpretation and direction of the dialogue. Through careful tense selection, speakers can clarify the timing of events, express certainty or speculation about future occurrences, reflect on past experiences, or emphasize the continuity or completion of actions, all of which serve to enhance the clarity, persuasiveness, or relational dynamics of the communication. Tense choice, therefore, is not merely a grammatical decision but a deliberate tool employed by adept language users to navigate conversations and achieve specific interactive aims.
routinized resources (projector construction)	Interactional grammatical device	Routinized resources refer to patterns, practices, or tools that have become standardized and regularly employed within specific contexts or activities. These resources are often developed through repeated use over time, leading to a level of automation or routine in their application. In organizational or social settings, routinized resources help in streamlining processes, reducing the need for decision-making about routine tasks, and ensuring consistency in actions and outcomes. They can include documented procedures, established workflows, habitual practices, or even common language and scripts used in interpersonal interactions.
subordinate clauses		Subordinate clauses, also known as dependent clauses,

		are groups of words that contain a subject and a verb but do not express a complete thought and therefore cannot stand alone as a sentence. They function within a sentence by providing additional information to the main clause, to which they are connected by subordinating conjunctions (such as "because," "although," "when," "if") or relative pronouns (such as "who," "which," "that"). Subordinate clauses serve various roles in sentences, including acting as adjectives, adverbs, or nouns, and are essential for adding complexity, detail, and nuance to communication. Their use enables speakers and writers to articulate relationships of cause and effect, contrast, condition, time, and more, enriching the expressiveness and depth of language.

A.4 Speaking Task Collection Instruction

Task One: Elicited Conversation

Elicited Conversation Task (ECVA)

Instruction: 在这部分中，你将看到一个话题。请就此话题与你的同伴展开讨论。

In this section, you will be given a topic. Based on the topic, I would like you to talk together with your partner.

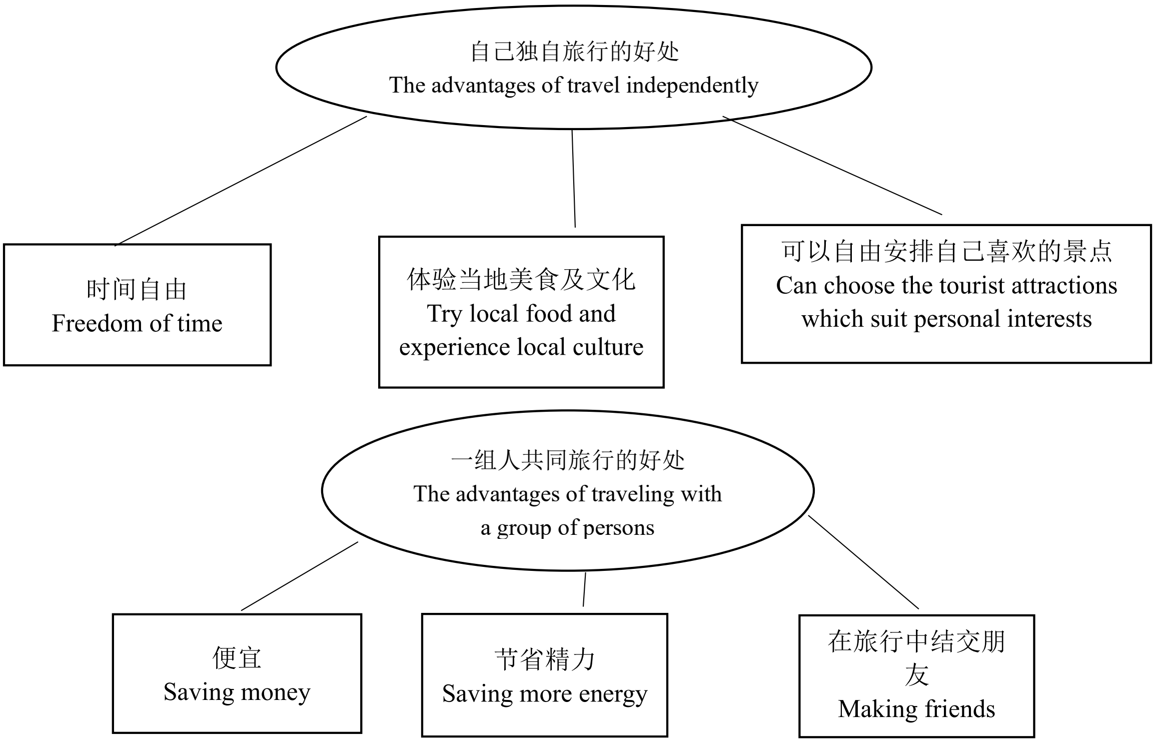
话题：目前，有很多人喜欢独自旅行，也有一些人喜欢和一组人共同去旅行。请讨论：

Nowadays, some people like traveling independently, while others like travels with a group of persons. Please discuss:

你喜欢自己独自出行旅游还是和一组人一起去旅行？为什么？
Do you like traveling individually or traveling with a group of persons? Why?

