Modelling retrospective evaluation of situational interdependence in conversations from its estimated real-time evaluation

Taichi Uno

Delft University of Technology

TU Delft

Modelling retrospective evaluation of situational interdependence in conversations from its estimated real-time evaluation

by

Taichi Uno

to obtain the degree of Master of Computer Science
at the Delft University of Technology,
to be defended publicly on Wednesday June 26th, 2024 at 10:00 AM.

Student number:	5056764	
Project duration:	November 1, 2023 – June 26, 2024	
Thesis committee:	Dr. B. Dudzik,	TU Delft, supervisor
	Dr. H. Hung,	TU Delft, supervisor
	Dr. C. R. M. M. Oertel,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

This Master's thesis marks the conclusion of my journey at Delft University of Technology. Over the past two years, I have had the privilege of immersing myself in a field that ignites my passion. The past nine months of working on this thesis have been challenging, but the support of those around me has made this accomplishment possible.

I am deeply grateful to my supervisors, Dr.Hayley Hung and Dr.Bernd Dudzik, for their invaluable guidance. Their insightful feedback and inspirational discussions during our weekly meetings were crucial to successfully completing this research. I also extend my appreciation to the members of the HOMI lab, who warmly welcomed me into their community. Their supportive environment was essential for sharing my thoughts and drawing inspiration from their research.

Finally, I extend sincere thanks to my family and friends, whose support has been a constant source of motivation throughout these past two years.

Taichi Uno
Delft, June 2024

Abstract

Understanding how users retrospectively evaluate their interactions with adaptive intelligent systems is crucial to improving their behaviours during interactions. Prior work has shown the potential to predict retrospective evaluations based on different real-time aspects of conversations, such as verbal cues and non-verbal behaviours. However, the relationship between how one retrospectively evaluates and the real-time evaluations in the moment of conversations remains unclear. This study investigates the relationship between real-time evaluations of a situation, using the Situational Interdependence Scale (SIS) framework, and its retrospective evaluations. We investigate the presence of the peak-end rule and a complex relationship that could be modelled using Long Short-Term Memory (LSTM) for each SIS dimension using the PACO dataset. Due to the absence of ground truth for real-time SIS evaluations, we also present a methodologically sound technical approach to utilize a Large Language Model (LLM) to estimate values for each SIS dimension for each spoken utterance in conversations. Analysis of the experiments revealed the absence of both the peak-end rule and an LSTM-modelled relationship across all dimensions of SIS. However, both types of models at least predict better than the average of the estimated real-time evaluation. This may be largely due to the inaccuracy of the estimated real-time SIS evaluations and the limited LLM's capability of labelling real-time SIS in conversational data. Future works may focus on improving the annotation of real-time SIS evaluations through human annotation or human-supervised few-shot learning of LLM, using other modalities in combinations with verbal content, and exploring other predictive models.

Contents

1

Introduction

How one remembers daily conversations or experiences is related to what happened during an event, but it has been found that this relationship could be complex [51]. This is because evaluating what happened at the moment of an event later is different from how one evaluates in real-time during an ongoing event due to temporal distances, which is the perceived time between the moment of evaluation and the moment of the stimulus events [9, 27]. In the scope of human-robot interactions, as interactive agents become more popular in our daily lives, one of their challenges is to accurately measure and improve the users' impressions towards such agents. The impressions here mean the retrospective evaluation of the interactions with the robots, which is also related to the evaluation of the robots themselves. Optimizing users' retrospective evaluation towards robots is important for adaptive robots in several aspects, for example, it shapes the future willingness to interact with agents [1, 6]. Traditionally, agents rely on explicit feedback from users, however, too frequent prompts for feedback can affect usability negatively [20]. If the systems can estimate users' retrospective evaluation of agents based on what happened during interactions, instead of directly asking, it allows them to adapt their behaviour from the estimated feedback accordingly. Additionally, not only in robot-human interactions but also in the field of psychology, it would be interesting to uncover the relationship between the feedback which is based on how one recollects past events and evaluate retrospectively, and the real-time experience, which could ultimately provide implications for human perceptions and cognitions. Investigating the relationship between real-time and retrospective evaluations in human-human conversations is an essential meaningful first step towards achieving this goal.

Under evolving situations, people are known to continuously and subjectively evaluate these changing situations. In the following, we refer to this as *real-time evaluation of a situation*. It is based on personal belief, relevance, and significance and people adjust their behaviour accordingly [22, 37], where such behaviours include non-verbal behaviours, such as body posture, facial expressions, or tone of voice, and verbal contents. Past research showed that triggers of behaviours of people in a situation lie in their ongoing internal evaluation of the current situation [22]. Its reverse is also considered true, which means that the real-time evaluation and the ongoing behaviours bidirectionally influence each other [35, 36]. Traditionally, psychologists attempted to conceptualize how individuals evaluate a situation and how a situation influences human behaviour from various perspectives and proposed frameworks such as emotional appraisal [36, 37], situation perception [29], and situational interdependence [16]. While *real-time evaluation of a situation* happens during the moment of an interaction, *the retrospective evaluation of a situation* happens after an interaction or a situation occurs, so there is a temporal distance between them, which changes the perception of past events [27]. To visually illustrate the real-time and retrospective evaluation of a situation, the schematic image of two people in a conversation is shown in Figure 1.1. While Person A and B are having a conversation, there is a real-time evaluation of the conversation by each of them which is affected by the change in the situation as time passes. On the other hand, the retrospective evaluation happens after an interaction, which is reflected and evaluated based on what happened during the moment of conversation.

There are several attempts at estimating how in real-time people evaluate situations by looking at

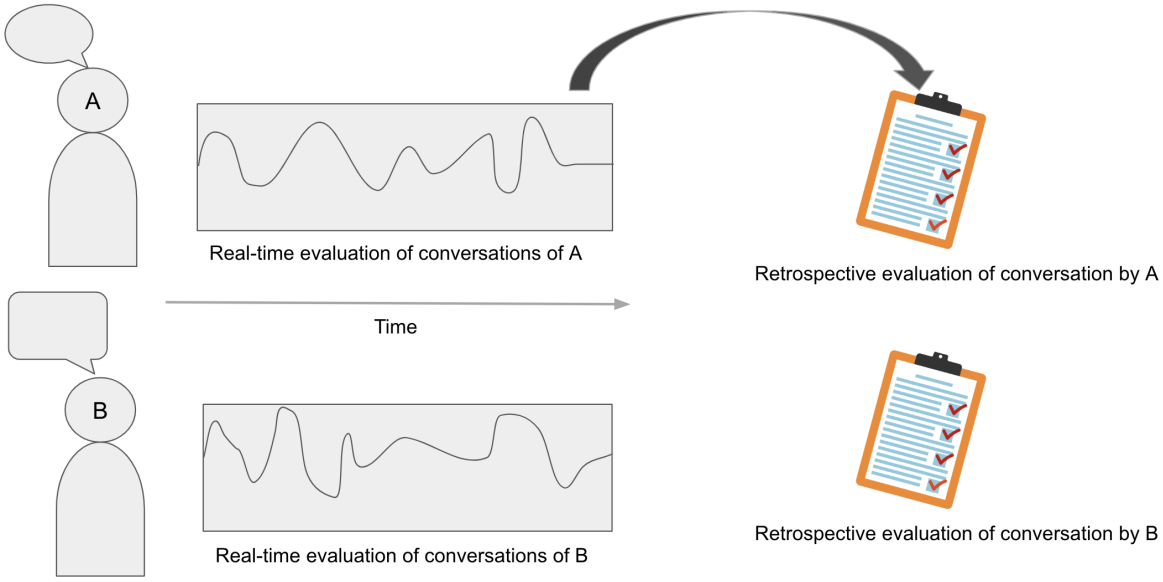


Figure 1.1: A schematic image of two people in a conversation. The grey box shows the real-time evaluation of the conversation, which changes as the change of situation as time passes. The retrospective evaluation of the conversation is likely to be linked with what happened and how one evaluated the situation during the conversation, which is symbolized with the black arrow on the top. However, this link between the two has not been established yet.

nonverbal behaviours [15] or verbal contents [54]. However, these studies did not look into how the real-time evaluation is related to the retrospective evaluation. Also, the link between the real-time behaviours at the moment and the retrospective evaluation is studied [10], but not the relationship with the real-time evaluation.

In the field of psychology, peak-end rule [14] is a well-known principle that explains the link between how one experiences the moment in real time and what is remembered. It is found to be generalized to other domains [2, 46], however, have not yet been investigated for the evaluation of a situation. In the scope of an attempt to link the retrospective evaluation and the real-time sentiment analysis, which can be seen as one kind of evaluation framework, a study has investigated predicting the customer satisfaction (CSAT) of a conversational agent from sentiment analysis of real-time verbal behaviour during users' conversations with the agent [20]. Although their studies' reported the successful performance of predicting CSAT, they did not investigate their relationship using the same evaluation framework for real-time evaluation and retrospective. Investigating a relationship between real-time and retrospective evaluations expands the possibility of identifying the factors lying in the moment of conversations influencing retrospective evaluations of situations.

One of the reasons for the lack of research in this relationship between the real-time evaluation of a situation and its retrospective evaluation would be that it is challenging to obtain accurate measures of the real-time evaluation of situations. For example, via self-report, the invasiveness of measurement tools cannot be ignored [33] as users have to report their evaluation of ongoing situations in real-time using some device, which would affect their evaluation of the situation. On the other hand, using third-party observation and annotation by examining observable signals, there is a limitation that what can be observed in terms of behaviours might not truly reflect the actual internal state of evaluation [25].

Despite these limitations in obtaining a real-time evaluation of a situation, LLM (Large language models) became gradually popular and recognized as a tool to augment the human annotation process [34, 57], which helps with obtaining annotated labels utilizing textual data. LLM has proven its capability in various natural language processing tasks [26], where verbal cues are one informative source to grasp the idea of human perceptions of a situation [37]. Specifically, in the scope of applying LLM to Computational Social Science, the use of LLM in a zero-shot manner, meaning without task-specific training, has demonstrated notable performance in various text labelling tasks, including utterance-wise sentiment analysis [57]. Although the reported study does not support the full replacement of human annotators by LLM, the idea of using LLM in estimating human perceptions is gradually gaining

recognition.

Acknowledging these research gaps and limitations, this research attempts to explore the relationship between the real-time and the retrospective evaluation of a situation, specifically in a conversation setting. To address the lack of ground truth for the real-time evaluation of a situation, we propose a methodologically sound approach that utilizes zero-shot learning LLM to estimate it from each utterance of the verbal contents of the conversation. This approach serves as a proxy for its actual real-time evaluation. Thus, the research question can be formulated as follows.

- To what extent is an individual's estimated real-time evaluation of situational interdependence during interaction indicative of their retrospective evaluation of situational interdependence?

We use the PACO dataset [24] that consists of video recordings of conversations between two people and questionnaire responses about the conversations. In their experiment, after a conversation, participants were asked to fill in a post-conversation questionnaire that included the retrospective evaluation of a situation in terms of SIS (Situational Interdependence Scale). SIS is proposed by Gerpott et al. [16], which consists of five dimensions to evaluate the situation and provides a way to quantify the perception of interdependence in a situation. Using this estimated real-time evaluation of SIS, the predictive models are implemented to explore the potential existence of relationships of 1) the peak-End rule and 2) a more complex relationship with its retrospective evaluation. Within the scope of this research, the contributions of this research are as follows.

- In order to tackle the challenge of obtaining real-time evaluation data, we present a technical approach to estimate the real-time evaluation of SIS from verbal contents of conversation using an LLM in a zero-shot manner.
- We present an investigation of exploring the relationship between the estimated real-time and the retrospective evaluation of SIS. More specifically, for each independent dimension of SIS, we investigate if the peak-end rule is present and if there is a more complex relationship that can be modelled using machine learning techniques.

2

Related Work

This chapter provides the insights from relevant literature. Firstly, frameworks of situation evaluation are explained to give a comprehensive overview of why it is interesting to research. In addition, Section 2.2 explains the prior works on using any real-time information, including real-time evaluation and other real-time information such as behaviours, to predict the retrospective evaluation of situations. Finally, the last section discusses the current state of the research on using LLMs to estimate the real-time evaluation of situations from verbal contents of conversations.

2.1. Situation evaluation frameworks

In the field of cognitive science and psychology, there are many theories about how people evaluate situations during interactions. This section is going to explore the implications of prior works on the evaluation of a situation and discuss different frameworks for situation evaluation.

Situation perception is a type of cognitive process that interprets and understands situations subjectively based on an individual's experiences, beliefs, and expectations [44]. The expressed emotions and behaviour can be indicative of how one perceives a situation [17, 18] as they are triggered based on the evaluation of the current situation and its reverse is also true [29]. Thus, the understanding of situation characteristics helps with not only just identifying the nature of situations, but also studying how people evaluate the situation.

