Master Thesis

Educational Science and Technology

Faculty of Behavioral Management and Social Science

Evaluating ChatGPT's Ability for Automated Dialogue Analysis

Andisheh Sedaghat

Student Number: 3123936

1st Supervisor: Pantelis M. Papadopoulos

2nd Supervisor: Loes Hogenkamp

November, 2024

Acknowledgments

This journey has been like traveling to unknown territory as I had to lean into a new subject, acquire new knowledge, and develop a new way of thinking. This could not have been made possible without all whom I am deeply grateful to.

First, I extend my deepest appreciation to my first supervisor, Pantelis Papadopoulos, who carefully and patiently guided every path throughout this journey. His constant mentorship and inspiration gave me the resilience to move ahead with a critical mind. I am also grateful for my second supervisor, Loes Hogenkamp, whose constructive feedback and deep insights greatly enhanced the quality of the finished project.

A special thank you to my best friends, who have each supported me in a unique way, carried me through this journey and taught me the true meaning of friendship.

To my family, whose constant support and their belief in me have been a source of strength, and I am forever grateful for their love and sacrifices.

Abstract

Shifting to online learning drives a new way of assessing where both social and individual contributions are involved. Discourse analysis focusing on learning analytics signifies the students' dialogue as a means for assessing learning, as language possesses the capacity to open up access to students' thoughts and understandings. The richness of students' dialogue can be further augmented when four goals, as highlighted in the Academically Productive Talk framework by Michaels & O'Connor (2012), are achieved by students. According to the framework, students are held accountable for sharing knowledge (Goal 1), listening attentively (Goal 2), deepening their reasoning (Goal 3), and engaging in collaborative thinking (Goal 4). However, the laborious nature of the manual analysis does not always make it appealing because of the time and effort it requires from researchers and educators. Hence, inspired by research into automatic analysis, this study explored the potential of ChatGPT, a conversational agent developed by OpenAI in dialogue analysis. Specifically, the study compared ChatGPT's analysis with that of the researcher's analysis by applying Cohen's Kappa to gauge how effectively ChatGPT can perform deductive coding, a qualitative analysis where researchers use pre-determined codebooks to label the data into a fixed set of codes. The primary deductive discourse analysis was conducted, grounded on a coding scheme derived from the four goals of the Academically Productive Talk framework. The study continued in two phases: the pilot study, where the prompt model was revised, tested, and finalized, and also the dialogue quality study, where the performance of ChatGPT's analysis was compared with researcher analysis by examining characteristics of dialogue. The characteristics were classified into three groups: long/short, balanced/unbalanced, and succinctness/verbosity.

The pilot study achieved fair to substantial agreements with researcher-coded results. Findings revealed that ChatGPT effectively identified key utterances that contribute to productive dialogue, showing its potential to support deductive data analysis when the limitation in specifying each individual goal is not a primary concern. Additionally, the results indicated that the presence of certain indicative vocabulary impacts ChatGPT's performance. The inter-rater reliability analysis, taking into account the characteristics of dialogues, showed that ChatGPT was particularly effective in coding instances where students shared knowledge (Goal 1) during verbose dialogues, listened attentively to each other (Goal 2) in short dialogues, and engaged in longer dialogues with high word counts and

balanced word distribution among students. Additionally, higher agreement levels were observed in short dialogues when students deepened their reasoning (Goal 3) and engaged in collaborative thinking (Goal 4). These insights first inform researchers and educators for better handling of the prompts and also customizing AI tools in designing educational, supportive tools. Second, it calls for further research to build on these initial findings and enhance the integration of AI in educational contexts.

Keywords: Student's dialogue, Discourse analysis, Academically Productive Talk, ChatGPT

Table of Contents

1. Introduction

The rapid advancement of technology and the rise in big data availability have transformed students' learning experience and introduced a new interface of education (Velez et al., 2022). This transformation has appeared with a shift to a deeper analysis of students' learning processes within the social settings (Kent & Rechavi, 2020) that represents an adaptation to a learning approach that emphasizes both individual efforts and social participation (Sfard, 1998). Given that, Learning analytics and discourse analysis, in particular, have received growing attention. Every digital behavior or action is accounted for and translated into information that supports creating a proactive and customized system of informing and assisting students in an educational setting (Kaliisa et al., 2022). Additionally, The digital footprints of texts from natural language interactions generated for or as a result of these activities during online tasks have made access to in-depth information for analysis. As language has the potential to reveal the processes that are involved in the "joint creation of meaning, knowledge, and understanding" (Littleton & Whitelock, 2005), analyzing such information aids in understanding how students learn by using talk and that quality of talk can predict educational success (Gilbert & Dabbagh, 2005).

According to the literature on effective dialogue, the Academically Productive Talk (APT) framework (Michaels & O'Connor, 2012, 2015) introduces targeted strategies known as "talk moves" that teachers use to prompt students to engage in constructive dialogue, ensuring four foundational goals central to academically productive talk are achieved. APT has four goals, which require students Goal 1 "to share, expand, and clarify their own thoughts" Goal 2 " to listen carefully to each other," Goal 3 "to deepen their reasoning," and Goal 4 "to actively engage with the reasoning of others" (Michaels & O'Connor, 2012).

In online group tasks, learners can be expected to engage in similarly productive talk, provided these goals are satisfied. However, such context creates barriers for students to receive consistent, direct monitoring as the presence of teachers and educators is limited or restricted, which can impact the collaborative process during online learning (Silalahi & Hutauruk,2020). Teachers may interact with students at different times online, and with students engaging in discussions across various breakout rooms, it is not feasible for a teacher to be present in multiple rooms at once to guide the talk that would thereafter render personalized feedback effectively. Thus, the post-evaluation of students' talk during a collaborative task is needed to assess how each student in a group contributes to the goals of productive talk and how much they benefit from this approach.

To compensate for the teacher's absence in guiding dialogues, conversational agents have been developed to automatically support and guide the students through real-time prompting, lessening the teacher's involvement. Such conversational agents, like Clair (Collaborative Learning Agent for Interactive Reasoning), developed with the principles derived from the Academically Productive Talk (APT) framework, are designed to stimulate productive talk by facilitating constant interventions in collaborative settings like what one's teacher would have done, for instance, by prompting the students to share or clarify their thoughts or to encourage them to reason or challenge other peers thinking while discussing during the collaborative task (de Araujo et al., 2024). However, subsequent comprehensive analyses are needed to assess the effectiveness of these automatic interventions by analyzing the students' utterances in chats against the APT framework to learn how effective the interventions were in helping students meet the four Goals for productive discussions.

The most commonly used approach to analyze students' utterances is discourse analysis which could be implemented either manually or automatically. Manual content analysis of students' dialogues requires considerable time and effort from researchers and educators for a reliable and valid evaluation and this discourages them from initiating such studies (Pilny et al., 2019). There have been efforts to leverage Artificial Intelligence (AI) to automate content analysis, such as Long Short-Term Memory (LSTM) networks (Chen et al., 2022), yet as they use supervised learning models that require manual intervention in the coding process with a large number of datasets beforehand, it makes them a time-consuming process and, thus a less appealing alternative. Moreover, since models are trained on specific datasets, they may not generalize across different educational topics, which limits their applicability in academic settings. On the other hand, more advanced models like ConSent—a novel machine-learning algorithm that runs with minimal human interventions—can adapt to various topics but are not able to visualize the analysis (de Araujo et al., 2023).

Open AI's conversational agent called ChatGPT has represented a significant advancement, adapting deep learning and generative pre-trained transformer (GPT) architecture to engage in natural human-like language by generating relevant and contextually appropriate responses to user commands known as prompts (Ouyang, 2022). Research by Zamfirescu-Pereira et al. (2023) and Fiannaca et al. (2023) indicates that the quality of these prompts affects the quality of output generated by ChatGPT.

The capability of ChatGPT allows effortless task accomplishments without requiring sophisticated programming knowledge or additional training data (Dang et al., 2023). Its exceptional ability to process natural language benefits fields that rely on textual analysis

(Chang et al., 2023; Radford et al., 2019). It has been noted that ChatGPT facilitates efficient analysis of qualitative datasets (Zhang et al., 2023), and its adaptability to generate and interpret complex sociological knowledge can accelerate the initial analysis phases (Kien Nguyen-Trung, 2024). Specifically, Research by Wang et al. (2023) on ChatGPT demonstrated its proficiency in annotating classroom dialogue and identifying specific students' talk moves. However, implicit talk moves were not identified. Such limitations of large language models (LLMs) in processing data become apparent when they have to deal with long and complex data (Ghosal et al., 2020), reflecting the dialogue challenges noted by Azizova (2023), where dialogues generally do not conform to a uniform structure, exhibiting variations in characteristics such as length, word count, and distribution of messages between two participants.