*(**温馨提示：**以下提供一些观点供你参考，在讨论中，你可以根据需要加入自己的新观点。你不必将图中所有的观点全部提及。请你们至少讨论 3 分钟，我会在适当的时候停止你们的讨论。)*

***(Note:** Some ideas are provided below. You can use these ideas in your discussions if you want. If you have some new ideas, you can use your own ideas in the discussion. You do not need to cover all the ideas provided below. Please discuss at least 3 minutes. I will stop you if necessary.)*



自己独自旅行的好处
The advantages of travel independently

时间自由
Freedom of time

体验当地美食及文化
Try local food and experience local culture

可以自由安排自己喜欢的景点
Can choose the tourist attractions which suit personal interests

一组人共同旅行的好处
The advantages of traveling with a group of persons

便宜
Saving money

节省精力
Saving more energy

在旅行中结交朋友
Making friends

Task Two: Role Play

Role Play Task One (VARP1)

Role A

提示：请你阅读以下场景，想象你自己身处此场景中。如果你有任何问题，请提出。

Note: Please read the following situation carefully and imagine yourself in the following scenario. If you have any questions, please feel free to ask.

你这个学期选择了中国文化课，这门课程要求你使用中文写一篇论文，但是你对自己的中文写作并不是那么的自信。你的好朋友中文水平非常高。你和你的朋友住在同一栋公寓。

现在，你想要去你朋友的房间请求他/她帮你改论文，你会怎么做？

You are enrolled in the Chinese culture course for this semester. One assessment task of this course is to write an essay in Chinese. However, you are not confident in your Chinese writing. Your friend's Chinese proficiency is very high. You and your friend live in the same apartment.

Now, you decide to go to your friend's room. How would you ask your friend to help you revise your essay?

Role B

提示：

请你阅读下述场景，你需要想象自己身处这个场景中。你需要以灵活自然的方式引导整个对话。在你的同伴提出请求之前，请不要接受或拒绝你同伴的请求。如果你有任何问题，请提出。

Note:

Please read the following scenario and imagine yourself in the following situation. It is your responsibility to respond to the task in a natural and flexible way. Before your interlocutor is on the record, please do not accept/ refuse your friends request. If you have any questions, please feel free to ask.

你这周需要写很多篇论文，下周就是提交这些论文的截止日期了。因此，你非常忙，并且没有足够的睡觉休息时间。你的同学们都知道你的中文水平非常好。

这时，和你同住在一栋公寓的朋友来敲你的房门。

You need to finish writing several papers this week, as the deadline for submitting the papers is next week. Therefore, you are very busy and do not have enough sleep hours. Your classmates know your Chinese proficiency is quite good.

Now, your friend, who lives in the same apartment, rings your doorbell.

Role Play Task Two (VARP2)
Role A

提示：
请你阅读以下场景，想象你自己身处此场景中。如果你有任何问题，请提出。

Note: Please read the following situation carefully and imagine yourself in the following scenario. If you have any questions, please feel free to ask.

你想要在网上报名参加汉语水平考试（HSK）。报名 HSK 考试需要使用中国银行卡在网上支付考试费用。但是，你没有中国的银行卡。你想要借你同朋友的银行卡来支付报名费。你的这位朋友和你住在同一栋公寓。

现在，你想要去你朋友的房间，请求他借给你银行卡，你会怎么做？

You want to register for HSK examination online. You need to make an online payment by using a Chinese credit card. However, you do not have a Chinese credit card. You want to borrow a credit card from your friend who lives in the same apartment.

Now, you decide to go to your friend’s room. How would ask your friend in real life to borrow the credit card?

Role B

提示：

请你阅读下述场景，你需要想象自己身处这个场景中。你需要以灵活自然的方式引导整个对话。在你的同伴提出请求之前，请不要接受或拒绝你同伴的请求。如果你有任何问题，请提出。

Note:

Please read the following scenario and imagine yourself in the following situation. It is your responsibility to respond to the task in a natural and flexible way. Before your interlocutor is on the record, please do not accept/ refuse your friends request. If you have any questions, please feel free to ask.