Emotion is one of the examples of how the evaluation of a situation influences people's behaviour or internal states. Appraisal theory of emotion conceptualized how emotions arise as a result of evaluating situations [37]. Within this framework, appraisal refers to the evaluative process of situations based on personal beliefs, relevance, and significance. The Component Process Model (CPM) of appraisal inherits from the broader appraisal theory of emotion [35]. In this framework, combinations of specific configurations of different appraisal components (i.e. relevance, implications/consequences, coping potential, and norm compatibility) define emotions [33, 36]. It is worth noting that different psychologists may argue different dimensions of the appraisal theory of emotion. With such frameworks of appraisal-based emotion theories, it is possible to capture nuanced emotions, instead of distinct emotional labels like in Ekman's six emotions [12].

Lastly, it is known that the variations of evaluations occur not only among different individuals but also within the same individual over time, influenced by changes in perception of the situation, and the situation itself [9]. The temporal distance between the moment of evaluation and the moment of the stimulus event has been shown to impact the evaluation. [47]. However, the theories of evaluations of situations are currently vague in how time, specifically the real-time evaluation based on constantly evolving situations, plays a role, instead, they often focus on studying different dimensions of situation characteristics.

2.2. Predicting retrospective evaluation from real-time information

Research by Dudzik et al. [10] investigated recognizing retrospective perceived situational interdependence in face-to-face negotiations by exploiting real-time facial expressions, upper body behaviour and non-verbal vocal behaviour during interaction. They built a model based on the Ridge Classifier to analyse multivariate time series of those behavioural features. Their main discovery is that, out of the dimensions of SIS, people's real-time behaviour seems to predict the evaluation of conflict of interest and power, while the conversation partner's behaviour is for CI (conflict of interest), FI (future independence) and IC (information certainty). Also, this research did not explicitly take the temporal dynamics of behaviour into account, instead, they constructed aggregated real-time behavioural features by using ROCKET (Random Convolutional Kernel Transformation) [8], which might lack the direct implications of how temporal dynamics of real-time behavioural features have significance to the retrospective situational interdependence.

In the field of psychology, there is a widely known phenomenon called the "Peak-end rule" proposed by Fredrickson and Kahneman, which states that people tend to evaluate the overall experience of an event based on its peak and the end rather than based on the sum or weighted average of overall experience [19]. It also implies that the duration of the event does not have much influence on how one evaluates the situation later. Its applicability has been investigated in wider domains, such as affective arousal, pain and perception of discomfort [14, 19].

However, Strijbosch et al. criticises the peak-end rule has only been traditionally investigated in rather simpler experiences and argues that the peak-end rule may not be robust for complex and heterogeneous experiences [42]. They tested the existence of the peak-end rule by asking participants to report their emotional arousal when they are watching a VR movie, which can be seen as a more continuous stimulus event in contrast to Fredrickson and Kahneman where they tested emotional valence (pleasantness) with a set of distinct video clips. These findings suggest that the peak-end rule may not be robust for cognitively more complex experiences.

A study by Kim, Levy, and Liu has attempted to predict customer satisfaction (CSAT) of a conversational agent from the sentiments and contents of the conversations with the conversational agent [20]. They used the conversations which naturally occurred between users and conversational agents. They extracted two types of features for each utterance turn from the raw conversation audio, namely automatic and human-annotated sentiment analysis. The automatic annotation analyses its sentiment embedding based on acoustic and lexical cues automatically and the human annotation involves manually evaluating each utterance in terms of activation, valence and satisfaction. These sequences of utterance-wise annotations are inputted to two types of machine learning models, BLSTM (Birectional long-short-term memory) and ν-SVR (Support Vector Regression), where BLSTM can well-capture the temporal dynamics of input sequences but the other one is static in time. They reported the importance of exploiting utterance-wise features to predict CSAT instead of aggregated features of a conversation as utterance-wise ones scored higher in correlation with CSAT. The BLSTM models with automatic sentiment features have performed the best, implying the temporal dynamics within each conversation are important to estimate CSAT. Although automatic annotation has outperformed the human annotation input, it also shows high predicting performance, which suggests that the human-annotated sentiment analysis per sentence is informative in predicting retrospective evaluation. Similarity to our research lies in taking an approach to annotate the conversation per utterance in terms of some dimensional spaces which possibly reflects on users' real-time evaluation of a conversation and predicts the retrospective evaluation of conversations. However, they used different frameworks for annotating real-time and retrospective evaluation, whereas in our study we used the same framework, namely SIS.

2.3. Estimating Real-time Evaluation using LLMs

This section discusses the present research landscape on estimating real-time evaluations of situations, including conversations or interactions, with a specific focus on utilizing LLM (Large Language Model). Several studies have explored using real-time verbal content to predict the numerical values of different frameworks of retrospective evaluations. The previously mentioned study by Kim, Levy, and Liu[20] showed the capability of real-time conversation audio cues and their sentiments, during interactions in predicting customer satisfaction. Their study has manually annotated the conversation utterances in the dimensions of activation, valence and satisfaction. As these prior works have shown, the verbal

content of a conversation is one of the rich information sources to estimate the speaker's internal states of real-time evaluations of situations [52].

As it is costly to annotate the conversations to extract such features manually, LLMs are one of the alternative approaches that would augment human annotators in such feature extraction tasks in the Computational Social Science field because they can map the input sequences, in this case, natural languages, non-linearly that may resemble human cognition and reasoning by utilizing their processing and memory capability [34, 57]. Transformer architecture [48], which is the core of LLMs nowadays, enables learning long-range dependencies in sequential data efficiently, which makes it suitable for various natural language processing tasks. As technological advancement, LLM became more recognised in the field of computational social science to partially automate human annotations [57]. Several attempts were made previously to quantify the components of psychological frameworks, especially in emotional appraisal, using LLMs [5, 13, 43, 53, 54].

One of the examples is to predict dimensions of emotional appraisal with zero-shot learning using the CovidET dataset [54, 55], which consists of covid-related posts on Reddit, which can be considered an outcome of retrospective evaluation of experienced events as they were presumably composed after the events rather than during them. Similar approaches have been adopted to estimate how one evaluated the experienced event in terms of emotions from text descriptions of specific scenarios [5, 43, 53]. These prior studies have underscored the potential of LLMs in extracting dimensions of a variant of the retrospective evaluation of a situation from textual data. However, to the best of our knowledge, there have been no attempts so far to be used in the context of situational interdependence so far.

One notable study conducted by Feng et al. [13] focused on estimating real-time emotions in the conversational setting by analyzing each utterance to identify emotions within a conversational context using a set of emotional labels. The study found that, compared to state-of-the-art supervised approaches, LLMs exhibited lower performance in zero-shot learning scenarios. However, performance notably improved with instruction-following demonstrations, indicating the effective utilization of prompts is essential. It underscores their greater generalizability to other natural language processing tasks and robustness to errors in automatic speech recognition. Overall, this study has shown the potential of employing LLMs within conversation settings to infer one's perception of a situation from conversational text data.

Regarding the criticism toward using LLM for the annotation process, it is reported that LLMs struggle with multiple-choice questions (MCQs) due to their bias in selecting options based on their positions [31, 56], which makes it not suitable for simulating annotation by selecting options. In addition, it is also found that LLM also has a bias in outputting numerical values, where they have a "favourite" number when outputting numbers [40]. It would be worth investigating in our research how the performance of LLM would change due to the format of MCQs in the prompt.

3

Approach

This chapter provides a brief overview of the experiment pipeline, including 1) the explanation of the dataset and its applicability and implications to this research, and 2) outlining the entire experiment setup to achieve our research goal.

3.1. Dataset

In this research, the PACO dataset is used [24]. Although it was collected for different purposes, it provides all the relevant information to our research aims, which is namely 1)video recordings of two participants having conversations and 2) the results of a post-questionnaire including self-reported retrospective evaluation of SIS for each dimension.

The PACO dataset was initially developed to model partner selection and explore its relationship with human impressions during social interactions [24]. The dataset includes recordings of 3-minute online conversations between two individuals over two different settings where people have different expectations towards their conversation partners. The settings include performing 1) a joint trust task (TRUST), which is a cooperative decision-making task with mixed motives, and 2) a joint competence task (COMP), where the outcome depends on the competence of two individuals. Half of the participants were selected for performing the joint trust task and the other for the joint competence task. Depending on the task the participants were assigned, the goal of the conversations would be different. For the TRUST task, the participant in the conversation is looking for someone who seems more warm, while those assigned to the COMP task try to find someone with higher competence.

Each participant was asked to have one-on-one online video conversations to find a suitable partner for one of the two tasks with 3-5 other participants and to fill out a questionnaire including a 10-item version of the Situational Interdependence Scale (SIS) after each conversation. As mentioned earlier, SIS proposed by Gerpott et al. [16] provides a way to quantify and reason the perceived interdependence of people in a situation. This scale is useful as it is known from previous research that people can recognize how their behaviours affect their own or interaction partners' outcomes and change their behaviour accordingly based on the real-time evaluation of situational interdependence at any given time [16]. This framework consists of the five dimensions of SIS as in Table 3.1. Two questions per dimension were asked, and participants had to rate each question/statement in terms of a 1-5 Likert scale. Some questions are reversed, which means that a higher value in the answer means a lower value for the degree of a dimension. The average score of the scores from two answers, reversed if necessary, is used as the value for each dimension.

The distribution of the Likert scale for each dimension of the retrospective evaluation is shown in Figure 3.1. As shown in this figure, for all dimensions, the Likert-scale distributions are imbalanced, which could cause biased predictive models. These imbalances may be due to the contextual settings of the conversations, more specifically the goal of the conversation. The goal of the participants during conversations is to figure out if the person they are talking to right now is suitable for the tasks carried out in the next stage of the experiment, and the same goes for their conversation partners. For that

Dimension	Description
Mutual Dependence	Degree of how much each person's outcomes are determined by how each person behaves in that situation
Conflict of Interest	Degree to which the behaviour that results in the best outcome for one individual results in the worst outcome for the other
Future Independence	Degree to which own and other's behaviour in the present situation can affect own and other's behaviour and outcomes in future interactions
Information Certainty	Degree to which a person knows their partner's preferred outcomes and how each person's actions influence each other's outcomes
Power	Degree to which an individual determines their own and others' outcomes, while others do not influence their own outcome

Table 3.1: Dimensions of situational interdependence and their definitions in short from Gerpott et al.[16]

reason, since both parties in a conversation share the same goal, it may be less likely to have CI between the two people, resulting in relatively lower scores as shown in 3.1b. Also, for the dimension of P, it may be less likely that one of the two who are in a conversation will dominate the conversation as the contextual setting of the experiment does not impose power imbalance, resulting in its peek in the middle (Likert scale of 3) indicating equal balance of power in Figure 3.1e. Similarly, the tendency of middle or higher values in the dimension of MD and IC as shown in Figure 3.1a could be that both people in a conversation are aware of how they behave in a situation can have an impact on the impression of their conversation partner. On the other hand, not everyone is perceived to be a suitable partner to perform the task or understand the other's expectation toward oneself accurately, leading to lower FI and IC, resulting in slightly lower peaks in contrast to MD as observed in Figure 3.1c and 3.1d.

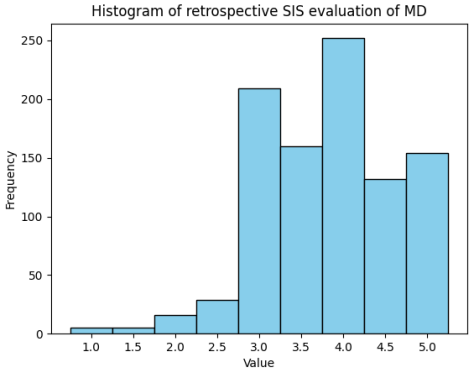
The videos of conversations in the PACO dataset have two kinds of recordings from two perspectives. Each video only captures audio-visual signals from one person's point of view, which is saved locally and on the cloud. This setup implies that the participant was watching the cloud videos of his/her conversation partners while interacting with each other. In this research, both people's perspectives of the local recordings in a conversation are used as it was observed that they contain higher-quality audio from the manual inspection. The audio is extracted from these video recordings and merged into one audio file, which contains audio of both parties in a conversation. This file is then passed to the audio transcription process, which will be detailed in Section 4.1.

3.2. Experimental setup

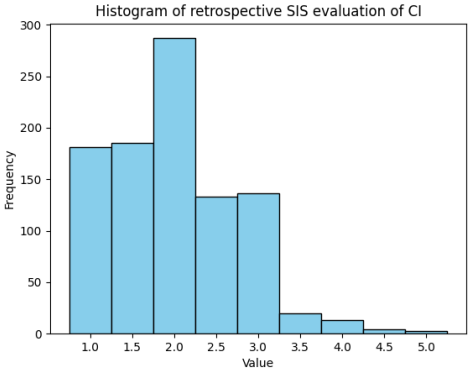
The experiment to test our hypotheses involves the following two steps;

1. Estimating the real-time evaluation of SIS from verbal contents of conversations.
2. Creating predictive models to map the real-time evaluation of SIS to the retrospective evaluation.