However, as newer and more complex generative models like GPTs (Generative Pre-trained Transformers) rapidly overtake earlier versions, the need for ongoing research into the efficacy of models in educational and academic settings becomes increasingly critical. Such investigations are key to understanding, to some extent, these tools, such as ChatGPT, can effectively function, and their decisions are valid and reliable, which is vital to be adapted into educational environments. Teachers and academic professionals should trust and see the added value of AI's integration to utilize it as a tool in teaching and research effectively. Therefore, more research is necessary to examine how the models effectively analyze and interpret data. As such, this study examines the potential of ChatGPT based on GPT-4 in analyzing and detecting the four goals of productive talk in students' dialogue within an online collaborative task.

2. Theoretical Framework

2.1 Classroom Dialogue

Dialogue is one of the forms of speech that requires at least two individuals to engage in the communication process while alternating roles of speaker and listener. Dialogue does not own a uniform structure, as it varies syntactically, semantically, and even pragmatically (Azizova, 2023). Classroom dialogue has a long history of research conducted by researchers working in different areas, such as conversation and discourse analysis. Individuals make sense of the world through efforts between their own understanding and those of others (Phillipson & Wegerif, 2019), and language is undoubtedly one of the tools available to create and pursue reasoned arguments. Significant attention was given to the sociocultural aspects of learning in theory developed by Vygotsky in 1978 where three components of

culture, shared language, and social contexts in which individuals live were stated as indicators of learning and development (Mercer, 2008). This signifies the role of dialogue in learning and development by taking the sociocultural aspect of learning (Vygotsky,1978), that is, cognitive development is not just the individual process but occurs by engaging in meaningful dialogue and collaborative learning experiences with others (Howe & Abedin, 2013).

Children exposed to varied social experiences do not have the same language exposure, so we cannot expect all children to develop the same language skills necessary for learning and reasoning. In that sense, not only classroom dialogue but also quality is valued for its impact on individuals' learning and development (Mercer, 2008). Students who apply dialogic skills where they reason, discuss, argue, and explain, as noted by Alexander (2008), improve their higher-order thinking, reasoning, and collaborative problem-solving (Howe & Abedin, 2013; Kiemer et al., 2015; Kuhn, 2015, 2018).

The Academically Productive Talk (APT) framework (Michaels and O'Connor, 2012) that is extended and built upon Accountable talk (Michaels et al. 2008) is a widely used strategy for supporting students to engage in an effective collaborative dialogue. According to Adamson et al. (2014), "APT is a classroom discussion facilitation approach that has grown out of instructional theories that emphasize the importance of social interaction in the development of mental processes, in particular ones that value engaging students in transactive exchanges." (p. 97). Academically Productive Talk provide specific conversational strategies referred to as "talk moves" that teachers can use to engage students in productive discussions. These strategies help achieve the four necessary goals central to academically productive talk. The four goals include

Goal 1: Individual students share, expand, and clarify their own thoughts.

Goal 2: Students listen carefully to one another.

Goal 3: Students deepen their reasoning such as by asking for evidence or reasoning.

Goal 4: Students engage with others' reasoning, such as agree/disagree and why.

2.2 Discourse Analysis

Learning analytics (LA) is defined as " the measurement, collection, analysis, and reporting of data about learners and their contexts to understand and improve learning and the environments in which it takes place" (Siemens et al., 2011), that shapes our teaching approaches and strategies (Knight, Shum, & Littleton, 2014). LA methods have been

developed to aid scientists, researchers, and academics in understanding learning behaviors that can further facilitate informed decision-making (Larrabee Sønderlund et al., 2019).

Within the field of learning analytics, discourse analysis focuses on dialogues, analyzing them with the understanding of the role of language in learning and development and also for the potential of language to unfold the processes of knowledge construction in learning. Some of the discourse analysis approaches aimed to understand the large amount of text generated in online courses and activities with the increased adoption of computer conferencing within distance learning courses. Schrire (2004) conducted a discourse analysis of students' online interactions by adopting an approach analogous to the Initiation, Response, and Follow-up (IRF) model for analyzing the messages between students and instructors to understand their implications for cognitive development and learning. Likewise, in a study by Lapadat (2007), discourse analysis was employed to investigate the text-based interactions in an online education course, focusing on the asynchronous discussions between students and teachers. The thematic analysis was used to trace instances of agreement and disagreement among participants, showing how language use and discourse strategies helped to maintain community, coherence, and negotiation within the online educational context.

The sociocultural perspective views learning as a byproduct of the dialogical process where the quality of dialogue can influence the likelihood of success or failure in education (Mercer, 2004). Taking sociocultural perspectives, discourse analysis can represent how the way learners discuss and interact is performative, specifically when they result in making new knowledge or alternating existing ones across different contexts and technologies (Knight et al., 2014). Thus, discourse-focused social learning analytics is an approach to understanding how people use talk to build knowledge (Mercer, 2004).

In this line, learning analytics researchers have based this approach on the analysis of textual interactions in online courses, aiming to analyze the text-based discourse in online educational settings for better insight into the quality of students' written and spoken exchange posted in online collaborative environments (Kaliisa et al., 2022), and basically for understanding the process of building knowledge in students (Ferguson, 2009). This is supported by Knight & Littleton (2015) as they advocate for deeper analysis that moves beyond mere surface-level measures of learning such as participation counts in a linguistic activity, instead propose identifying "linguistic proxies" during co-constructive activities like how students construct reason or arguments that support high-quality discourse for learning contexts.

2.3 Discourse Analysis Models

Focusing on discourse analysis, Mercer (2004) introduced the quantitative approach in discourse analysis as a static coding of utterances into mutually exclusive categories and then described using statistical analysis. In contrast, qualitative analysis is beyond merely counting words to examining language and takes deeper consideration of talk while applying subjective interpretation through a systematic process of coding (Hsieh & Shannon, 2005). Through this approach, the actual talk is maintained as the primary data throughout the analysis so that researchers can examine the development of shared understanding in detail and track the likelihood of misunderstandings and differing perspectives. A qualitative method can be used in the conversational argumentation context to examine how justifications, i.e., reasons are given, then combining these qualitative insights with quantitative methods to identify statistical relationships between argumentative practices and variables like age, grade, and gender (Luginbühl et al., 2021).

One of the methodologies within discourse analysis, known as deductive coding, involves labeling all the data based on a codebook. It is a top-down process that begins with researchers developing a codebook with key variables from the chosen theory, and then the variables can be operationally defined through descriptions, subcategories, and examples based on the research focus or theory (Hsieh & Shannon, 2005).

The labeling data into a set of codes based on a predefined codebook (deductive coding) can range from fully manual to fully automated, varying in the degree of human involvement in the process. Manual coding may take much time to deal with a single transcript. Selecting and training qualified coders for effective analysis requires significant effort and time, yet achieving reliability among coders is not easy due to different abilities and understanding (Webb et al., 2019). While participants involved, like teachers and students, need to get timely feedback on the quality of dialogue, manual coding may not be feasible (Wang et al., 2014).

2.3.1 Automated Models on Discourse Analysis

One solution to address the limitations of manual coding is automated discourse analysis, referring to the coding of messages with the use of computer algorithms. The origin of automated discourse analysis dates way back to when experiments were almost limited to producing basic text statistics, such as word counting (Stone, Dunphy, Smith, & Ogilvie, 1966 as cited in Scharkow, 2017). A series of methods were later on developed to categorize the text, among these, the dictionary and rule-based methods being one of the earliest

proposed and adopted. These techniques employ predefined handcrafted dictionaries and sets of rules to interpret and analyze data (Scharkow, 2017).

In 2007, Erken and Janssen proposed automated techniques for coding dialogue acts. They applied the rule-based approach to predict communicative functions including five categories -argumentative (indicating a line of argumentation or reasoning), responsive (e.g., confirmations, denials, and answers), informative (transfer of information), elicitative (questions or proposals requiring a response), and imperative (commands).

Rule-based methods take much manual effort early in the process, particularly in generating the rules, which is a time-consuming and complex task for effective applications. The decisions and interpretations that researchers usually bring to their work, based on their knowledge and expertise, are rarely identified in the analyses produced by these methods due to their static approach. The static rules of the method may not adapt well to variations in new texts that weren't accounted for during the rule-defined phase (Varadarajan, Kasravi, & Feldman, 2003; Scharkow, 2013).

Given the emergence of machine-learning techniques, there has been a shift toward more flexible and innovative solutions in processing and analyzing large datasets, reducing the time required compared to traditional methods (Sarker et al., 2021. Such techniques have been leveraged for more efficient solutions to data analysis.