你有一张信用卡，你可以使用这张信用卡在网上付款。你不想把你的信用卡借给别人，因为你觉得告诉别人你银行卡密码十分不安全。

这时，和你同住在一栋公寓的朋友来敲你的房门。

You have a credit card, so you can finish the online payment by using this card. You do not want to lend others your credit card, as you think telling the passwords to other persons is unsafe.

Now, your friend, who lives in the same apartment, rings your doorbell.

A.5 BERT Configuration

We give detailed experimental settings in the following table. All the experiments can run via a single NVIDIA GeForce RTX 4070 GPU within a reasonable time.

Table 9: Experimental Details

Parameters	Configurations
Max Length	128
Tokenizer	Bert Pretrained Tokenizer
Encoder	BERT-base-uncase
Activation Function	GELU
Batch Size	32
Learning Rate	5e-5
Loss Function	CrossEntropyLoss
Optimizer	Adam
Epoch	15

A.6 Prompts for GPT-4o

The following Table 10 shows the prompts for annotating micro-level features with GPT-4o.

A.7 Prompts for GPT-4o Dialogue Overall Evaluation

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

A.8 Overall Score Description and Definitions

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

A.9 Four Interactivity Aspects Definitions and Descriptions

Table 13 shows the descriptions and definitions for dialogue's overall quality score.

A.10 Top-k feature computation method

This is reproduced from (Gao et al., 2024). Given that a trained LR, NB and RF classifier all provide weights to indicate the importance of each feature, for each classifier, we first compute *common* micro-level features f_{c} across the four interactivity labels:

$$\begin{aligned} f_{\text{c}} =& \text{top5}\big(\text{top10}(f_{\text{topic}}) \cap \text{top10}(f_{\text{tone}}) \\ & \cap \text{top10}(f_{\text{opening}}) \cap \text{top10}(f_{\text{closing}})\big) \end{aligned}$$

where topk is a function that returns the best k items given by their weights, f_{topic} denote the set of micro-level features with their weights for predicting the topic management interactivity label.

For micro-level features that are specific to each of the marco-level interactivity aspects. To that end, for each classifier we compute interactivity-specific features, e.g., for topic management, as follows:

$$\text{top10}(f_{\text{topic}}) - f_c \tag{1}$$

B Experimental Result

Span	**Description and Example**
Reference Word	A linguistic term used to avoid repetition and link different parts. **Example:** **[**你**]**到哪里去 (Where are [you] going?)
Noun & Verb Collocation	Words or phrases that habitually occur together, forming a strong and natural linguistic association. **Example:** 能**[**帮我一个忙**]**吗? (Can you [do me a favor]?)
Code-switching	The alternation between languages within a single conversation. **Example:** 我觉得, **[well]**, 这个是对的 (I think, well, this is right.)
Negotiation of Meaning	Interactive process where speakers clarify and confirm understanding. **Example:** SPK1: 那你是没有做吗? SPK2: **[**的确**]**, 我一点也没学 (SPK1: So you didn't do it? SPK2: [Indeed], I haven't learned at all.)
Tense Choice to Show Interactive Aims	Using verb tenses to fulfil specific communicative goals. **Example:** 我**[**现在**]**还不能, 因为我还有很多工作 (I can't [right now] because I still have a lot of work.)
Routinized Resources	Prefabricated linguistic elements are used to manage dialogue interactions efficiently. **Example:** **[**你说的你**]** ([As you said])
Subordinate Clauses	Clauses that provide additional information to the main clause. **Example:** 你是数学好, **[**但是**]** 我是因为需要自己努力 (You are good at math, [but] because I need to work hard myself.)
Backchannels	Brief responses from a listener, such as "uh-huh". **Example:** **[**哎哎哎。**]** ([Yep, yep, yep.])
Question Responses	Replies in a dialogue that directly or answer a preceding question. **Example:** [SPK2 你既然有反驳的能力, 你还是自学吧] (SPK2: **[Since you have the self-learning skills, you shall teach yourself.]**)
Formulaic Responses	Conventional phrases in dialogue to respond in familiar situations. **Example:** [差不了多少]**([Roughly the same])**
Collaborative Finishes	Instances in a dialogue where one speaker completes another speaker's sentence or thought. **Example:** SPK1:[好嘞再见啊] SPK2:[再见了您] (**SPK1: [Alright, See you.] SPK2:[Goodbye.]**)
Epistemic Copulas	Phrases that express a speaker's degree of certainty about a statement, often using verbs like "is" or "seems". **Example:** [一个人去还是觉得有点变扭] (**[I felt wired to go there alone]**)
Epistemic Modals	Modal verbs or phrases that express a speaker's judgment about the possibility, such as "might," "must,". **Example:** [你应该自己学会这些中文知识的] (**[You should learn these Chinese by yourself.]**)
Adjectives & adverbs of possibility	Adjectives or adverbs to show possibility, like "Possibly". **Example:** [我也许回去故乡] (**[I maybe go back to hometown.]**)
Non-factive Verb Phrase	Expressions that use verbs to convey statements without asserting them as true; verbs "think," "believe," or "seem." **Example:** [我姑且能跟上吧] (**[I can barely follow the progress.]**)
Impersonal Subject	An impersonal subject (such as "it" or "there") is followed by a non-factive verb and a noun phrase, often express opinions. **Example:** [这不好说吧] (**[It's hard to tell.]**)
Feedback in Next Turn	Using next turn to respond other speaker. **Example:** [我认为你有道理] (**[You words make sense.]**)