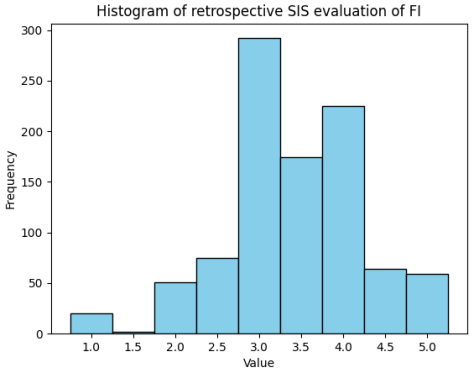
The overall pipeline consisting of these two steps is visualized in Figure 3.2. Firstly, conversation audio files are transcribed with speaker labels in time order. The transcription is then processed to generate the real-time evaluation of SIS. LLM is used here to generalize this label for each utterance of a speaker, where the input prompts ask to act as a person in a conversation and estimate each dimension of SIS by filling the same questionnaire as what participants completed, which we call this output the *estimated real-time evaluation of SIS*. Each dimension of this output from LLM will be the input for the predictive models, generating predicted values for each dimension of the *retrospective evaluation of SIS*. The models are trained using the dataset which contains the retrospective evaluation of SIS of speakers. Finally, the predicted outputs are compared with actual data and further analysed to test the potential existence of hypothesised relationships, which will be present later. The detailed description of these steps is explained in Chapter 4.



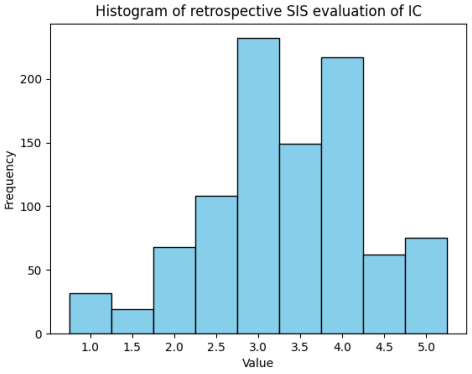
(a) MD (Mutual Dependence)



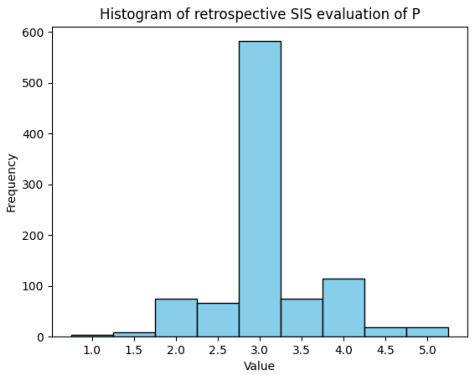
(b) CI (Conflict of Interest)



(c) FI (Future Interdependence)



(d) IC (Information Certainty)



(e) P (Power)

Figure 3.1: Histograms of Likert-scale distributions of the retrospective evaluation for each dimension of SIS

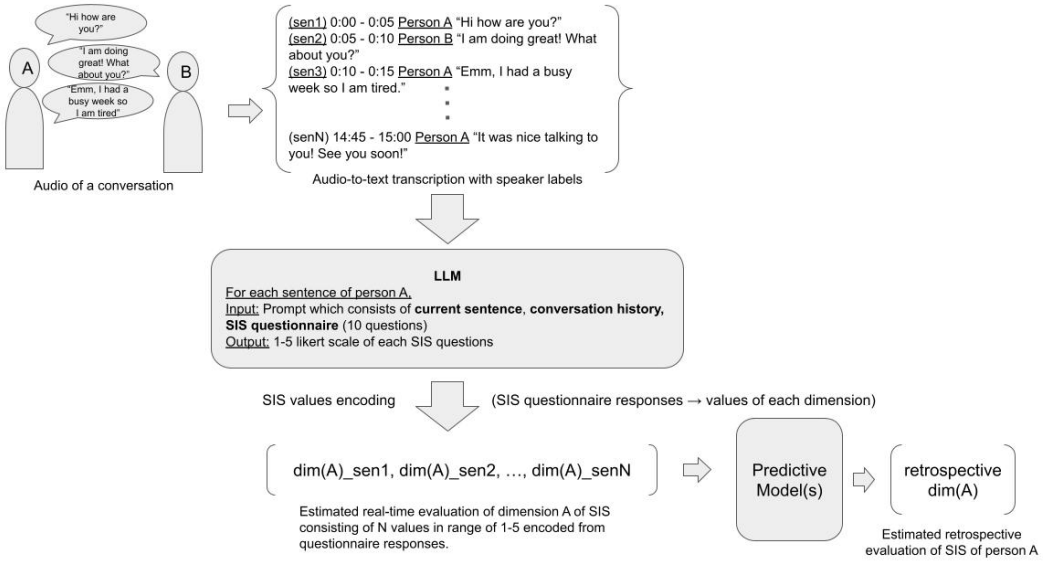


Figure 3.2: Schematic Representation of the pipeline

4

Methodology

4.1. Estimation of real-time evaluation

4.1.1. LLM setup

To estimate the real-time evaluation of SIS, a Large Language Model (LLM) is used to analyze the spoken content of conversations and generate estimated values for each dimension of SIS. The input of the LLM will be a prompt which contains 1) an utterance from a speaker at a time, 2) the conversation history, and 3) a set of questions to answer each time to output values for each dimension of SIS. This process is repeated from the first utterance until the last utterance of the speaker in a conversation. Due to the lack of ground truth data of the real-time evaluation of SIS, the model is used in a zero-shot manner. It is a way to utilize LLMs without specific training or fine-tuning their parameters for a specific task, which is in this case, labelling each dimension of SIS at a given sentence in a conversation. This is possible because of the high generalization ability of LLMs which is backed up by its huge training set and LLM is known to perform well in such unseen natural language processing tasks [39]. In this research, Llama2-chat[45] is used. It is a state-of-the-art LLM that is pre-trained and fine-tuned with 7 to 70 billion parameters optimized particularly for dialogue applications. This model has demonstrated promising zero-shot performance across various natural language generation and processing tasks [45]. The adaptability of Llama2-chat to unseen tasks makes it an ideal tool for our research. Given the limited time and resources, this study uses a 7 billion parameter configuration.

4.1.2. Preprocessing

The first step of the pipeline is to transcribe the conversation audio. Whisper X [3], a fast automatic speech recognition system with word-level time stamps and speaker diarization, is used to produce the transcriptions of the conversation video clips. While transcription generates what has been said, speaker diarization is a process of segmenting a conversation audio clip based on who is speaking. The specific model was the ideal tool for this study because it is based on the state-of-the-art automatic speech recognition model which offers promising performance and it detects turn-taking with speaker diarization as it is crucial in our setup to recognize the sequence of who spoke what when [3].

4.1.3. Prompt structure

Prompts are important for guiding LLM responses and ensuring the quality of the output. This research drew inspiration from the design used by Feng et al.[13], whose work is relevant as it focused on estimating emotional appraisal for each new utterance within the conversation settings. This approach makes use of the conversation history as the context for future queries, reflecting the dynamic nature of real conversation settings where responses might be influenced by multiple preceding sentences rather than just the current one. While Feng et al.[13] also used in-context learning, we could not implement it due to the absence of ground truth labels for real-time evaluation of SIS in the PACO dataset. Additionally, the techniques described in a study by White et al.[50] about creating effective prompts are utilized throughout our prompt. The following patterns are used to create the prompt; "meta language creation pattern", which clarifies the use of languages within a prompt, "persona pattern", which tells

the LLM to simulate a certain persona, "template pattern", which constrains the output format for better post-process, and "context manager pattern", which specifies the context to consider when producing output.

The design of the prompt used in our study is shown in Table 4.1. There are mainly two parts, Task Definition and Query. Task Definition consists of informing the LLM to act as the person named *ID of the speaker* in a conversation and providing the context that it is a conversational setting with one conversation partner whose name is *ID of the conversation partner* (persona pattern and context manager pattern). In addition, *context* specifics their conversation objectives. In the PACO dataset experiment, half of the participants were assigned to carry out a task in the future where conversation partners' competence is important, while the other half did a task where conversation partners' warmth is important.

Next, the description of the objective of the task is presented, namely to answer 10 questions, which will later be provided in the Query part, at the moment of a sentence by the speaker. As following the approach of the aforementioned study by Feng et al.[13], the conversation history up until the sentence of interest is included in the prompt as *hisotry* (context manager pattern) and the spoken utterance by the speaker of interest. Lastly, since the last part of the Query part describes the conversation partner as "person X", the last sentence of the Task Definition part explicitly clarifies that person X is indeed the conversation partner (meta language creation pattern). We create this prompt for each spoken utterance of the speaker of interest. For example, for a conversation of Person A with his conversation partner, when creating the real-time evaluation of SIS of Person A in a conversation, the prompts are created for each spoken utterance by Person A while the conversation history contains the dialogue between Person A and his conversation partner. This is because verbal content as a behaviour reflects upon the internal evaluation of the person as supported by prior works as mentioned before [35]. It is worth noting that it is also true that what the conversation partner said also influences the evaluation but in this experiment, we do not evaluate the utterance of the conversation partner when evaluating the SIS of Person A.

The Query part contains the questions that the LLM has to answer to get the real-time evaluation of SIS values for each dimension. To keep the consistency with the retrospective evaluation of SIS, the same set of questions and instruction scripts are provided as in the data collection of the PACO dataset [24]. This set of questions is the scaled version of the original SIS questionnaire by Gerpott et al.[16], which contains two questions per dimension, where one of them is a reverse coded question. The LLM selects a textual label from the five listed options for each question to report its answer. The value for each dimension of SIS is calculated by averaging the two corresponding answer values on a scale of 1 ("strongly disagree" or "definitely person X") to 5 ("strongly agree" or "definitely myself"), and it is flipped for the reverse coded questions so 1 ("strongly agree" or "definitely myself") to 5 ("strongly disagree" or "definitely person X"). Following these 10 questions with instructions, the prompt ensures the output format for easier post-processing of the data (template pattern).

The current prompt structure is the result of several iterations of tries and errors of prompt engineering as described in Appendix A. This appendix section discusses the previous approaches to estimate real-time evaluation using different prompts alongside their limitations and considerations that led to this final design and its context.

4.2. Predictive models

All of the implemented predictive models take a dimension of the estimated real-time evaluation of SIS as its input and produce the dimension of the estimated retrospective evaluation of SIS in the range of 1 to 5. They are implemented to test the following hypothesis. Firstly, for each of the dimensions of SIS, the "peak-end rule" is present. Fredrickson and Kahneman [14] argues that the average of the affective values of its peak and the end of an episode can be used to approximate the global evaluation of the entire episode, which is known as the "peak-end rule". It is also found that this rule is observable in other domains than affective values [21], suggesting the potential similar correlation in situational interdependence. Another hypothesis is that for each dimension of SIS, beyond the peak-end rule, there exists a more complex relationship between the real-time and the retrospective evaluation. Such a relationship could be modelled using machine learning models that capture the dependency within the sequential input data and output the retrospective evaluation. For this purpose, the long short-term

Component	Prompt body
Task Definition	Act as Person *{ID of the speaker}*. You are now having a conversation with Person *{ID of the conversation partner}* to find a partner to carry out a task with in the future where your partner's *{context}* is important. You are asked to answer the following 10 questions at the moment you said *{sentence}* given this conversation history. *{history}*. From now on, person X means your conversation partner, Person *{ID of the conversation partner}*.
Query	Here, you are asked to rate the interaction you just took part in. We are interested in your personal (subjective) impression of the situation. Thus, we ask you to be as honest as possible and describe the situation by using the following scale: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly Agree 1. What each of us does in this situation affects the other. 2. Our preferred outcomes in this situation are conflicting. 3. How we behave now will have consequences for future outcomes. 4. We both know what the other wants. 5. Whatever each of us does in this situation, our actions will not affect the other's outcome. 6. We can both obtain our preferred outcomes. 7. Our future interactions are not affected by the outcomes of this situation. 8. I don't think the other knows what I want. For each item, please think of the same conversation and indicate how the following statements describe the specific situation. Definitely person X, Maybe person X, Neither person X nor myself, Maybe myself, Definitely myself 9. Who do you feel had more power to determine their own outcomes in this situation? 10. Who has the least amount of influence on the outcomes of this situation? Use the template to answer in JSON format. You do not need to provide explanation. {"Q1" : SCALE, "Q2" : SCALE, "Q3" : SCALE, ... , "Q10" : SCALE}

Table 4.1: A template of the prompt for estimating real-time evaluation of SIS at a given sentence given the conversation history

memory (LSTM) network is implemented to test this hypothesis because this architecture is known to be efficient in learning sequential data dependency compared to traditional recurrent neural networks (RNNs).