From speech recognition to robot control and natural language processing, machine learning techniques are applied to train systems for desired input-output behavior.

Machine learning techniques have been researched for analyzing student discourse in educational settings. Within that, Min et al. (2019) leveraged the Long Short-Term Memory (LSTM) technique in machine learning to predict dialogue breakdowns between students and chatbots during classroom interactions. In online learning, machine learning techniques have also been used to classify learners' speech and dialogue acts (Samei et al., 2014; Lin et al., 2022). Inspired by exploratory talk and sociocultural perspectives, transformer techniques were leveraged to stimulate students to engage more in exploratory talk by automatically detecting their dialogue into cumulative, disputation, and exploratory (Ubani & Nielsen, 2022). In a study by de Araujo et al. (2023), researchers used ConSent, a new machine-learning algorithm, to analyze the students' chat within the Academically Productive Talk (APT) framework. This model is aimed at automatically analyzing the students' chat to inform the timely interventions for the conversational agent.

In reviewing the application of machine learning techniques for automated analysis of student discourse, it is, however, not easy to learn and track the reasoning behind system

decisions and predictive analysis. This can influence human-AI collaboration if teachers and educators cannot trust or rely on the system's recommendations. Transparency is key in AI ethics, defined by interpretability and understandability (Vainio-Pekka et al., 2023). There is no doubt that it is important to understand the decision-making processes resulting in specific actions by AI (Singh et al., 2023). Explainability helps understand how and why AI algorithms decide or function in such a way, and that can build appropriate levels of trust (Zerilli et al., 2022). However, the explainability has to fit the users' knowledge level and context as the AI experts and typical users seek different types of explanations depending on the context (Mohseni et al., 2021). Buschek et al. (2022) suggest that while there is a common focus on the concepts of transparency or explainability of intelligent systems, we should explore how users perceive, understand, and interact with AI systems (Buschek et al., 2022). Natural language chatbot interfaces can affect transparency if the system engages users by stimulating a natural conversation, explaining why or how such a certain decision is made, and, for example, providing more clarifications in response to the user inquiry (Hernandez-Bocanegra & Ziegler, 2023).

2.3.2 ChatGPT

OpenAI released its ChatGPT conversational agent in November 2022 (https://openai com/ChatGPT), which benefits from the use of large-scale text data and significant human-coded samples in training data (Stiennon et al., 2020; Ouyang et al., 2022). ChatGPT is not designed for any specific need or task. However, its potential to converse in human language has had a broad impact on tasks like text generation, translation, and data analysis.

When using ChatGPT, the interaction starts with the user initiating the conversation by entering the instructions or commands, commonly known as "prompts," that stimulate the model for the desired responses from AI systems. A prompt defines the context for the conversation and specifies what is important to generate. According to Schmidt et al. (2024, p. 2), the prompt is key in "tuning the model performance, enhancing the quality of interaction, and achieving user satisfaction", that is, the quality of the prompt ensures the quality of output. Toward that, specific techniques have been introduced to produce outputs with better contextual relevance. For instance, in the few-shot prompt technique, ChatGPT is given a few specific examples relevant to the task that guides ChatGPT on how to respond to the command that's considered appropriate based on the example's context (Zhao et al., 2021). A chain of thought prompting involves guiding the model to generate a sequence of reasoning steps akin to the human thought process (Wei et al., 2022). Role-playing scenarios

encourage LLMs' to adopt specific roles and integrate relevant knowledge (Gao, 2023). Additionally, factors such as spelling, unclear text, and the number of examples can affect the quality of the output. The prompt size should also be considered (Wu et al., 2022). While these techniques can be applied across broad contexts, their efficacy relies on being customized to specific contextual requirements and integrating domain-specific knowledge (Wang & Jin, 2023). Similarly, Zhang et al. (2023) proposed a flexible framework for thematic analysis, emphasizing the necessity of adjusting prompts to the variation of the context relevant to the analysis.

The use of ChatGPT for initial coding in qualitative thematic analysis was explored by Turobov et al. (2024). The study revealed two-fold findings: While ChatGPT showed the potential to aid researchers in identifying the significant codes, details, and patterns within a complex dataset, the importance of human supervision and intervention was acknowledged as ChatGPT's inclination to produce descriptive rather than interpretive outputs, error making in quotations and code naming. Turobov et al. (2024) also linked ChatGPT's lack of inferential reasoning to its tendency to produce descriptive output. In contrast, Wachinger et al. (2024) further noted that ChatGPT could engage in complex interpretive analysis linking the identified themes to broader theories, but with the presence of the researcher, who should critically assess how applicable and useful a specific theory is to the data for avoiding mismatching.

In a deductive coding study with GPT-3 and a pre-determined codebook, the result demonstrated fair to substantial agreement with experts, suggesting that GPT-3 could effectively assist researchers, especially in handling large datasets (Xiao et al., 2023). The study also compared the different prompt designs. It was concluded that when using a codebook with examples, a *codebook-centered design,* the model showed better results than an *example-centered design,* which focuses on explaining the reasons behind each example. When comparing different techniques, the study found that the *one-shot setting*—where the model was given an example for each code—outperformed both *zero-shot* and *few-shot* settings.; however, increasing the number of examples did not improve the performance, as there was no significant difference when the model was given five examples in a few-shot setting compared to one example in a one-shot setting. This implies that the model could benefit from the examples, but the number of examples does not necessarily improve its coding accuracy. Huang et al. (2023) noted that explanations behind the classification were perceived to be clearer than those annotated by humans. However, despite the explanations provided by ChatGPT for its decisions that might enhance transparency, a question arose:

Could this convincing behavior pose a risk, where its decision could be incorrect yet persuasive?

GPT's potential, compared to other models in deductive coding for detecting specific talk moves, highlights both promises and challenges. ChatGPT showed the potential to identify talk moves, such as utterances that relate to one another, particularly when indicators are explicitly stated in the prompt, and also it offers a friendly interpretability configuration interface compared to another model in Wang et al. (2023) study. However, as observed by Wang et al. (2023), it did not perform effectively in predicting implicit talk moves.

Ghosal et al. (2020) identified that neural network models and attention mechanisms that are employed to identify the contextual clues from the context text have issues with coreference resolution and understanding complex and long-chained inferences. Zhang et al. (2023) pointed out the role of the memory reading operation in LLMs, such as ChatGPT, which is responsible for extracting necessary information from training data for reasoning and decision-making. Given the existing vast amount of memory and the likelihood that many are not directly relevant to the current context, selecting and extracting information based on its relevance and other task-specific factors can be challenging. This might be compounded in contexts involving dialogues, which, as Azizova (2023) noted, are basically unpredictable, with multiple speakers taking their turns without clear patterns and variation in the number of spoken words, length, and duration of the talk.

Across the strengths and limitations identified in existing literature, ChatGPT still presents a compelling means for further research. Research literature indicated that no validated research has examined the performance of the recent advanced ChatGPT developed based on the GPT-4 model in deductive coding to analyze students' dialogue. Research is needed to understand how different characteristics of text, specifically dialogue transcripts, such as length, number of words, and one dominancy, can affect ChatGPT's analysis performance.

2.4 Research Model and Questions

This study aims to examine the potential of ChatGPT 4 in analyzing and detecting four goals of productive talk in students' dialogue within an online collaborative task. Therefore, the following research questions are posed:

RQ1: How effective is ChatGPT 4 in terms of validity and consistency when detecting Academically Productive Talk during collaborative learning discussions?

RQ2: How effectively does ChatGPT-4 detect utterances when students share, expand, and clarify their thoughts (Goal 1) during collaborative learning discussions, across various dialogue characteristics?

RQ3: How effectively does ChatGPT-4 detect instances and moments where students are listening carefully to one another (Goal 2) during collaborative learning discussions, across various dialogue characteristics?

RQ4: How effectively does ChatGPT-4 detect instances where students deepen of reasoning (Goal 3) during collaborative learning discussions across various dialogue characteristics?

RQ5: How effectively does ChatGPT-4 detect instances where students are showing engagement with other reasoning (Goal 4) during collaborative learning discussions across various dialogue characteristics?

3. Method

3.1 Research Design

An exploratory research design was conducted to explore and evaluate the potential of ChatGPT based on GPT-4 in analyzing students' dialogue within an online collaborative task.

Firstly, The research employed a discourse analysis using a mixed method, quantitative and qualitative, coupled with a deductive approach. The qualitative deductive analysis involved a deep analysis to seek detailed evidence and interpret how students construct reason and develop collaborative understanding as explained by the four goals of the Academically Productive Talk framework (Michaels & O'Connor, 2012) using a predefined coding scheme. The quantitative deductive analysis was also incorporated, classifying the utterances according to the same predefined coding scheme derived from the four goals and measuring the frequency of observed goals achieved in the dialogues. The manual analysis and ChatGPT's analysis were then compared to assess the extent of agreement in the labeling of the dialogue transcripts.