Table 10: LLM Annotation Prompts for CSL Dialogue Span annotation for Mirco-level features

Field	Description
CONVERSATION	A dialogue of second language Chinese conversation.
Output Fields	- **score**: The score of the interactivity of the Chinese second language dialogue (1 to 5). - **rationale**: The reason why and how the score is made.
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.

Table 11: LLM Dialogue Overall Dialogue Quality Evaluation Prompts

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

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

Response Type	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Epistemic Copulas	0.997	0.929	0.843	0.884
Formulaic Responses	0.976	0.875	0.781	0.825
Question-based Responses	0.986	0.834	0.591	0.892
Non-factive Verb Phrase Structure	0.999	0.000	0.000	0.000
Impersonal Subject + Non-factive Verb + NP	0.997	0.909	0.243	0.384
Reference Word	0.990	0.985	0.989	0.987
Routinized Resources	0.983	0.783	0.706	0.742
Noun & Verb Collocation in Proper Form	0.985	0.970	0.958	0.964
Collaborative Finishes	0.995	0.802	0.631	0.706
Tense Choice to Indicate Interactive Aims	0.994	0.957	0.930	0.943
Negotiation of Meaning	0.991	0.849	0.713	0.775
Code-switching for Communicative Purposes	0.999	0.976	0.954	0.965
Feedback in the Next Turn	0.972	0.842	0.833	0.838
Epistemic Modals	0.997	0.909	0.945	0.926
Backchannels	0.982	0.824	0.731	0.875
Subordinate Clauses	0.990	0.960	0.937	0.948
Adv & Adj Expressing	0.964	0.863	0.831	0.847

Table 14: Predicted performance of micro-level features on fine-tune BERT

Response Type	Accuracy	Precision	Recall	F1 Score
Epistemic Copulas	0.995	0.700	0.736	0.717
Formulaic Responses	0.951	0.559	0.579	0.669
Question-based Responses	0.972	0.500	0.426	0.460
Non-factive Verb Phrase Structure	0.999	0.000	0.000	0.000
Impersonal Subject + Non-factive Verb + NP	0.998	0.000	0.000	0.000
Reference Word	0.987	0.982	0.983	0.982
Routinized Resources	0.971	0.564	0.448	0.500
Noun & Verb Collocation in Proper Form	0.972	0.953	0.900	0.826
Collaborative Finishes	0.988	0.565	0.406	0.472
Tense Choice to Indicate Interactive Aims	0.990	0.915	0.826	0.768
Negotiation of Meaning	0.971	0.560	0.462	0.506
Code-switching for Communicative Purposes	0.999	1.000	0.941	0.869
Feedback in the Next Turn	0.951	0.679	0.741	0.708
Epistemic Modals	0.995	0.973	0.770	0.860
Backchannels	0.975	0.697	0.588	0.638
Subordinate Clauses	0.964	0.863	0.831	0.847
Adv & Adj Expressing	0.901	0.731	0.706	0.847

Table 15: F1 performance of micro-level span annotation by GPT-4o

Models	Topic	Tone	Opening	Closing
BERT (raw dialogue)	0.414	0.401	0.414	0.379
GPT-4 (raw dialogue)	0.553	0.533	0.585	0.557
BERT+BERT (on annotated data)	0.987	0.990	0.993	0.978
BERT+BERT (based on BERT predicted micro-level features)	0.836	0.855	0.836	0.830
GPT-4+GPT-4 (on annotated data)	0.761	0.749	0.812	0.809

Table 16: F1 performance of Marco-level four interactivity aspects' score prediction across different model versions