To test the hypothesis of the peak-end rule, two types of peak-end rule models (*peak_end*) are implemented. The rule states that the overall experience can be modelled by averaging its peak values and the end [19]. In this study, we implement the peak-end rule for each dimension and take the maximum value as "peak" and the value of the last sentence by a speaker as "end", and the average of these two values is returned based. It can be formulated as,

$$Y_{\text{pred}} = \frac{\max(X) + X_n}{2}$$

, where Y_{pred} represents the predictive output of the peak-end model, and X is the input vector with its length n that contains the values of estimated real-time evaluation for each dimension of SIS. Also, following the approach of Trofymchuk, Liz, and Trofimchuk [46], a linear regression model that learns the weight of the peak and the end is also implemented (*peak_end_reg*). This can be expressed as,

$$Y_{\text{pred}} = w_{\max} * max(X) + w_{\text{end}} * X_n + b$$

, where $w_{\max}$ and w_{end} represents the weight for the peak and the end values respectively and b is the intercept which is also learned during training. To test the significance of the combination of the peak and the end, two additional models were also implemented where one only outputs the peak (*peak_only*),

$$Y_{pred} = max(X)$$

and the other only outputs the end (*end_only*)

$$Y_{pred} = X_n$$

.

To test the hypothesis of the complex relationship, the Long Short-Term Memory network (LSTM) is chosen. The LSTM's ability to capture long dependencies over time is well-suited for our task, given that the input is sequential data of real-time evaluation per sentence. To tackle the problem of unequal lengths of the input sequence, two types of LSTM, *lstm_pad* and *lstm_length_varying* are implemented. Both these two models account for the different number of spoken sentences per person and conversation. The first model (*lstm_pad*) pads the input sequence to match the lengths with the longest sequence so that it can be input into the model. While the padding techniques are widely used to solve the issue of unequal lengths sequences in many machine learning tasks, the problem with this approach would be, for example, that the network might misinterpret the padded sequence and also the padded sequence does not hold semantic meaning as they are not the actual meaningful input information. This can be a cause of a biased network. To mitigate such problems with padding, the other model is trained without padding the sequence (*lstm_length_varying*). This model is implemented by grouping and batching the input sequences according to their lengths, also called bucketing, which means that each training batch only contains the sequences with the same lengths. However, this approach might struggle with the performance as the sequence lengths are unequally distributed, which could harm the performance or make the training process inefficient. For each iteration of training of *lstm_pad* model, the number of hidden states and the number of epochs are tuned using the techniques of hyperparameter tuning. These LSTM models are implemented with Tensorflow [23]. For *lstm_length_varying* model, the hidden states of 16, and the 10 epochs are chosen as hyperparameters.

Finally, the dummy model (*dummy*) is implemented, which returns the average value of the retrospective SIS of the training set as its predicted value. This dummy model serves as the null hypothesis, which helps evaluate if the two models above show improvements and their significance to test the hypotheses and whether the hypothesized relationships are present. Additionally, a baseline model (*base_line*) is also implemented, where it takes the mean of the input sequence.

4.3. Evaluation

To compare these three models, the performance of each model for each dimension is measured in terms of R^2, which is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_{pred})^2}{\sum_{i=1}^{n}(y_i - \mu)^2}$$

, where indicates the number of observations, y_{pred} refers to predicted value and y_i is its ground truth, and μ is the class average of a dimension. R^2 indicates how the model explains the variability in the target variable. $R^2 = 0$ implies that the performance is the same as the class average, meaning it does not make prediction better than *dummy*. A higher value of R^2, where its maximum value is 1, suggests a better fit of the model, meaning the model explains well the variability in the target variable. R^2 is the best fit to test our hypotheses because our interest lies in whether or what type of relationship there is between the real-time evaluation and the retrospective evaluation of SIS.

Furthermore, to test if a model has made a statistically significant improvement, paired t-tests are carried out. It checks if the model has made a statistically significant difference given the mean and the variance of its performance compared to the comparison target model. In other words, it provides us meaningful insight into whether the performance improvement is by a random chance or due to the better fit of the model.

In this study, Welch's t-test is applied as we do not assume that the average performance scores do not have the same variance [49] for different models. The underlying assumptions of Welch's t-test are 1) the observations are independent and 2) residuals are normally distributed.

For the first assumption of independence, each participant engages in three conversations with three different people but each conversation is treated as one data point. There could be personal bias in how one evaluates SIS in real-time and retrospectively so not all data points are independent. Also, the participants were assigned to two different contextual settings, where the contextual setting itself could have an impact, which makes it not independent. However, in this experiment, these are ignored as we carry out 10 times 10 iterations of cross-validation as detailed later, which makes 100 different observations with different test-training splits, which we believe is large enough to discard the assumption. Regarding the second assumption of Welch's t-test, for each model in pair t-tests, the Shapiro-Wilk test is performed to test the normality [38].

The p-values less than 0.05 indicate that the variance of the result is not because of random chance. The t-statistics value can be positive and negative, where the positive one suggests the model has resulted in better prediction performance, while the negative one indicates that the model performed worse than the comparison target. To test the first hypothesis, *peak_end* and *peak_end_reg* models are compared against *dummy* model. On the other hand, for the second hypothesis, *lstm_pad* and *lstm_length_varying* are compared with *dummy* model. The performance of each model is measured by 10 iterations of 10-fold cross-validation, meaning 100 observations in total, with random shuffling as these hyperparameters of cross-validations have been found to have high replicability of the result [4], where high replicability of evaluation contributes to the higher reproducibility of the research. Here, randomly stuffing means that we do not explicitly divide based on the participants or contextual settings.

5 Result

5.1. Estimating the estimated real-time evaluation of SIS

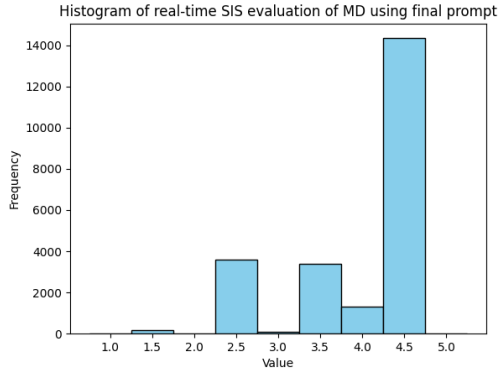
Out of 888 outputs of estimated real-time evaluation from the LLM, 766 of them have outputted in the correct format, which is then processed to calculate the value of each dimension of SIS. Figure 5.1 shows the distributions of labels of the estimated real-time evaluation for each dimension. For the dimension of MD, the sharp peak is at 4.5 as in Figure 5.1a. Similarly, for CI, most labels are distributed in 2.5 shown in Figure 5.1b. For the dimension of FI, it showed its strong peak at 3.5. Dimension of IC and P showed the spread in the labels compared to the rest of the dimensions. The dimension of IC has its peak at 4.0 but there are also values of 3.5 and 3.0. For the dimension of P, it is shown the most varying labels ranging from 1.5 to 4.0, but interestingly the values of 2.5 have occurred much less than its neighbouring bins of 2.0 and 3.0. These imbalanced distributions of real-time evaluation of SIS could be a cause of biased predictive models. Due to the absence of the ground truth of the estimated real-time evaluation, it is not possible to test the accuracy of how much the estimation reflects the actual real-time evaluation of SIS by participants in the moment of conversations.

5.2. Predicting the retrospective from the estimated real-time evaluation of SIS

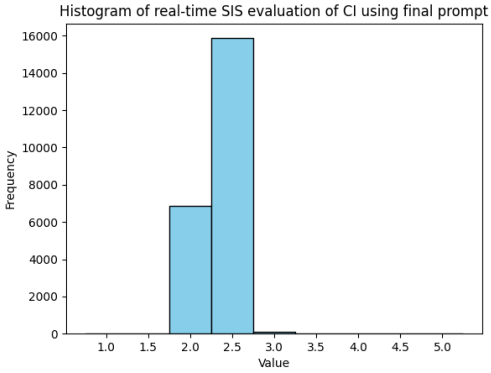
Given the correctly formatted estimated real-time evaluation of SIS (766 outputs), the performance of 10 times randomly resampled 10-fold cross-validation is shown in Table 5.1. Their visualization as an error-bar plot to illustrate the differences in performance and their variance across different models are presented in Figure 5.2. Their x-axis shows the names of the predictive models described in Section 4.2 for all figures.

Overall, for all models across all dimensions, average R^2 has resulted in negative values, suggesting a low degree of the model fit and that the performance was worse than the simple class average of each dimension. In addition, as it is visible in Figure 5.2, some of the models showed high variance. For heuristic models (*peak_end*, *peak_only* and *end_only*) where they output constant values based on the estimated real-time evaluations, it is most likely because of imbalanced estimated real-time evaluation data and the retrospective evaluation as these three models do not learn and adjust the outputs from the data. *lstm_length_varying* in dimensions of MD and FI showed relatively high variance compared to those in the other three dimensions, which could imply that the models in MD and FI failed to capture the underlying pattern.

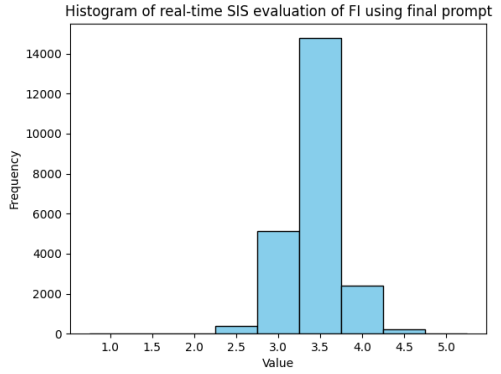
In the scope of peak-end rule models (*peak_end*, *peak_only*, *end_only* and *peak_end_reg*), *peak_end*, *peak_only* and *end_only* models have high variance across all dimensions compared to *peak_end_reg*. Noticeably, in the dimension of P, *peak_only* and *end_only* have performed significantly worse than in the other dimensions, probably because the estimated real-time evaluation of dimension P was varied compared to the other dimensions. For all dimensions, *peak_end_reg* outperformed the rest of peak-end rule models, suggesting that weights of the peak, the end, and the intercept term play important roles in predicting



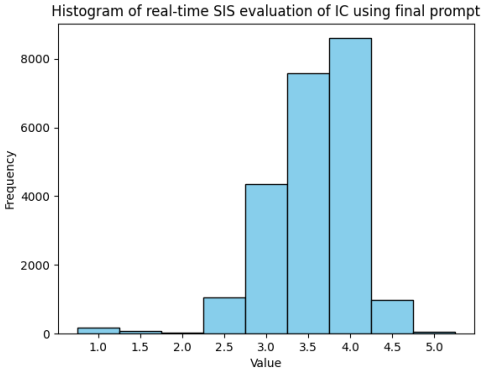
(a) MD (Mutual Dependence)



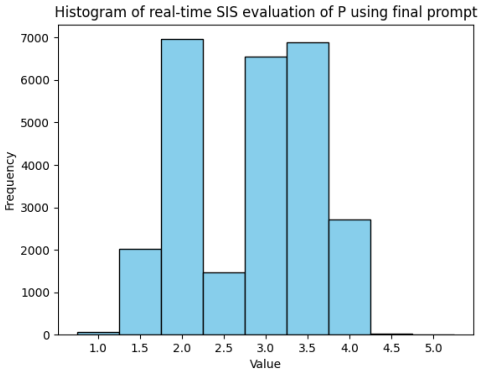
(b) CI (Conflict of Interest)



(c) FI (Future Interdependence)



(d) IC (Information Certainty)



(e) P (Power)

Figure 5.1: Histograms of distributions of estimated real-time evaluation for each dimension of SIS

Model	r2_mean	r2_std
peak_end_reg	-0.01283	0.03058
peak_end	-0.50435	0.13572
peak_only	-0.76691	0.25914
end_only	-0.88505	0.24135
lstm_pad	-0.03214	0.04543
lstm_length_varying	-0.23561	0.25796
base_line	-0.24333	0.09122
dummy	-0.01425	0.02170

(a) MD (Mutual Dependence)

Model	r2_mean	r2_std
peak_end_reg	-0.01359	0.02532
peak_end	-0.50959	0.24549
peak_only	-0.70497	0.22669
end_only	-0.39587	0.19802
lstm_pad	-0.04676	0.09348
lstm_length_varying	-0.10529	0.07974
base_line	-0.25756	0.11370
dummy	-0.01574	0.02065

(b) CI (Conflict of Interest)

Model	r2_mean	r2_std
peak_end_reg	-0.01840	0.02371
peak_end	-0.17407	0.08140
peak_only	-0.59842	0.17525
end_only	-0.27332	0.12240
lstm_pad	-0.02980	0.03786
lstm_length_varying	-0.56738	0.27748
base_line	-0.03993	0.03502
dummy	-0.01517	0.02138