3.2 Material

The presented research used existing online discourse transcripts that consisted of 46 transcripts of students' dialogues recorded in Dutch during five 9th-grade biology classes at two public high schools in a small city in the East of the Netherlands. The data was initially collected for a study aiming to investigate how their recent design of automated interventions -an analytics-based Collaborative Learning Agent for Interactive Reasoning (Clair)- could improve the productivity of student dialogue (de Araujo et al., 2024). All personal information

gathered is kept anonymous, and the research has been approved by the University of Twente Ethics Committee.

3.2.1 Dialogue Transcript Content

Each dialogue transcript documented a 50-minute dialogue between students who were grouped in pairs and instructed to discuss through chat while working on a digital inquiry-based science task. The task emphasizes collaboration and discussion as it progresses.

The task was delivered over two lessons through an online learning ecosystem (https://golabz.eu/) that covers the role of enzymes in the digestive system. Overall, the dialogue transcripts first began with students who read the content and shared the names and functions of enzymes in the digestive system. Following that, they were tasked to formulate the hypotheses and talk through their initial assumptions about the impact of temperature on the process of breaking starch by salvia. Students then had to discuss the observations from the findings of an online lab experiment engagement related to the reaction between saliva and starch at different temperatures. In the end, students were required to reflect on their steps and discoveries found during the lesson.

3.2.2 Dialogue Transcript Selection

The dataset of 46 dialogue transcripts was initially divided into two sets: the randomly selected 10 transcripts were used for the pilot study, which was further split into two subsets of 8 for extracting examples and 2 for prompting and testing. After analyzing the remaining 36 dialogue transcripts based on length, distribution of messages, and average word count, five transcripts were selected featuring in Length (Long/Short), Distribution of messages (Balanced/Unbalanced), and Average word count (Succinctness/Verbosity). The procedure behind the selection will be detailed in the process section.

3.3 Process

The study began by translating the entire dataset into English, as the data was originally recorded in Dutch during the initial data collection at two public schools in the Netherlands. To ensure the reliability of translation, Deepl Translator, a neural machine learning translation tool, was used for its ability to translate complex sentence structures while maintaining contextual integrity, as indicated by Cambedda et al. (2021).

Next, the four goals (Michaels & O'Connor, 2012) were adapted as the main themes, and the first ten transcripts were analyzed by the researcher accordingly to understand how the four goals manifested across the utterances. Following this initial analysis, it was decided

to develop a codebook where the goals were operationalized and made specific to utterances (current study). Therefore, for each goal, the criteria were defined and detailed along with examples (Appendix A).

The analysis proceeded in Atlas. ti, with the same first ten transcripts coded by the researcher according to the criteria. From the analysis, the utterances coded under each goal were collected for further sampling necessary for the codebook and prompt model.

Among the ten dialogues initially coded by the researcher, the two that were found challenging were subjected to recoding by an additional rater. The rater was provided with the codebook and necessary background information. First, the second rater coded the two transcripts, and the Interrater-reliability for each transcript, in relation to each goal, was calculated by comparing the codes assigned by the researcher and those assigned by the second rater.

Second, Given that the codebook was also part of the prompt model, the clarity of the codebook needed to be assured. Thus, discrepancies observed in the coding between the researcher and the second rater were addressed through reflective discussions as further explanations were sought. During the talks, it was attempted to learn more about the reasons behind the misalignment between codes. It helped with an in-depth review of the codebook, which required detailed modifications to address any ambiguities, and ensured its clarity when used in prompting.

A pilot study was subsequently conducted to explore effective strategies for the prompt model while evaluating the validity and consistency of ChatGPT's performance in detecting Academically Productive Talk (Four Goals).

3.3.1 Pilot Study

The pilot study had three steps:

First, Inspired by research into prompting techniques (Zhao et al., 2021; Wei et al., 2022; Gao, 2023), different iterations and testing were made to achieve the desired outcomes. Each of the four goals was sketched through a single prompt model, each given its corresponding criteria and samples but using the overall same structure throughout, which resulted in consistent outputs and, thereby, a more solid and reliable analysis. Uploading the entire dialogue transcript seemed to create issues with ChatGPT as it skipped the utterances, did not fully examine the transcripts, and also deviated from the instructions. Therefore, the dialogue was uploaded in batches of the 10-minute size and continued until the entire dialogue had been processed.

The outputs generated by ChatGPT from each prompt iteration were compared with the researcher's analysis. Based on the results, the prompt model was revised, and the desired features and strategies were carefully mapped out. (Appendix B).

Strategies and Revised Approach to Finalizing the Prompt Model

The strategies that showed better alignment and higher consistency throughout the four prompt models were identified and chosen as common elements in each model.

Table 1 displays strategies and examples common in the four prompt models.

Table 1

Strategies & Examples Used in Prompt Model

Strategies	Examples of prompt
Role-playing	Your role is to be an academic expert in Qualitative deductive Analysis, aiming to help teachers.
Background context	The following text explaining the context in which the dialogues occur.
Goal of task	1. Read and comprehend the user's uploaded dialogue. 2. Analyze the entire dialogue and identify where students share their ideas, knowledge, and observations about the digestion concepts and it related concepts with their peers without explaining the underlying reasons.
Few shots	Student-2 2022-12-05 09:46:46 because otherwise you have a big bump in your stomach and then suddenly nothing which makes you suddenly very hungry right Student-1 2022-12-05 09:47:24 no it is because otherwise the nutrients cannot pass through your blood
Chain of thought	Step 1: Read the following text explaining the context in which the dialogues occur.

Transparency	Provide a clear reason for why the utterance has been coded in a certain way.
Format (output)	Output: Present a table with the following columns: Column 1: Username Column 2: Timestamp Column 3: Utterance Column 4: Students engage with other peer's reasoning and thoughts Column 5: Reason

Along with the strategies common to the four prompt models, revisions made to each prompt model related to each goal,

In the first iteration of the prompt model for Goal 1, ChatGPT was tasked with identifying the utterances that met any criteria specified in the model, therefore coding them under the title "S*tudents share, expand, and clarify their thoughts".* Upon analyzing the ChatGPT's labeling, it was seen that ChatGPT primarily focused on these three words, *share, expand, and clarify,* mentioned in the title, and overlooked the criteria and examples provided in the prompt model. ChatGPT's narrow emphasis on just the title limited its attention to the detailed criteria and examples, and coding was merely based on ChatGPT's interpretation of those words. Thus, ChatGPT either failed to code the utterances originally identified by the researcher or coded them in ways that were irrelevant and unnecessary. This overemphasizing of such words might be partly due to the bias of language models, such as favoring answers commonly seen in the data it was trained on (Zhao et al., 2021). Removing the specific title of the goal and replacing it in the prompt model with "Code Goal 1" helped streamline the labeling process. This approach reduced the distractions or misalignments in the interpretation made by ChatGPT, so directing its attention to the special keywords that define the criteria as suggested by Hadi et al. (2023).

For Goal 3, the prompt model used the same design strategy, instructing ChatGPT to identify utterances that met any specified criteria in the model. However, the analysis indicated that including criteria and examples in the prompt model did not make a difference in ChatGPT's labeling for Goal 3, where the focus is on students deepening their reasoning. This might be attributed to the close semantic value between the words in the goal's title,

such as 'deepen' and 'reasoning,' and those used in the criteria, which likely triggered the same classification related to the goal. Nevertheless, in both iterations, whether incorporating the criteria and examples or not, ChatGPT failed to recognize the implicit cues embedded in the dialogues, similar to findings reported by Wang et al. (2023).

As to Goals 2 and 4, including both the title of the goal and a combination of criteria and examples in the prompt model showed an effective strategy. The presence of these elements influenced ChatGPT's labeling behavior. Removing either the title or the criteria, and examples changed how effectively and accurately ChatGPT labeled the data.

Second, after finalizing the prompt model, Cohen's Kappa test was applied to measure the inter-rater reliability (the level of agreement) between the coding performed by ChatGPT and the researcher on two sets of transcripts. Following that, an in-depth analysis was made to explore where ChatGPT's coding did not align with the researcher's coding.

Third, the consistency across two attempts with the finalized prompt model was also calculated by applying Cohen's Kappa.

3.3.2 Dialogue Quality Study

With the pilot study completed and the prompt model finalized, the study moved forward by factoring different dialogue qualities into the analysis.

In the following, a detailed explanation is provided regarding the step-by-step process from data preparation to analysis.