(c) FI (Future Interdependence)

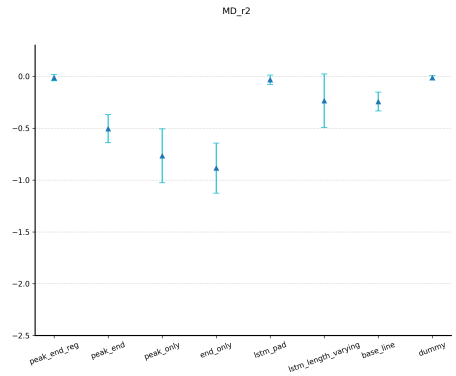
Model	r2_mean	r2_std
peak_end_reg	-0.01957	0.02365
peak_end	-0.51263	0.16376
peak_only	-1.18541	0.28657
end_only	-0.50324	0.17668
lstm_pad	-0.02466	0.03606
lstm_length_varying	-0.18577	0.12899
base_line	-0.11839	0.07251
dummy	-0.01125	0.01236

(d) IC (Information Certainty)

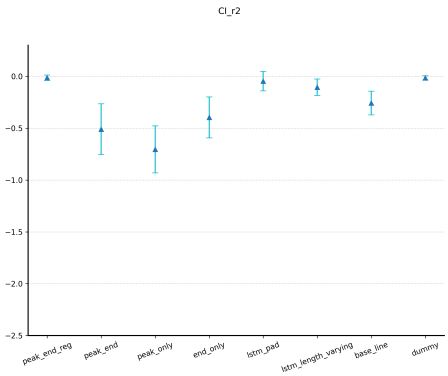
Model	r2_mean	r2_std
peak_end_reg	-0.01716	0.03482
peak_end	-0.50511	0.14838
peak_only	-1.85003	0.51372
end_only	-1.53530	0.41435
lstm_pad	-0.03480	0.04932
lstm_length_varying	-0.11302	0.09406
base_line	-0.46190	0.16043
dummy	-0.01311	0.01744

(e) P (Power)

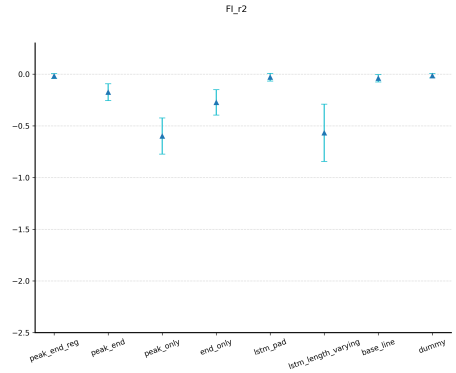
Table 5.1: Comparison of R^2 for each dimension of SIS across different models.



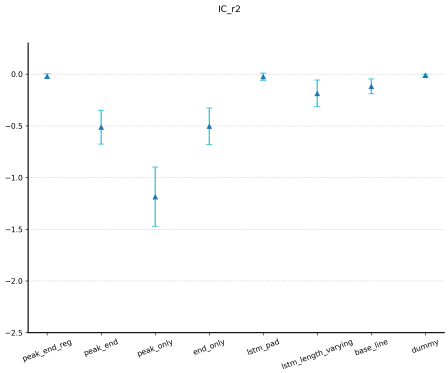
(a) MD (Mutual Dependence)



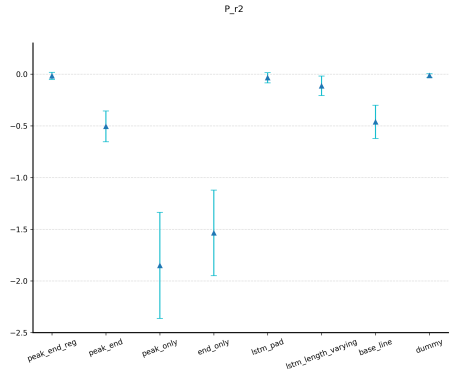
(b) CI (Conflict of Interest)



(c) FI (Future Interdependence)



(d) IC (Information Certainty)



(e) P (Power)

Figure 5.2: The error bar plot of R^2 for each dimension of SIS across different models (from the left, *peak_end*, *peak_end_reg*, *peak_only*, *end_only*, *lstm_pad*, *lstm_length_varying*, *base_line*, and *dummy*). The point shows the average and the error bar shows the standard error.

the retrospective evaluation from the estimated real-time evaluation.

Regarding LSTM models, *lstm_length_varying* models in the dimensions of MD and FI have shown high variance in their performance as mentioned earlier. *lstm_length_varying* models in CI and P have outperformed *base_line* while they did not in the rest of the dimension. On the other hand, *lstm_pad* has resulted in comparable performance as *dummy* and *peak_end_reg* for all dimensions and outperformed *base_line*. Lastly, *lstm_pad* has outperformed *lstm_length_varying* in all dimensions, suggesting that for predicting retrospective evaluation of SIS for all dimensions from estimated real-time evaluation of SIS, it is better to pad the input sequences, instead training in batches with the same length with the current configuration of LSTM. This could be because the imbalanced variation of the conversation lengths was high, which makes the training process ineffective by learning from the inputs with the same lengths.

In order to illustrate the significance of the predictive models, the results for t-tests are shown in Table 5.2. The normality of the results of these selected models is validated by the Shapiro-Wilk test as shown in Table B.1. The comparison of *peak_end* and *peak_end_reg* against *dummy* and *base_line* have been measured to test our first hypothesis of whether the peak-end rule is present in the relationship between the (estimated) real-time evaluation of SIS and its retrospective evaluation. In addition, the effects of ablations were also tested by comparing *peak_end* against *peak_only* and *end_only*. Next to that, to test our second hypothesis of the more complex relationship between the real-time and the retrospective evaluation of SIS, we conducted the t-tests of *lstm_pad* and *lstm_length_varying* against *dummy* and *base_line*.

Model 1	Model 2	p-value	t-statistics
peak_end	dummy	0.00000	-35.48038
peak_end_reg	dummy	0.70828	0.37477
peak_end	base_line	0.00000	-15.88240
peak_end_reg	base_line	0.00000	23.83804
peak_end	end_only	0.00000	13.67978
peak_end	peak_only	0.00000	8.93047
lstm_pad	dummy	0.00055	-3.53711
lstm_length_varying	dummy	0.00000	-8.50826
lstm_pad	base_line	0.00000	20.61939
lstm_length_varying	base_line	0.77942	0.28069

(a) MD (Mutual Dependence)

Model 1	Model 2	p-value	t-statistics
peak_end	dummy	0.00000	-19.94565
peak_end_reg	dummy	0.51444	0.65316
peak_end	base_line	0.00000	-9.26912
peak_end_reg	base_line	0.00000	20.83891
peak_end	end_only	0.00043	-3.58743
peak_end	peak_only	0.00000	5.81812
lstm_pad	dummy	0.00167	-3.22405
lstm_length_varying	dummy	0.00000	-10.81755
lstm_pad	base_line	0.00000	14.24921
lstm_length_varying	base_line	0.00000	10.90916

(b) CI (Conflict of Interest)

Model 1	Model 2	p-value	t-statistics
peak_end	dummy	0.00000	-18.78667
peak_end_reg	dummy	0.31492	-1.00753
peak_end	base_line	0.00000	-15.06242
peak_end_reg	base_line	0.00000	5.06409
peak_end	end_only	0.00000	6.71837
peak_end	peak_only	0.00000	21.85053
lstm_pad	dummy	0.00102	-3.34712
lstm_length_varying	dummy	0.00000	-19.74260
lstm_pad	base_line	0.05208	1.95431
lstm_length_varying	base_line	0.00000	-18.76448

(c) FI (Future Interdependence)

Model 1	Model 2	p-value	t-statistics
peak_end	dummy	0.00000	-30.37803
peak_end_reg	dummy	0.00229	-3.10370
peak_end	base_line	0.00000	-21.90320
peak_end_reg	base_line	0.00000	12.89158
peak_end	end_only	0.69832	-0.38816
peak_end	peak_only	0.00000	20.28149
lstm_pad	dummy	0.00065	-3.50058
lstm_length_varying	dummy	0.00000	-13.40111
lstm_pad	base_line	0.00000	11.51670
lstm_length_varying	base_line	0.00001	-4.53098

(d) IC (Information Certainty)

Model 1	Model 2	p-value	t-statistics
peak_end	dummy	0.00000	-32.76564
peak_end_reg	dummy	0.30245	-1.03486
peak_end	base_line	0.05055	-1.96732
peak_end_reg	base_line	0.00000	26.95479
peak_end	end_only	0.00000	23.28981
peak_end	peak_only	0.00000	25.02560
lstm_pad	dummy	0.00007	-4.12469
lstm_length_varying	dummy	0.00000	-10.39117
lstm_pad	base_line	0.00000	25.31933
lstm_length_varying	base_line	0.00000	18.66586

(e) P (Power)

Table 5.2: T-test results of the predictive Model 1 against Model 2 for each dimension of SIS.

Regarding the t-tests of peak-end rule models, for all dimensions, *peak_end* for all dimensions did

not result in positive statistical significance against *dummy* and *base_line*, meaning *peak_end* did not outperform *dummy* and *base_line* in all dimensions. The highest performance in comparison to those baseline models was achieved in the dimension of P, where it only showed negative statistical significance against *dummy* and comparable performance with *base_line*. Also, *peak_end_reg* against *dummy* did not show statistical significance for all dimensions except for IC, where it showed negative statistical significance. This indicates that the performance of *peak_end_reg* and *dummy* has shown comparable performance for MD, CI, FI and P and worse in the dimensions of IC. However, *peak_end_reg* showed positive statistical significance against *base_line* in all dimensions. These results suggest that *dummy* performs the best compared to all of the peak-end rule models, which align with the analysis of R^2 values themselves as mentioned earlier in this section. However, the weighted peak-end rule in estimated real-time evaluation is a better predictor than the average values of the estimated real-time evaluations. On the other hand, the simple peak-end rule model which takes the average of the peak and the end values is not present in the relationship between the estimated real-time evaluations and the retrospective evaluations.

In the scope of ablations of peak-end rule, *peak_end* has shown positive statistical significance against only using the peak (*peak_only*) in all dimensions. In contrast, it showed positive statistical significance against *end_only* only in the dimensions of MD FI and P, while negative significance in CI. In the dimension of IC, it did not show any statistical significance. These results show that for the dimensions of CI, the end of the estimated real-time evaluation is a better predictor than its peak or the average of the peak and the end. In the dimension of MD, FI and P, *peak_end* has shown positive statistical significance against both *peak_only* and *end_only*, suggesting that the peak-end rule explains the retrospective evaluation at least better than their ablations.

From the results of t-tests of LSTM models, there is no statistical significance for all LSTM models in all dimensions against *dummy*, indicating that LSTM failed to capture the underlying relationship between the estimated real-time evaluation of SIS and its retrospective evaluation. In t-tests against *base_line*, *lstm_pad* models in all dimensions have outperformed *base_line* except for FI, where the t-test results suggest the variation of the performance is due to a random chance. The *lstm_length_varying* model only in the dimensions of P and CI outperformed statistically significantly and underperformed in IC and FI dimensions. For the dimension of MD, the t-test results did not show statistical significance. These t-test results of LSTM models suggest that the hypothesized relationship was not present, or the current configurations of the LSTM models in our experimental setup were not able to capture the relationship effectively. However, similar to peak-end models, *lstm_pad* models in all dimensions and *lstm_length_varying* in P and CI were able to perform better than *base_line* across all dimensions, suggesting that they predict at least better than the average value of the estimated real-time evaluation.

5.3. Additional investigations of estimating SIS using LLM

5.3.1. Estimating summary evaluation

Given the poor performance of the predictive models across all dimensions, the assessment of the reliability and capability of LLM in estimating real-time evaluation of SIS is carried out to test the validity of our approach. A new prompt is designed to estimate the "summary evaluation" of SIS. This prompt asks LLM to act as a person in the conversation and answer the same sets of 10 questions about SIS given the entire conversation history. Our expectation for this prompt is that it would simulate the retrospective evaluation of SIS.