Step 1: The three groups of Short/long, Balanced/Unbalanced, and Succinctness/Verbosity were defined by studying three variables.

The variables were the total number of messages per dialogue, the ratio of messages exchanged by students in a dialogue, and the average number of words per dialogue.

The 36 dialogue transcripts were formatted in an Excel spreadsheet with three columns: username, timestamp, and text. For the analysis, dialogues were selected to control the confounding effects. For example, pairs of dialogues were chosen that had close similar totals in the number of messages and the average number of words per dialogue but were different in the ratio of messages exchanged by students.

1. The total number of messages per dialogue was counted as it indicated the length of each dialogue. After taking the average, those dialogues with a message number above the average were placed under Long, and those below the average fell under the Short group of dialogues.

2. To assess whether the dialogue was balanced or unbalanced, the ratio of messages exchanged by each student in a dialogue was calculated. The difference in these ratios across all dialogues was then computed. With the average taken, dialogues with a difference greater than the average were classified as Unbalanced, while those with a difference less than the average were classified as Balanced.
3. The quality of the conversation in terms of wordiness was identified by calculating the average number of words in each dialogue. After taking the average of the average of word counts across all dialogues, the dialogues with a higher value than the average were placed under Verbosity, and those with a closer value to the average were classified as Succinctness.

Table 2

Analysis of Dialogue Characteristics: Assessing Message Numbers, Message Difference, and Average Word Counts

Variable	M	*(SD)*	min	max
Message numbers	149.88	*(77.61)*	28	364
Message difference	10.35	*(7.89)*	0.12	26.53
Average word counts	4.5	*(0.82)*	3.5	6.48

The characteristics of these five transcripts are summarized in Table 3.

Table 3

Comparative Analysis of Dialogue Characteristics Across Different Categories

Category	Number of Messages	Message Differences	Average Word Counts per Dialogue	*(SD)*
Long	203	5.41	3.94	*(3.56)*
Short	101	5.27	3.89	(2.91)
Balanced	203	8.37	4.08	(4.24)
Unbalanced	236	16.95	4.79	(6.50)
Succinctness	203	8.37	4.08	(4.24)
Verbosity	211	0.80	6.48	(12.51)

Step 2: Following the dialogues classification, one dialogue transcript was selected for each variable from each group. A manual analysis was then conducted to establish a reference. The same two uncoded transcripts were uploaded in batches of the 10-minute size together with the prompt model specifically designed for each goal.

Step 3: The final phase involved first comparing the results generated by ChatGPT with those from the manual analysis. Inter-reliability for each transcript was calculated in relation to the manual analysis. The findings were then compared to understand how the quality attributes of the dialogues influence the ChatGPT's performance.

4. Results

4.1 Pilot-testing

As described in the method session, the two transcripts that were initially identified as challenging for analysis underwent a re-coding conducted by another rater.

In the first step, Cohen's Kappa was calculated to compare the coding consistency between the researcher and the rater for each of the four goals. The second step was testing the prompt models designed for prompting ChatGPT with the same transcripts. The model that showed better performance was then analyzed using Cohen's Kappa to assess the agreement levels between ChatGPT's coding and that of the researcher.

The inter-reliability of the re-coding, the selected model, and its consistency across four goals are presented.

4.1.1 Goal 1: Descriptive Analysis for Transcripts 1 & 2

In this analysis of Goal 1 from Transcript (1) in Table 4, the inter-rater reliability involving a rater, ChatGPT's first attempt, and a consistency attempt with ChatGPT was evaluated using Cohen's Kappa, revealing moderate to substantial agreement levels among the analysis. The percentage agreements were high, ranging from 93.75% to 94.38. The Kappa statistics between the researcher and the rater —0.51, ChatGPT's first attempt 0.54, and ChatGPT's second attempt 0.61— illustrate a consistent application of judgment by the rater and ChatGPT, indicating moderate agreement.

Table 4

Transcript 1/Goal 1 (Total no of quotes = 160)

Agreement: Researcher vs Rater/ChatGPT$_{attempt1}$/ChatGPT$_{attempt2}$	
Researcher (n=11)	Others

	Rater (n=11)	$ChatGPT_{attempt1}$ (n=10)	$ChatGPT_{attempt2}$ (n=17)
A. Coded only by the Researcher	5	5	2
B. Coded only by the other coder (i.e., Rater or ChatGPT)	5	4	8
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	6	6	9
D. Not coded by anyone	144	145	141
E. Agreement (C + D)	150	151	150
F. Percentage of agreement (E/Total)	93.75%	94.38%	93.75%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.51	.54	.61

In the "Inter-Rater Reliability Analysis - Goal 1 for Transcript (5) in Table 2, the data demonstrated high levels of agreement across three measures—Rater, the first attempt with ChatGPT, and the consistency check in the second attempt. The percentage agreement was the same at 96.46% for the Rater and the second attempt, slightly increasing to 96.48% in the first attempt with ChatGPT. The Kappa statistics suggested moderate agreement with values of .48 for Rater and higher values of .58 for both attempts with ChatGPT.

Table 5

Transcript 2/Goal 1 (Total no of quotes = 226)

Agreement: Researcher vs Rater/$ChatGPT_{attempt1}$/$ChatGPT_{attempt2}$			
	Others		
Researcher (n=8)	Rater (n=8)	$ChatGPT_{attempt1}$ (n=11)	$ChatGPT_{attempt2}$ (n=12)
A. Coded only by the Researcher	4	2	2
B. Coded only by the other coder (i.e., Rater or ChatGPT)	4	5	6
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	4	6	6
D. Not coded by anyone	214	213	212
E. Agreement (C + D)	218	219	218

F. Percentage of agreement (E/Total)	93.17%	91.56%	91.53%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.48	.58	.58

4.1.2 Goal 1: Qualitative Analysis for Transcripts 1 & 2

In two analyses by ChatGPT regarding the first and second attempts the utterances coded under Goal 1 "*The students share, clarify, and expand their thoughts"* were identified under Goal 3 "*The students deepen their reasonings*" in the researcher analysis. ChatGPT's coding emphasized the student's lack of in-depth reasoning in their explanations. For instance, the second utterance was labeled as Goal 1, however, Student 1 provided reasons for the earlier statement as follows:

Username	Quote	Code	Reason
Student-1 Example 1	*"the enzyme can now react with the next particle of nutrient is it I think"*	Goal 1	Shares an understanding of enzyme activity without explaining the biochemical processes involved.
Student-1 Example 2	*"because it detaches from the enzyme and then new nutrients can be put back in"*	Goal 1	Explains a process related to enzyme function but without deepening into the biochemical specifics. This provides further explanation about the process by which enzymes release one nutrient particle to interact with another, sharing relevant digestion knowledge without going into the deeper scientific mechanisms of enzyme activity.

In another example, in the reason given by ChatGPT, although it is explicitly indicated that student 1 provides a reason, it adds a justification for not having scientific depth, thus labeling it as Goal 1.

Username	Quote	Code	Reason
Student-1	*"so you also don't get pain from the big food chunks"*	Goal 1	Adds a reason for the task-related decision, sharing a practical aspect without scientific depth.

In analyses where students share inaccurate knowledge, ChatGPT could effectively identify it. Although the accuracy was not focused within the criteria defined by the code

book, ChatGPT did not code the utterance as Goal 1. For example, this utterance was not labeled in either the first or second attempt due to the inaccurate information.

Username	Quote	Code	Reason
Student-2 $_{attempt1}$	*"I think it doesn't contain the bacteria to destroy other kinds of substances"*	Not labeled	Speculates on enzyme function, lacks clarity and accuracy in explanation.
Student-2 $_{attempt2}$	*"I think it doesn't contain the bacteria to destroy other kinds of substances"*	Not labeled	Attempts to describe what enzymes do not do, incorrectly shares understanding of enzyme function. Lacking accurate reasoning.

4.1.3 Goal 2: Descriptive Analysis for Transcripts 1 & 2

In Goal 2, the inter-rater reliability analysis (see Table 6) involving a researcher and a rater demonstrated an observed agreement of 86.25%, which showed a substantial agreement as reflected by a Kappa of 0.68. In contrast, the first attempt with ChatGPT marked a decrease in observed agreement to 75.00%, leading to a moderate Kappa of .42. The second attempt with ChatGPT maintained the same level of observed agreement and resulted in a Kappa of .45.