In order to evaluate the performance, we carried out a 10-fold cross-validation and used MAE (Mean Average Error) as the performance metric. MAE indicates the average difference between the predictive values and the actual value, thus the lower the better. In order to compare the estimated summary evaluation, a dummy model is implemented, which returns the class average of the training set. The results of comparing 714 summary evaluation estimations and the dummy model using the retrospective data from the PACO dataset as ground truth are shown in Table 5.3. As shown in the table, for all dimensions, the dummy model outperformed the estimated summary evaluation. Among the estimated summary evaluations, the performance of the dimension of IC is the worst (MAE mean of 0.88968), followed by MD with 0.82725. In contrast, the dimension of CI has scored the best performance (MAE mean of 0.70948) but is still lower than the dummy model. These results show that the dimension of IC is the hardest for the LLM with the summary evaluation prompt to estimate its value while it performs

	Dummy model		Estimated Summary Evaluaiton	
Dimension	mean	std	mean	std
MD	0.64651	0.02367	0.82725	0.05802
CI	0.56944	0.04882	0.70948	0.04154
FI	0.66347	0.04281	0.72749	0.04407
IC	0.76367	0.08077	0.88968	0.10577
P	0.43304	0.04694	0.74531	0.10731

Table 5.3: The results of the 10-fold cross-validation of comparing the performance of the dummy model and the estimated summary evaluation in terms of MAE and its mean and standard deviation (std)

best in predicting the dimension of CI and the class average predicts better for all dimensions. These results pose questions in estimating SIS using LLM, where the problems could be inherited from the capability of the LLM or the prompt design.

Based on these results of an attempt to estimate the summary evaluation, it can be expected that these labels for real-time evaluation of SIS might not fully reflect the actual evaluation of participants. Especially, the labels for real-time evaluation of the dimensions of IC and MD might have a higher chance that it does not reflect the actual data given their performance in estimating summary evaluation.

5.3.2. Manual Inspection of the real-time evaluations

A manual inspection of the labels is conducted to assess the reliability of the estimated evaluation of SIS, for both the real-time and the summary. Ideally, this process should involve multiple people to judge fairly. However, due to the limited resources, it is carried out by the author for a limited number of samples. We inspect the best and the worst-performing instances for each dimension in both the summary and the estimated real-time evaluation. For the estimated real-time evaluation, we selected the instances based on the performance of *peak_end_reg* model as it was the best performing model. We inspect the outputs based on whether we can retrace the output labels, if they seem plausible and if the ground truth seems plausible.

Manual inspection of estimated real-time evaluation

Overall, the LLM seems to struggle with changing the outputs based on each utterance. For the questions listed earlier in the prompt (Q1-5), their outputs are usually output a constant value, which is a sign that it failed to reflect the dynamic nature of the conversation for each incoming new utterance. On the other hand, for questions which were listed later in the prompt had shown a variance in their responses. It is also seen that the LLM sometimes outputs the responses in the wrong format or did not select from the correct lists of options. For example, it outputs "definitely person X" for the questions (Q1-Q7 in the prompt described in Table 4.1) where the expected answers are from "strongly agree" to "strongly disagree" or sometimes even output numerically where "strongly agree" seems is 5 and "strongly disagree" is 1.

Also, it was hard to retrace LLM's output earlier in the conversation, meaning the LLM seemed to struggle with analysing the situation and evaluating the questions from conversations. It could be because the conversation is mostly just greeting, introducing themselves and chitchatting, mainly about the crowdsourcing website and their experience with participating in similar experiments. Especially solely based on transcribed text, such contents of conversations made it hard to infer useful information related to situational interdependence. Additionally, some of the conversations were the entire conversation are only about their self-introductions not only in the beginning. In all of the manually inspected outputs, there was no extreme utterance by a speaker which would drastically influence the real-time evaluation of the SIS of the speaker, which makes it reasonable that the LLM's outputs did not change a lot by a new utterance.

For dimensions of MD, CI, FI and IC, the LLM outputs seem consistent, meaning the responses for a question and reverse-coded question were roughly matching. For dimensions of P, it was not the case.

This implies that the LLM seem to have failed to evaluate the conversation in terms of the questions related to the power balance.

Finally, the ground truth values did not seem plausible in some cases from looking at the transcribed text. It was especially noticeable in examples of dimensions of P. For example, in an example in which the prediction of the summary evaluation of P did not match with the retrospective evaluation by participants, the self-reported retrospective evaluation was 1, indicating that the person who rated it felt that his/her conversation partner had more power during the conversation. However, from the estimated real-time evaluation and manually following the transcribed text, none of the two in the conversation seem to have more power.

In summary, the manual inspection has revealed that though most of the labels in real-time seem plausible for both examples where the prediction went well and did not, except in the dimension of P, the transcribed text alone might not be the perfect modality to infer the situational interdependence given that most conversations did not contain much useful information to infer the situational interdependence.

Manual Inspection of estimated summary evaluations

Overall, similar to the problem in the real-time evaluation, most of the conversation primarily focused on their basic self-introduction, background, or general impression towards the crowdsourcing portal and the experiment. This might have limited the LLM to infer situational interdependence from the transcribed conversation. Consequently, this could have made it hard for LLM to estimate the summary evaluation of SIS.

Although the conversation did not seem to contain much relevant information when evaluating SIS, the responses to the questions of the dimensions of MD, CI, FI and IC were at least consistent across both cases of accurate and inaccurate estimations. In the scope of the dimension of P, the answers were inconsistent in both accurate and inaccurate estimation cases, which might suggest that the estimated summary evaluation of P was unsuccessful. This aligns with the observations of the estimated real-time evaluation, where the dimension of P seems to be challenging for the LLM to predict from the given information.

6

Discussion

This chapter discusses the results presented in the previous chapter further to elaborate on the implications of our studies. The research question of our study was "To what extent is an individual's estimated real-time evaluation of situational interdependence during interaction indicative of their retrospective evaluation of situational interdependence?". In order to answer this question, it was necessary to estimate the real-time evaluation of SIS using LLM due to the lack of ground truth. Hence, the reliability and validity of the approach of the estimated real-time evaluation of SIS is discussed by elaborating on the performance of modelling the summary evaluation and its comparison against the retrospective evaluation. Subsequently, given the research question, two hypotheses were formulated regarding the relationship between the retrospective evaluation of SIS and its estimated real-time evaluation, namely testing the existence of 1) peak-end rule and 2) a more complex relationship. After discussing the implications of the experiment results on these two hypotheses, we also discuss the limitations of this study and suggestions for future works.

6.1. Estimating real-time evaluation of SIS

The poor performance of LLM in estimating summary evaluation highlights two potential challenges. First, temporal distance might have affected participants' perceptions of situational interdependence, as suggested by prior research [9, 27]. Unlike participants answering a retrospective questionnaire, the LLM prompt did not explicitly model this temporal distance. Second, it is reported that LLMs struggle with multiple-choice questions (MCQs) and output biases in positions of options [31, 56]. This aligns with our prompt asking to output by selecting listed options, potentially causing bias. While LLMs started gaining recognition for the annotation process in Computational Social Science, which justifies our initial choice of taking this approach, their current limitations in handling temporal distances when simulating human participants and MCQs might suggest the unreliability for estimating real-time SIS evaluation using the LLM based on their performance with summary evaluation. While our exploration suggests the potential for LLMs in real-time SIS annotation, their limitations should be taken into consideration before making further implications.

6.2. Peak-End rule

Our analysis did not support the presence of the peak-end rule in the relationship between the retrospective evaluation of SIS and its estimated real-time evaluation for all dimensions. However, both the simple peak-end rule (*peak_end*)and the weighted peak-end rule model (*peak_end_reg*) have shown better performance than taking the average of the estimated real-time evaluation (*base_line*) for all dimensions.

There are several potential reasons for this poor performance of the peak-end rule, including the imbalance of both retrospective evaluation and the reliability of the estimated real-time evaluation as it could make the models biased to certain values, and amplifies the performance of the dummy model, which only outputs the class average. While these problems with the biased data could apply to all

models in the experiment, the limited generalizability of the peak-end rule could be a reason specifically for the peak-end rule models. Although several prior works reported the presence of the peak-end rule across different domains [14, 19], the absence of peak-end rule in our experiment aligns with the findings by Strijbosch et al. [42], which discusses the robustness of peak-end rule in more cognitively complex experimental situations. They argue that the peak-end rule was tested in simple and single-dimensional experiments traditionally (i.e. [14, 19]), thus it cannot be generalized to more complex or heterogeneous experiences. Our experiment focused on evaluating situational interdependence of conversations, which would be characterized as rather more multifaced experiences where the situations are constantly varying based on such as the conversation partner's behaviour, which falls into the category of more complex and heterogeneous experiences, where the peak-end rule might not apply.

6.3. Complex relationship

To test our second hypothesis of whether there is a complex relationship, we have conducted the experiment using *lstm_pad* and *lstm_length_varying*. However, none of the LSTM models was able to predict the retrospective evaluation of SIS from its estimated real-time evaluation, which shows our hypothesis was not rejected from this experiment. Despite the poor performance, they at least outperformed a model which simply takes the average of the estimated real-time evaluation.

Our results for the LSTM-modelled relationship contradict the finding that the LSTM model could indeed capture the retrospective evaluation of customer satisfaction (CSAT) from the human-annotated real-time evaluation of verbal cues in terms of activation, valence and satisfaction per sentence [20]. The potential reasons for that would be inherited from the difference in how SIS and CSAT are evaluated retrospectively internally in humans. CSAT score is measured by asking participants to rate "How do you feel about speaking with this social robot again?" in 1-5 Likert scale.

In addition, the imbalanced data in the estimated real-time evaluation of SIS and its retrospective evaluation, similar to the peak-end rule models as described earlier, could have influenced the model performance. Especially for complex neural network models like LSTM, the imbalanced training data could hinder its performance, which could have been the case as Figure 3.1 and 5.1 illustrates the imbalanced distributions of the labels. Such imbalanced data of the retrospective evaluation might also have inflated the performance of the dummy model, which outputs the class average labels as the predicted value. This could have influenced our analysis of R^2 values and the t-tests against this model.

6.4. Limitations and Future work

First of all, the limitations and future works of using an LLM in annotating the estimated real-time evaluation are presented. We assumed that the estimated real-time evaluation of SIS using zero-shot LLM reflects reality to some extent, which might not have been the case. Although our attempt to reliably obtain values of the estimated real-time evaluation through prompt engineering and its validity check using summary evaluation, given the lack of ground truth, it is still unclear to what extent it represents the actual real-time evaluation of SIS during conversations. Ideally, a proper manual annotation would be useful. But given that it requires a lot of resources, LLM could still be a good alternative but the support of humans is necessary as the current technology of LLM does not match up to the point where LLM can fully replace human annotators [57]. For example, with a small amount of reliable ground truth, LLM could outperform zero-shot learning by using few-shot learning [13]. Few-shot learning is a type of in-context learning for LLM which is provided with a few sample input-output pairs in a prompt so LLM can learn from them. Lastly regarding LLM, we have used Llama-2-7b-chat, which is one of the state-of-the-art LLM models. It would be interesting to test with its larger models as it has shown better performance in most of NLP tasks [45] and also with other state-of-the-art models such as GPT4 [28] or FLAN-T5 [7].

Additionally, we estimated the real-time evaluation of SIS solely using LLM from verbal content. Inputting verbal contents directly and letting models learn the intermediate embedding could be a different approach and could perform better as prior works have reported [20]. Also, using combinations of different modalities of data instead of only verbal content, such as nonverbal vocal cues or facial expressions, could be effective in accurately estimating the real-time evaluation as purely looking at verbal contents could lose some nuanced semantics expressed in other modalities of behavioural cues

[11, 30].

Regarding predicting the retrospective evaluation of SIS from the real-time evaluation, we have tested only with LSTM to model the relationship for our second hypothesis. It would be interesting to use different types of models as the assumption we made when using LSTM is that the temporal dynamics in the real-time evaluation are important elements in predicting the retrospective evaluation. Furthermore, as it has been shown that the duration does not matter in the retrospective evaluation [14, 21], future works can also test with the models that do not consider the assumption about temporality.

Finally, while this research specifically focused on individual dimensions independently to investigate the possible existence of a relationship, there could be an inter-dimensional relationship between the estimated real-time and the retrospective evaluation of SIS. In other words, a combination of multiple dimensions of real-time SIS might be indicative of one or multiple dimensions of the retrospective evaluation of SIS. Therefore, it would be interesting to investigate this in the future.

7

Conclusion

This study has investigated the relationship between the real-time evaluation of SIS (situational interdependence scale) [16] and its retrospective evaluation. We hypothesized the presence of the peak-end rule [14] and a complex relationship which can be modelled using LSTM (Long-short-term memory), independently for each dimension of SIS. A set of predictive models was implemented to test these hypotheses in the PACO dataset [24]. Given the lack of the ground truth of real-time evaluation of SIS, Llama2-7b-chat [45], one of the state-of-the-art LLMs, along with our designed prompt is used to estimate the real-time evaluation of SIS for each dimension of SIS for each spoken sentence by a speaker in conversations. From experiments, both hypotheses were rejected. However, both peak-end rule and LSTM models scored higher performance compared to a baseline model, which takes the average of the estimated real-time evaluation. All models seemed to struggle with the skewed data for the estimated real-time and retrospective evaluation of SIS. One of the potential causes could be due to the accuracy of the estimated real-time evaluation of SIS as it is uncertain to what extent it reflects the actual evaluation. The reliability of the usage of LLM in estimating SIS was further investigated by modelling the summary evaluation of SIS. The investigation revealed the difficulty of evaluating SIS for the LLM from a transcribed text of the conversation, especially when it is mainly about a random topic, such as self-introductions. It also showed the LLM's limited capability of labelling them correctly. In terms of the peak-end rule, prior research showed that the peak-end rule is only present in simple cognitive evaluation experiments, which supports the rejection of the peak-end rule hypothesis in the conversational setting [42]. On the other hand, LSTM was reported to be able to capture the relationship between the real-time evaluation of other multidimensional frameworks and retrospective evaluation [20], which contradicts our findings. Future work includes improved annotation of estimated real-time evaluation of SIS by creating human annotation or applying human-supervised few-shot learning of LLM, combining other modalities, usage of other types of predictive models, and looking into inter-dimensional relationships.