Table 6

Transcript 1/Goal 2 (Total no of quotes = 160)

Agreement: Researcher vs Rater/ChatGPT$_{attempt1}$/ChatGPT$_{attempt2}$			
	Others		
Researcher (n=54)	Rater (n=46)	ChatGPT$_{attempt1}$ (n=46)	ChatGPT$_{attempt2}$ (n=56)
A. Coded only by the Researcher	15	24	19
B. Coded only by the other coder (i.e., Rater or ChatGPT)	7	16	21
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	39	30	35
D. Not coded by anyone	99	90	85
E. Agreement (C + D)	138	120	120
F. Percentage of agreement (E/Total)	86.25%	75.00%	75.00%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.68	.42	.45

In Goal 2 of the second transcript, the inter-rater reliability analysis (see Table 7), the observed agreement between a researcher and a rater was 89.82%, reflected by a Kappa of .69, indicating substantial agreement beyond chance. The first attempt with ChatGPT showed a decrease in observed agreement to 83.33%, with a Kappa of .47, suggesting moderate agreement. Despite the first attempt, the second attempt showed a fair agreement, a further reduction in observed agreement to 78.76%, and a Kappa of .34 was observed.

Table 7

Transcript 2/Goal 2 (Total no of quotes = 226)

Agreement: Researcher vs Rater/ChatGPT$_{attempt1}$/ChatGPT$_{attempt2}$			
	Others		
Researcher (n=48)	Rater (n=37)	ChatGPT$_{attempt1}$ (n=42)	ChatGPT$_{attempt2}$ (n=44)
A. Coded only by the Researcher	17	22	26
B. Coded only by the other coder (i.e., Rater or ChatGPT)	6	16	22
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	31	26	22
D. Not coded by anyone	172	162	156
E. Agreement (C + D)	203	190	176
F. Percentage of agreement (E/Total)	89.82%	83.33%	78.57%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.69	.47	.34

4.1.4 Goal 2: Qualitative Analysis for Transcripts 1 & 2

The interactions between students emerge as questions and answers. In ChatGPT's analysis, these questions were labeled under Goal 2, *"Students listen to each other carefully,"* as an initial step in listening behavior. However, this categorization sometimes led to incorrect labeling, particularly when responses from peers didn't directly follow the questions posed.

For example, in one instance, student 2 sought the opinion of student 1, but no responses were observed from student 1, as the dialogue progressed.

Username	Quote	Code	Reason
Student-2 attempt1	*"what do you think yourself"*	Students listen to each other carefully	Solicits Student-1's opinion, demonstrating engagement and value for peer input in forming a collective understanding.
Student-2 attempt2	*"what do you think yourself"*	Students listen to each other carefully	Prompts peer's opinion or thought, inviting substantive discussion.

4.1.5 Goal 3: Descriptive Analysis for Transcripts 1 & 2

The inter-rater reliability analysis involving three measures (see Table 8)—between a researcher and a rater, and two subsequent attempts with ChatGPT—consistently demonstrated high observed agreements. In the initial analysis between the researcher and the rater, the observed agreement was 95.63% with a Cohen's Kappa of .45, indicating moderate agreement. The subsequent analysis with ChatGPT also showed high levels of agreement; the first attempt recorded an observed agreement of 96.25% with a Kappa of .61, and the second attempt maintained high agreement at 95.00% with the same Kappa of .61, both indicating substantial agreement.

Table 8

Transcript 1/Goal 3 (Total no of quotes = 160)

Agreement: Researcher vs Rater/ChatGPT$_{attempt1}$/ChatGPT$_{attempt2}$			
	Others		
Researcher (n=10)	Rater (n=3)	ChatGPT$_{attempt1}$ (n=6)	ChatGPT$_{attempt2}$ (n=13)
A. Coded only by the Researcher	7	5	3
B. Coded only by the other coder (i.e., Rater or ChatGPT)	0	1	5
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	3	5	7
D. Not coded by anyone	150	149	145
E. Agreement (C + D)	153	154	152
F. Percentage of agreement (E/Total)	95.63%	96.25%	95.00%

Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.45	.61	.61

The inter-rater reliability for transcript (2) in Table 9 demonstrated high observed agreements across three measurements. In the rater analysis, observed agreements reached 98.67% with a Cohen's Kappa of .66, indicating substantial agreement. This pattern of high agreement was also observed in the first attempt with ChatGPT, where the observed agreement reported a value of 99.56%, with a kappa of .89 suggesting almost perfect agreement. The second attempt with ChatGPT also maintained a high level of observed agreement at 98.67%, with a Kappa of .72 indicating substantial agreement. Expected agreements by chance were also high, supporting that the high agreement levels were not random.

Table 9

Transcript 2/Goal 3 (Total no of quotes = 226)

Agreement: Researcher vs Rater/$ChatGPT_{attempt1}$/$ChatGPT_{attempt2}$			
	Others		
Researcher (n=5)	Rater (n=4)	$ChatGPT_{attempt1}$ (n=4)	$ChatGPT_{attempt2}$ (n=6)
A. Coded only by the Researcher	2	1	1
B. Coded only by the other coder (i.e., Rater or ChatGPT)	1	0	2
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	3	4	4
D. Not coded by anyone	220	221	219
E. Agreement (C + D)	223	225	223
F. Percentage of agreement (E/Total)	98.67%	99.56%	98.67%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.66	.89	.72

4.1.6 Goal 3: Qualitative Analysis for Transcripts 1 & 2

The number of utterances coded by ChatGPT with Goal 3 increased from the first attempt to the second attempt. This increase in categorization was more apparent in the transcript (1) from 1 code in the first attempt to 5 in the second attempt, reflecting a potential for over-categorization and inconsistency in ChatGPT's coding decisions during the analysis.

For example, However, ChatGPT recognized the context by tracking back to two previous utterances about the presence or absence of starch, in the first attempt, it did not label an utterance where student 2 asked about observations during the experiment, stating a lack of reasoning or explanation. In the second attempt, this same utterance was labeled as Goal 4, but with a reason contrary to that provided in the first attempt.

Username	**Quote**	**Code**	**Reason**
Student-2	*"no I do not share an opinion"*	Not labeled	Response to facilitator's question, lacks explanation or reasoning about the topic.
Student-1	*"you have to see if there is starch in it"*	Not labeled	Instructional comment, lacks reasoning or scientific explanation about why this is necessary.
Student-2	*"yes but how can you tell"*	Not labeled	Question indicating uncertainty, lacks explanation or reasoning about the method to determine starch presence.
Student-2 attempt1	*"does the saliva then turn black or brown"*	Not labeled	Inquiry about a reaction outcome, could indicate a hypothesis but lacks reasoning or explanation about why this change would indicate the presence or absence of starch.
Student-2 attempt2	*"does the saliva then turn black or brown"*	Students deepen their understanding	Trying to connect the concept with observable outcomes. This shows a deeper inquiry into understanding the physical changes that indicate the presence of starch, rather than merely knowing that a change occurs.

Additionally, utterances initially labeled as Goal 1 *"Students share, clarify, and expand their thoughts"* due to a perceived lack of reasoning were later classified as Goal 3 "*Students deepen their reasoning."*. Similarly, utterances originally categorized as Goal 4 *"Students engage with other reasoning"* were not interpreted without considering the fact the reason provided by students is a reflection of their own statement or in response to their peer's idea.

Username	Quote	Code	Reason
Student-2	*"because eventually it becomes a digestive product but the nutrient splits in two"*	Goal 1	Attempts to explain a task-related concept but does not delve deeply into the biochemical process involved.
Student-2	*"because eventually it becomes a digestive product but the nutrient splits in two"*	Students deepen their understanding	Student-2 explains how a nutrient eventually divides, providing a reasoning that deepens understanding of digestion products.

Username	Quote	Code	Reason
Student-1	*"so you also don't get pain from the big food chunks"*	Goal 1	Adds a reason for the task-related decision, sharing a practical aspect without scientific depth.
Student-1	*"so you also don't get pain from the big food chunks"*	Students deepen their understanding	Provides reasoning linking food size to physical comfort, suggesting why smaller pieces are beneficial, enhancing understanding of digestion.

Comparing the analyses of transcripts one and two, the kappa values for transcript two were higher, indicating substantial agreement, in contrast to transcript one, which showed moderate agreement. Transcript 1, which contained 160 utterances from students, included 14 utterances reflecting task-related scientific processes talk, while transcript 2, with 246 utterances, contained only 5 such utterances. The limited amount of scientific content in transcript 2 yielded higher agreement values by reducing the potential for over-categorization.

4.1.7 Goal 4: Descriptive Analysis for Transcripts 1 & 2

The inter-rater reliability analysis for Goal 4 - Transcript (1) in Table 10 indicated that the observed agreement by a rater was 93.75%. Subsequent measures with ChatGPT and the Consistency measure showed observed agreements of 90.63% and 89.38%, respectively. Cohen's Kappa values ranged from moderate to fair. As to the rater, kappa reported a value of .41, indicating moderate agreement. The initial attempt with ChatGPT similarly reported a Kappa of .43. The subsequent attempt, however, indicated a fair agreement with a Kappa value of .36, demonstrating a slight variation in consistency between different measures.

Table 10

Transcript 1/Goal 4 (Total no of quotes = 160)

Agreement: Researcher vs Rater/ChatGPT$_{attempt1}$/ChatGPT$_{attempt2}$			
	Others		
Researcher (n=12)	Rater (n=6)	ChatGPT$_{attempt1}$ (n=17)	ChatGPT$_{attempt2}$ (n=17)
A. Coded only by the Researcher	8	5	6
B. Coded only by the other coder (i.e., Rater or ChatGPT)	2	10	11
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	4	7	6
D. Not coded by anyone	146	138	137
E. Agreement (C + D)	150	145	143
F. Percentage of agreement (E/Total)	93.75%	90.63%	89.38%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.41	.43	.36

The inter-rater reliability analysis for transcript (2) in Table 11 demonstrated that the observed agreement for the rater was 96.02%, while both subsequent attempts showed an observed agreement of 95.13%.Cohen's Kappa values are indicative of substantial to moderate agreement.

The first attempt reported a Kappa value of .62, indicating substantial agreement. However, the first attempt with ChatGPT showed a lower Kappa value of .5, representing moderate agreement. In the consistency measure, the Kappa value was .57, falling between the rater and the first attempt, indicating a higher moderate agreement.

Table 11

Transcript 2/Goal 4 (Total no of quotes = 226)

Agreement: Researcher vs Rater/ChatGPT$_{attempt1}$/ChatGPT$_{attempt2}$			
	Others		
Researcher (n=14)	Rater (n=11)	ChatGPT$_{attempt1}$ (n=9)	ChatGPT$_{attempt2}$ (n=13)
A. Coded only by the Researcher	6	8	6

B. Coded only by the other coder (i.e., Rater or ChatGPT)	3	3	5
C. Coded by both the Researcher and the other coder (i.e., Rater or ChatGPT)	8	6	8
D. Not coded by anyone	209	209	207
E. Agreement (C + D)	217	215	215
F. Percentage of agreement (E/Total)	96.02%	95.13%	95.13%
Interrater reliability between the Researcher and the other coder (Cohen's kappa)	.62	.5	.57

4.1.8 Goal 4: Qualitative Analysis for Transcripts 1 & 2

In transcripts 1 and 2, 14 and 5 utterances, respectively, describe scientific processes related to the task. ChatGPT's analysis identified 5 of these 14 utterances in transcript 1, and all 5 utterances in transcript 2, as labeled under Goal 4 "*Students engage with other reasoning*." However, these utterances represented reasons given by students for their ideas or knowledge shared to solve the task, indicating a mismatch in coding. This pattern persisted in both analyses by ChatGPT.

For example, student 2 provided a reason for a statement previously made, yet ChatGPT categorized this under Goal 4, stating the reason for its decision was because the student provided a reasoning explanation that contributed to a deeper understanding of the task. This categorization conflicted with the criteria in the codebook for the prompt model, as it more accurately aligned with Goal 3 *"Students deepen their reasoning."*

Username	**Quote**	**Code**	**Reason**
Student-2 attempt1	*"because eventually it becomes a digestive product but the nutrient splits in two"*	Students engage with other peer's reasoning and thoughts.	Provides a reasoning explanation, contributing to a deeper understanding of the task.
Student-2 attempt2	*"because eventually it becomes a digestive*	Students engage with other	Explanation adds depth to the understanding of nutrients' role in digestion,

Username	Quote	Code	Reason
	product but the nutrient splits in two"	peer's reasoning and thoughts.	directly involves reasoning about the concept.

Taking another example, student 2 responded to a question in the task about making a hypothesis. But tracking back at the earlier statements, it is shown that student 1, except only one inquiry- " That's how you mean?" - does not contribute any specific idea only uttering some random words such as "okay'' and "what". The first statement by student 2 could be explained by Goal 4, however, it may not be clear due to some missing information not being shared by students. But following that, the dialogue was only dominated by student 2 revising and reflecting on the hypothesis-making process that better aligned with Goal 3 "*Students deepen their reasoning."*. However, if the labeling assumed a link back to the first statement, the reason provided does not support the coding decision.

Username	Quote	Code	Reason
Student-2	*"I think instead of starch saliva should be"*	Students engage with other peer's reasoning and thoughts.	Proposes a revision to the hypothesis, encouraging reevaluation and deeper understanding of the task concept.
Student-2	*the starch must then still behind breaks down*	Students engage with other peer's reasoning and thoughts.	Continues hypothesis development by refining the scientific process description.
Student-2 attempt1	*" the starch breaks down when the saliva is around 37 degrees"*	Students engage with others reasoning and thoughts.	Specifics about the condition under which starch breakdown occurs, reflecting an understanding of enzyme activity and temperature's role in digestion.

Username	Quote	Code	Reason
Student-2 attempt2	*"the starch breaks down when the saliva is around 37 degrees"*	Students engage with others reasoning and thoughts.	Refines hypothesis by explicitly linking temperature to biochemical reaction, enhancing content understanding.

4.2 Dialogue Quality Study

This section presents the results of the ChatGPT deductive analysis, taking into account the qualities of dialogue. Three groups—Balanced/Unbalanced, Verbosity/Succinctness, and Long/Short—were identified and tested with five transcripts.

4.2.1 Descriptive Analysis for Balanced & Unbalanced Dialogues

In Goal 1(see Table 12), In the balanced dialogue, the observed agreement was at 96.06%, with the Kappa statistic of .67 indicating substantial agreement. Conversely, in the unbalanced dialogue, the observed agreement with a slight increase was reported at 97.88%. The Kappa increased to .75, indicating very substantial agreement.

In Goal 2 (see Table 12), both the balanced and unbalanced dialogues revealed observed agreements, at 67.49% and 66.67% respectively, indicating a lower level of alignment between ChatGPT and the researcher analysis compared to Goal 1. For the balanced dialogue, the Kappa statistic of .31 suggests fair agreement beyond chance. In the unbalanced dialogue, the Kappa was .35, still indicating fair agreement. These observations indicated that regardless of whether the dialogues are balanced or unbalanced, there remains a low level of alignment between ChatGPT's analysis and the code book, hence with the researcher analysis, in identifying instances where students listen carefully to each other.

In Goal 3 (see Table 12), for the balanced dialogue, there was an observed agreement of 98.53%. The Kappa of .56 suggested moderate agreement beyond chance. Conversely, the unbalanced dialogue showed a similar observed agreement of 97.93% but led to a lower Kappa of .28, indicating fair agreement.

In Goal 4 (see Table 12), The balanced dialogue showed an observed agreement of 95.10%. The Kappa of .56 indicated a moderate agreement. The unbalanced dialogue exhibited a lower observed agreement of 87.60% with a Kappa of .49 which still presented a moderate agreement beyond. These observations showed a better performance for ChatGPT

in detecting Goal 3 where students deepen their reasoning in the balanced dialogue compared to a slight decrease in agreement under unbalanced dialogue.

Table 12

The Inter-rater Reliability Analysis for Both Balanced and Unbalanced Quality Across Four

	Balanced			Unbalanced		
	Observed Agreements	Percentage	κ	Observed Agreements	Percentage	κ
G1	195	96.06%	.67	231	97.88%	.75
G2	137	67.49%	.31	136	66.67%	.35
G3	201	98.53%	.56	236	97.93%	.28
G4	194	95.10%	.56	219	87.60%	.49

4.2.2 Qualitative Analysis for Balanced & Unbalanced Dialogues

Overall, from analyzing the dialogues, it was observed that in some dialogues, students either shared or directly referred to the scientific content exactly provided in the task. This content typically explained task-related scientific terms and concepts, as well as the reasons for their functioning, which aided the students in solving the tasks.

The balanced dialogue contained 8 utterances in which students articulated the piece of knowledge about the related-task scientific process while no such exact scientific content was shared during the conversation. However, the word “Because” was stated five times out of 8 utterances. The word “Because” can be indicative of reasoning and thereby deepening the reasoning referred to the Goal 3. From the five utterances with the word “Because” four utterances were labeled correctly as Goal 3, which perfectly aligned with all utterances labeled as Goal 3 in the researcher analysis, meaning there was little room for misinterpretation, over-categorization, or disagreement.