References

[1] M. Ahmad, Omar Mubin, and Joanne Orlando. "A Systematic Review of Adaptivity in Human-Robot Interaction". In: *Multimodal Technol. Interact.* 1 (2017), p. 14. DOI: 10.3390/MTI1030014.

[2] Elaine Domingues Alves et al. "Might high-intensity interval exercise be remembered as more pleasurable? An attempt to test the peak-end rule in the exercise context". In: *Perceptual and motor skills* 128.4 (2021), pp. 1586–1606.

[3] Max Bain et al. "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio". In: *INTERSPEECH 2023* (2023).

[4] Remco R Bouckaert and Eibe Frank. "Evaluating the replicability of significance tests for comparing learning algorithms". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2004, pp. 3–12.

[5] Joost Broekens et al. *Fine-grained Affective Processing Capabilities Emerging from Large Language Models*. 2023. arXiv: 2309.01664 [cs.CL].

[6] Meia Chita-Tegmark, Janet M Ackerman, and Matthias Scheutz. "Effects of assistive robot behavior on impressions of patient psychological attributes: Vignette-based human-robot interaction study". In: *Journal of medical Internet research* 21.6 (2019), e13729.

[7] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. DOI: 10.48550/ARXIV.2210.11416. URL: https://arxiv.org/abs/2210.11416.

[8] Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. "MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing Machinery, 2021, pp. 248–257. ISBN: 9781450383325. DOI: 10.1145/3447548.3467231. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3447548.3467231.

[9] Bernd Dudzik and Joost Broekens. *A Valid Self-Report is Never Late, Nor is it Early: On Considering the "Right" Temporal Distance for Assessing Emotional Experience*. 2023. arXiv: 2302.02821 [cs.HC].

[10] Bernd Dudzik et al. "Recognizing Perceived Interdependence in Face-to-Face Negotiations through Multimodal Analysis of Nonverbal Behavior". In: ICMI '21. Montréal, QC, Canada: Association for Computing Machinery, 2021, pp. 121–130. ISBN: 9781450384810. DOI: 10.1145/3462244.3479935. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3462244.3479935.

[11] Robin IM Dunbar et al. "Nonverbal auditory cues allow relationship quality to be inferred during conversations". In: *Journal of Nonverbal Behavior* 46.1 (2022), pp. 1–18.

[12] Paul Ekman. *Facial expressions of emotion: New findings, new questions*. 1992.

[13] Shutong Feng et al. *Affect Recognition in Conversations Using Large Language Models*. 2023. arXiv: 2309.12881 [cs.CL].

[14] Barbara L Fredrickson and Daniel Kahneman. "Duration neglect in retrospective evaluations of affective episodes." In: *Journal of personality and social psychology* 65.1 (1993), p. 45.

[15] Patrick Gebhard et al. "Marssi: Model of appraisal, regulation, and social signal interpretation". In: (2018).

[16] Fabiola H Gerpott et al. "How do people think about interdependence? A multidimensional model of subjective outcome interdependence." In: *Journal of personality and social psychology* 115.4 (2018), p. 716.

[17] Ursula Hess et al. "The bidirectional influence of emotion expressions and context: emotion expressions, situational information and real-world knowledge combine to inform observers' judgments of both the emotion expressions and the situation". In: *Cognition and Emotion* 34.3 (2020), pp. 539–552.

[18] Kai T Horstmann and Matthias Ziegler. "Situational perception and affect: Barking up the wrong tree?" In: *Personality and Individual Differences* 136 (2019), pp. 132–139.

[19] Daniel Kahneman. "Evaluation by moments: Past and future". In: *Choices, values, and frames* (2000), pp. 693–708.

[20] Yelin Kim, Joshua Levy, and Yang Liu. "Speech sentiment and customer satisfaction estimation in socialbot conversations". In: *arXiv preprint arXiv:2008.12376* (2020).

[21] Thomas Langer, Rakesh Sarin, and Martin Weber. "The retrospective evaluation of payment sequences: duration neglect and peak-and-end effects". In: *Journal of Economic Behavior & Organization* 58.1 (2005), pp. 157–175.

[22] Bharati Limbu, Gemma Unwin, and Shoumitro Deb. "Comprehensive assessment of triggers for behaviours of concern scale (CATS): Initial development". In: *International Journal of Environmental Research and Public Health* 18.20 (2021), p. 10674.

[23] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[24] Tiffany Matej Hrkalovic. "Designing Hybrid Intelligence Techniques for Facilitating Collaboration Informed by Social Science". In: *Proceedings of the 2022 International Conference on Multimodal Interaction*. ICMI '22. Bengaluru, India: Association for Computing Machinery, 2022, pp. 679–684. ISBN: 9781450393904. DOI: 10.1145/3536221.3557032. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3536221.3557032.

[25] Alexander Maye et al. "Subjective evaluation of performance in a collaborative task is better predicted from autonomic response than from true achievements". In: *Frontiers in Human Neuroscience* 14 (2020), p. 234.

[26] Humza Naveed et al. "A Comprehensive Overview of Large Language Models". In: *ArXiv* abs/2307.06435 (2023). DOI: 10.48550/arXiv.2307.06435.

[27] Donald A Norman. "THE WAY I SEE IT Memory is more important than actuality". In: *Interactions* 16.2 (2009), pp. 24–26.

[28] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL].

[29] John F Rauthmann et al. "The Situational Eight DIAMONDS: a taxonomy of major dimensions of situation characteristics." In: *Journal of Personality and Social Psychology* 107.4 (2014), p. 677.

[30] Monica Ann Riordan. *The use of verbal and nonverbal cues in computer-mediated communication: When and why?* The University of Memphis, 2011.

[31] Joshua Robinson, Christopher Michael Rytting, and David Wingate. *Leveraging Large Language Models for Multiple Choice Question Answering*. 2023. arXiv: 2210.12353 [cs.CL].

[32] Pranab Sahoo et al. "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications". In: *arXiv preprint arXiv:2402.07927* (2024).

[33] David Sander, Didier Grandjean, and Klaus R Scherer. "A systems approach to appraisal mechanisms in emotion". In: *Neural networks* 18.4 (2005), pp. 317–352.

[34] Giuseppe Sartori and Graziella Orrù. "Language models and psychological sciences". In: *Frontiers in Psychology* 14 (2023), p. 1279317.

[35] Klaus R Scherer. "The dynamic architecture of emotion: Evidence for the component process model". In: *Cognition and emotion* 23.7 (2009), pp. 1307–1351.

[36] Klaus R Scherer. "The nature and dynamics of relevance and valence appraisals: Theoretical advances and recent evidence". In: *Emotion review* 5.2 (2013), pp. 150–162.

[37] Klaus R Scherer. "What are emotions? And how can they be measured?" In: *Social science information* 44.4 (2005), pp. 695–729.

[38] S. S. Shapiro and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)". In: *Biometrika* 52.3/4 (1965), pp. 591–611. ISSN: 00063444. URL: http://www.jstor.org/stable/2333709 (visited on 05/27/2024).

[39] Sonish Sivarajkumar et al. "An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study". In: *JMIR Medical Informatics* 12 (2024), e55318.

[40] Juan Ignacio Specht. *42: GPT's answer to life, the universe, and everything*. Oct. 2023. URL: https://www.leniolabs.com/artificial-intelligence/2023/10/04/42-GPTs-answer-to-Life-the-Universe-and-Everything/.

[41] Karthik Sreedhar and Lydia Chilton. "Simulating Human Strategic Behavior: Comparing Single and Multi-agent LLMs". In: *arXiv preprint arXiv:2402.08189* (2024).

[42] Wim Strijbosch et al. "From experience to memory: On the robustness of the peak-and-end-rule for complex, heterogeneous experiences". In: *Frontiers in psychology* 10 (2019), p. 1705.

[43] Ala N Tak and Jonathan Gratch. "Is GPT a Computational Model of Emotion?" In: *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2023, pp. 1–8.

[44] Jurriaan L Tekoppele, Ilona E De Hooge, and Hans CM van Trijp. "We've got a situation here!–How situation-perception dimensions and appraisal dimensions of emotion overlap". In: *Personality and Individual Differences* 200 (2023), p. 111878.

[45] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[46] Olena Trofymchuk, Eduardo Liz, and Sergei Trofimchuk. "The peak-end rule and its dynamic realization through differential equations with maxima". In: *Nonlinearity* 36.1 (2022), p. 507.

[47] Yaacov Trope and Nira Liberman. "Temporal construal." In: *Psychological review* 110.3 (2003), p. 403.

[48] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[49] Bernard L Welch. "The generalization of 'STUDENT'S'problem when several different population varlances are involved". In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[50] Jules White et al. "A prompt pattern catalog to enhance prompt engineering with chatgpt". In: *arXiv preprint arXiv:2302.11382* (2023).

[51] Natalie A Wyer, Timothy J Hollins, and Sabine Pahl. "Remembering social events: a construal level approach". In: *Personality and Social Psychology Bulletin* 48.8 (2022), pp. 1238–1254.

[52] Michael Yeomans et al. "A practical guide to conversation research: How to study what people say to each other". In: *Advances in Methods and Practices in Psychological Science* 6.4 (2023), p. 25152459231183919.

[53] Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. "Investigating Large Language Models' Perception of Emotion Using Appraisal Theory". In: *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 2023, pp. 1–8. DOI: 10.1109/ACIIW59127.2023.10388194.

[54] Hongli Zhan, Desmond Ong, and Junyi Jessy Li. "Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14418–14446. DOI: 10.18653/v1/2023.findings-emnlp.962. URL: https://aclanthology.org/2023.findings-emnlp.962.

[55] Hongli Zhan et al. "Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9436–9453. URL: https://aclanthology.org/2022.emnlp-main.642.

[56] Chujie Zheng et al. "Large Language Models Are Not Robust Multiple Choice Selectors". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=shr9PXz7T0.

[57] Caleb Ziems et al. "Can large language models transform computational social science?" In: *Computational Linguistics* 50.1 (2024), pp. 237–291.

A

Prompt engineering

A.1. Initial Prompt

This chapter expands the work involving formulating the prompt to LLM for estimating real-time evaluation of SIS until converging to the final version prompt as detailed in Section 4.1. The motivation for the formulation and limitations are discussed for each prompt.

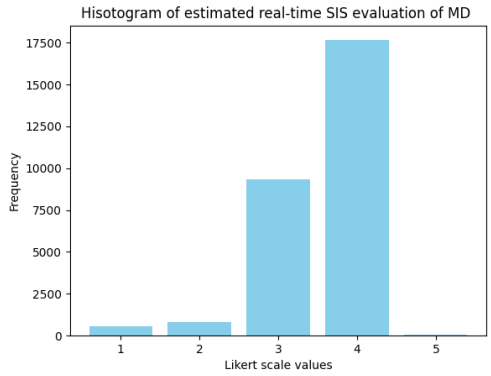
The task of this LLM pipeline was to estimate the real-time evaluation of SIS for each utterance of a speaker in a conversation. To be more specific, it requires generating the values for each dimension of SIS in each sentence given the conversation history. The initial prompt used in our study is shown in Table A.1. The prompt structure is inspired by Feng et al.[13], where they attempted to annotate emotion labels per sentence using an LLM, where the only difference with our works is the output labels are in terms of six distinct emotion labels, instead of numerical values for each dimension of SIS.

In the initial attempt, the prompt consists of two key elements: the context and the specific query that prompts the LLM to generate an output. The context includes definitions of each dimension of SIS (as outlined in Table 3.1) and the conversation history leading up to, but not including, the utterance of the speaker that will be evaluated. For each dimension of SIS, the value is estimated on a Likert scale of 1-5 to ensure alignment with the post-interaction questionnaire used in the two datasets. As this research does not look into the reasoning of LLM's outputs, it is specified only to return the numerical value of each dimension in the prompt.

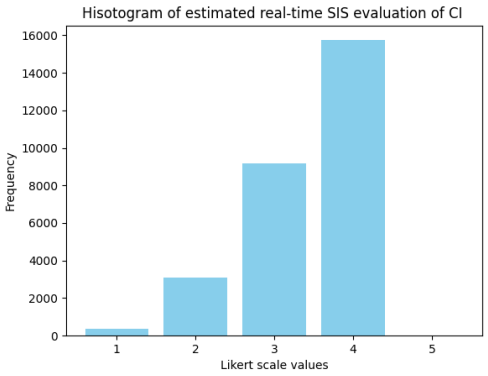
The distributions of the real-time evaluation SIS for each dimension using the old prompt are shown in Figure A.1. The histogram shows strong peaks in FI and IC at 4.0 and there are only a few labels with other values. The range of the values in FI and IC drastically differ from the retrospective evaluation shown in Figure 3.1. Given this unrealistic disparity in the estimated real-time and retrospective evaluations, it seems not to be plausible.

Prompt type	**Prompt template**
Task Definition	"Situational Interdependence" is defined in terms of [Mutual Dependence : (definition), Conflict of Interest : (definition), ...]
Query	Given the dialogue history between PersonA and PersonB : [PersonA: ..., PersonB: ..., ...], Analyse the extent of each element of situational interdependence in the next utterance of PersonA " ... " on a scale from 1 to 5, with 1 being "Extremely low" and 5 being "Extremely high"? {Task Definition} Please provide your answer as in the following example; MD:[num],CI:[num],FI:[num],IC:[num],P:[num]

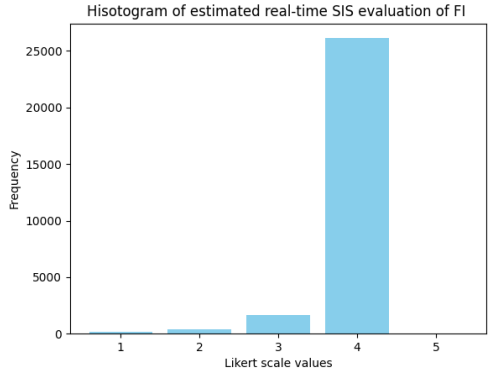
Table A.1: Template of the initial prompt



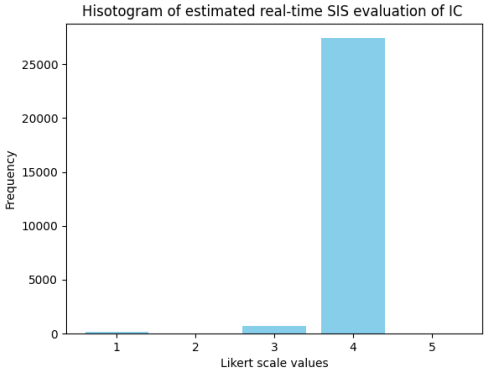
(a) MD (Mutual Dependence)



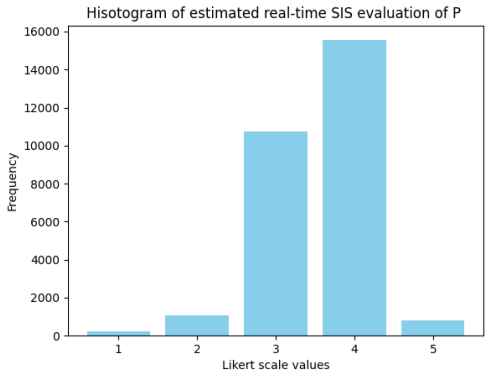
(b) CI (Conflict of Interest)



(c) FI (Future Interdependence)



(d) IC (Information Certainty)



(e) P (Power)

Figure A.1: Histograms of distributions of estimated real-time evaluation for each dimension of SIS using the initial version prompt.

The potential problem with this prompt is that it was questionable if these provided definitions of SIS are interpretable for the LLM, while the participants of the PACO dataset got 10 questions to answer, which were later processed to output the values for each dimension of SIS. Additionally, unlike the final version of our prompt, it does not incorporate explicit perspective-taking, which might pose a question of whether LLM evaluated each dimension based on the perspective of the speaker.

A.2. Second version prompt

To make the output comparable with the retrospective data and better simulate the speaker in a situation, we have designed the second version which does not provide the raw definitions which might be hard to interpret by LLMs. The prompt formulation is outlined in Table A.2. The LLM are now only asked to answer the questions the same as the participants, which could be easier to interpret by an LLM. In addition, explicit perspective-taking is implemented by explicitly mentioning in the prompt "act as the person". Explicitly describing the role that LLM has to "mimic" would increase the accuracy of their outputs as several studies agree upon [32, 50].

The histograms of the estimated real-time evaluation using this prompt are shown in Figure A.2. The histograms show more variety in their output values compared to those in Figure A.1. Interestingly, the peaks of MD, CI and FI have shifted by 1.0, from 4.0 using the old prompt to 3.0 using this prompt, meaning the LLM now outputs more natural outputs. Comparing with the retrospective evaluation histograms in Figure 3.1, it is observed that the peaks in the estimated real-time evaluation of CI and FI have shifted closer to the peaks in the retrospective one, while MD has shifted away from the peak. By using the same questions to label and calculate the scores for each dimension, the data seems to be more plausible but it still remains unclear whether this estimated real-time actually reflects the reality due to the lack of the ground truth.

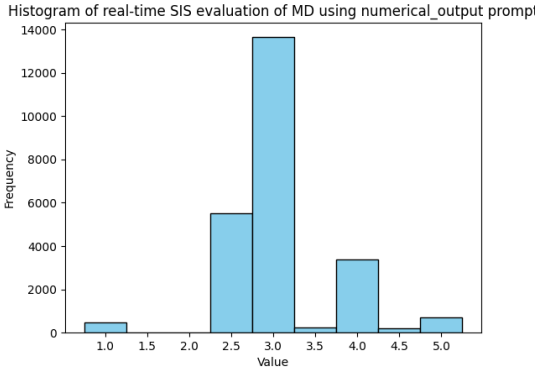
However, the problem of the bias of the numerical output still remains. As mentioned earlier, it is reported that LLMs have biases in outputting numerical values, where it seems that LLM has a "favourite" number [40]. Therefore, in the final version of the prompt, we removed the numbers associated with the different options and let the LLM only output the textual labels. Additionally, it does not contain a description of the contextual setting. By providing it explicitly, we expect the LLM to have more information about the participants, leading to more precise participant modelling. It could be beneficial in the context of zero-shot learning [41, 50].

A.3. Prompting per question

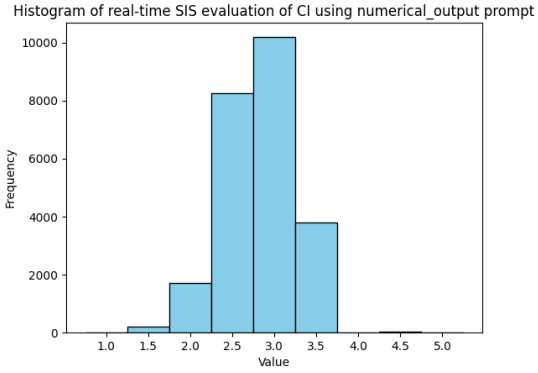
Prior works around prompt engineering claim that by splitting long and complex prompts into smaller chunks, it is possible to enhance the performance of LLM, called the Chain of Thoughts (CoT) technique [32]. As a part of prompt engineering for estimating real-time evaluation of SIS, we have attempted a similar approach to enhance the reliability of the responses. Instead of asking all 10 questions of SIS in a single prompt as in the final version prompt, the prompt contains the same contextual information but LLM has to answer only one question at a time. After its trial, however, manual inspection of the outputs revealed a bias towards extreme responses. Specifically, most responses for Q1-8 were "Strongly agree" and Q9 and 10 were "Definitely Person X". Because there are two questions for each dimension of SIS and one of them is a reversed coded question, this resulted in most of the SIS labels being at 3.0. The fact that LLM's responses did not change for reversed coded questions suggests that the prompt did not succeed in generating consistent answers. This contradicts the findings of prior works where splitting prompts into smaller chunks has been reported in higher performance. Therefore, we have not integrated this approach into the final version prompt.

Component	Prompt body
Task Definition	Act as Person *{ID of the speaker}*. You are asked to answer the following 10 questions at the moment you said *{sentence}* given this conversation history. *{history}*. From now on, person X means your conversation partner, Person *{ID of the conversation partner}*.
Query	Here, you are asked to rate the interaction you just took part in. We are interested in your personal (subjective) impression of the situation. Thus, we ask you to be as honest as possible and describe the situation by using the following scale: Strongly disagree = 1, Somewhat disagree = 2, Neither agree nor disagree = 3, Somewhat agree = 4, Strongly Agree = 5 1. What each of us does in this situation affects the other. 2. Our preferred outcomes in this situation are conflicting. 3. How we behave now will have consequences for future outcomes. 4. We both know what the other wants. 5. Whatever each of us does in this situation, our actions will not affect the other's outcome. 6. We can both obtain our preferred outcomes. 7. Our future interactions are not affected by the outcomes of this situation. 8. I don't think the other knows what I want. For each item, please think of the same conversation and indicate how the following statements describe the specific situation. Definitely person X = 1, Maybe person X = 2, Neither person X nor myself = 3, Maybe myself = 4, Definitely myself = 5 9. Who do you feel had more power to determine their own outcomes in this situation? 10. Who has the least amount of influence on the outcomes of this situation? Use the template to answer in JSON format. You do not need to provide explanation. {"Q1" : SCORE, "Q2" : SCORE, "Q3" : SCORE, ... , "Q10" : SCORE}

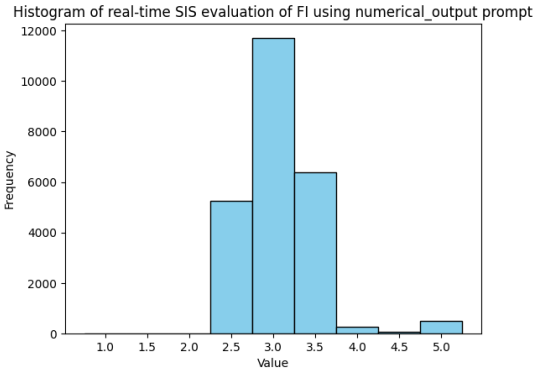
Table A.2: Template of the second version prompt



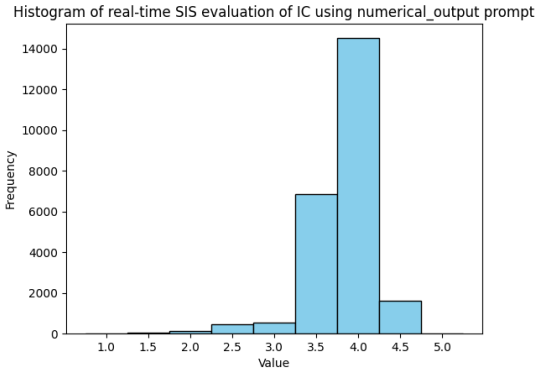
(a) MD (Mutual Dependence)



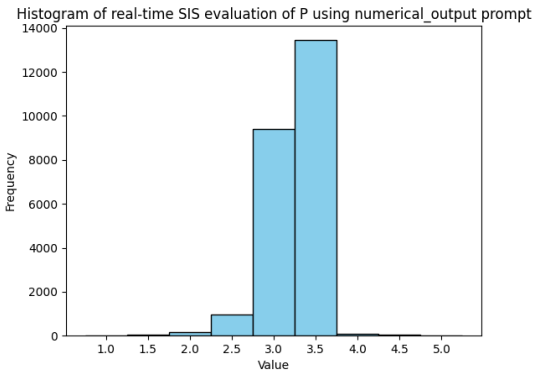
(b) CI (Conflict of Interest)



(c) FI (Future Interdependence)



(d) IC (Information Certainty)



(e) P (Power)

Figure A.2: Histograms of distributions of estimated real-time evaluation for each dimension of SIS using a prompt using the second version prompt

B

Supplementary data

B.1. Shapiro-Wilk test results

	Shapiro-Wilk statistics				
Model	**MD**	**CI**	**FI**	**IC**	**P**
peak_end	0.97396	0.86653	0.98998	0.96776	0.98344
dummy	0.66157	0.72062	0.67403	0.81166	0.75110
peak_end_reg	0.84214	0.80257	0.83719	0.89533	0.78403
base_line	0.97919	0.95719	0.98559	0.96059	0.96151
end_only	0.98963	0.94140	0.97980	0.97923	0.95653
peak_only	0.82847	0.97854	0.97376	0.98509	0.96193
lstm_pad	0.71038	0.45529	0.79117	0.71116	0.71957
lstm_length_varying	0.80180	0.93811	0.98533	0.96589	0.91198

Table B.1: Shapiro-Wilk test results for model performance measured in R^2. A statistic value closer to 1 indicates greater normality of the distribution.