The unbalanced dialogue had 10 utterances containing 3 utterances in which students shared those of the exact scientific content given in the task and were labeled correctly as Goal 1, reflected in a higher Kappa value for Goal 1 compared to balanced dialogue. However, the frequency of the word “Because” is 0 in the unbalanced dialogue, and those of such 3 utterances were wrongly labeled again as Goal 4, mistakenly justified by ChatGPT that students provided the reason for their opinions in its given interpretation for its coding. This increased the incorrect distribution of ChatGPT’s coding leading to lower kappa value.

4.2.3 Descriptive Analysis for Succinct & Verbose Dialogues

In Goal 1 (see Table 13), both the succinct and verbosity dialogue showed high levels of agreement. For the succinct dialogue, there was an observed agreement of 96.06% with a Kappa of .67, suggesting substantial. Conversely, the verbosity dialogue showed an even higher observed agreement at 98.58% with a Kappa of .86, indicative of almost perfect agreement. Thus, the verbosity dialogue showed better performance compared to the succinct dialogue.

In Goal 2 (see Table 13), regarding succinct dialogue, the observed agreement was relatively low at 67.49% with a Kappa of .30. Conversely, with verbosity dialogue, the observed agreement rose to 91.00%, and the Kappa also led to .58, reflecting moderate agreement. For Goal 2, the verbosity dialogue performed better, achieving both a higher observed agreement and Kappa value.

In Goal 3 (see Table 13) for succinct dialogue, there was a high observed agreement rate of 98.53%, and the Kappa statistic of .56 indicated only moderate agreement beyond chance. In the verbosity dialogue, the observed agreement slightly increased to 99.05%, and the Kappa decreased to .49, still reflecting moderate agreement. For Goal 3, both dialogue styles achieved moderate agreement.

In Goal 4 (see Table 13), both the succinct and verbose dialogues showed observed agreements of 95.10% and 97.16%, respectively. The Kappa statistic remained constant at .56 in both types of dialogues. This indicated moderate agreement beyond chance, suggesting consistent performance across the succinct and verbosity dialogues.

Table 13

The Inter-rater Reliability Analysis for Both Succinctness/Verbosity Quality Across Four Goals

	Succinctness			Verbosity		
	Observed Agreement	Percentage	κ	Observed Agreement	Percentage	κ
G1	195	96.06%	.67	208	98.58%	.86
G2	137	67.49%	.31	192	91.00%	.58
G3	201	98.53%	.56	209	99.05%	.49
G4	194	95.10%	.56	205	97.16%	.56

4.2.4 Qualitative Analysis for Succinct & Verbosity Dialogues

The contrast between succinctness and verbosity in dialogues was less influenced by their significant difference in standard deviation for the word distribution and also the imbalanced or balanced distribution of words sent by each student in each message. The succinct dialogue, with an average word count of 4.2 and a standard deviation of 4.24, showed a lower average word count and standard deviation than the verbose dialogue, which had a higher average word count of 6.48 and a significantly larger standard deviation of 12.51. However, in the verbosity dialogue, the number of words was closely matched, with student 1 sending 502 words and student 2 contributing 461 words. This could be understood from the sentence structure where both students shared more complete sentences. In contrast, the succinct dialogue showed an imbalanced pattern, with student 1 sending only 234 words, including brief responses like "yes," "no," and "could be," while student 2 sent 549 words. This imbalance was reflected in the Kappa value for Goal 2, where ChatGPT demonstrated better coding accuracy in the verbosity dialogue.

4.2.5 Descriptive Analysis for Long & Short Dialogues

In Goal 1 (see Table 14), both the Long and Short dialogues exhibited high levels of agreement among raters. The Long Dialogue recorded an observed agreement of 97.04% and a Kappa statistic of .61. Conversely, the Short dialogue achieved an observed agreement of 99.01%, with a Kappa statistic of .66. These results indicated that, regardless of the length of the dialogue, there was substantial agreement in identifying instances where students share, clarify, and expand their ideas through interaction.

In Goal 2 (see Table 14), the Long dialogue showed a drop in observed agreement to 86.70% with a Kappa of .43. This lower Kappa, compared to Goal 1, indicated only moderate agreement beyond chance. In contrast, the Short dialogue displayed 90.10% observed agreement, and the higher Kappa of .66 demonstrated substantial agreement. This suggests that ChatGPT performed better with shorter dialogues regarding Goal 2.

In Goal 3 (see Table 14), the Long dialogue exhibited a high observed agreement of 98.52% and with a Kappa of .66, which reflected substantial agreement. Conversely, the Short dialogue achieved perfect observed agreement at 100%, resulting in a Kappa of 1.00 This represented the almost perfect agreement between ChatGPT and the researcher’s analysis when the dialogue was short.

In Goal 4 (see Table 14), the Long dialogue demonstrated an observed agreement at 95.57%. The resulting Kappa of 0.64 indicated substantial agreement beyond chance.

Meanwhile, the Short dialogue showed an almost perfect observed agreement of times where students engage with their peer's thoughts at 98.02%, with a substantial agreement at Kappa of 0.74. Therefore, the Short dialogue was more effective in Goal 4 compared to the Long dialogue.

Table 14

The Inter-rater Reliability Analysis for Both Long/Short Quality Across Four Goals:

	Long			Short		
	Observed Agreement	Percentage	κ	Observed Agreement	Percentage	κ
G1	197	97.04%	.61	100	99.01%	.66
G2	176	86.70%	.43	91	90.10%	.66
G3	200	98.52%	.66	102	100.00%	1.00
G4	194	95.57%	.6	99	98.02%	.74

4.2.6 Qualitative Analysis for Long & Short Dialogues

The long dialogue, consisting of 203 utterances, included seven utterances where students share their knowledge, incorporate reasoning, or engage with their peers' thoughts. In contrast, the short dialogue, with 101 utterances, featured only two utterances where students negotiated scientific information related to the task. These two utterances convey very brief information, with no deepening observed. In ChatGPT's analysis of goal 3, particularly no utterances were categorized under Goal 3, which aligned perfectly with the researcher's analysis, as reflected by a kappa value of 1.00.

5. Discussion

This chapter begins by discussing the important findings gained from assessing the validity and consistency of ChatGPT's analysis in identifying the utterances that contribute to the four goals of Academically Productive Talk during collaborative learning discussions (RQ1). It then explores the impact of dialogue characteristics on ChatGPT's analysis in terms of Goal 1 (RQ2), Goal 2 (RQ3), Goal 3 (RQ3), and Goal 4(RQ4) is discussed.

5.1 RQ1: Evaluating the Validity and Consistency of ChatGPT-4 in Detecting Academically Productive Talk During Collaborative Learning

The findings indicate the feasibility of using ChatGPT for detecting Goals. In a deductive coding task, by combining GPT-4 and a codebook, ChatGPT achieved fair to substantial agreement with the researcher. Thus, one can realize that ChatGPT has the potential to detect key utterances and interactions in relation to productive talk regardless of the specific goal accuracy. In particular, in a comparative analysis between four goals, the performance of ChatGPT was more prominent when predicting Goal 3 with achieving a higher agreement on both transcripts. This can suggest that ChatGPT detects more accurate utterances in a dialogue when students deepen their reasoning; likely, the analysis is facilitated by the presence of indicative vocabulary such as "Because," which is commonly used for times when deepening the reasoning.

Revising the prompt models improved the agreement for Goal 1 to moderate and achieved substantial agreement for Goal 3. However, the identified discrepancies need to be addressed. Upon close examination of the analysis, an inconsistency was observed between the researcher's and ChatGPT's labeling for Goal 1, where ChatGPT mistakenly categorized utterances with Goal 1, which the researcher had identified with Goal 3.

In its classification, ChatGPT suggested that students were merely explaining the process without deepening into the process. However, those utterances labeled under Goal 1 for lack of in-depth reasoning were later classified as Goal 3 "*Students deepen their reasoning*."

This misalignment might be attributed to ChatGPT's tendency to match set criteria regardless of their actual fit (Wachinger et al., 2024). However, this phenomenon did not seem to occur for Goal 3, which can be related to the better alignment of ChatGPT's memory(trained data) with goal definition, possibly not leaving for over-interpretation when a deeper analysis of implicit cues was not needed. This can be convincingly argued with a higher agreement achieved by Goal 3.

Compared to other goals, the likelihood of frequency of Goals 2 and 4 was higher with utterances not necessarily reflecting enough explicit information, and the criteria provided for these goals had more components to ensure that the conditions for meeting these goals held an inclusive cover. In the analysis, ChatGPT demonstrated the potential to detect the utterances where students relate to each other as it forms the basis of two Goals namely *Students listen to each other carefully* and *Students engage with other reasoning*, however, it struggled to

identify the specific references or some related responses within the context reflecting the true listening and engaging.

Context is central to NLP research. According to Poria et al. (2017), in dialogue analysis, context is explained by the surrounding utterances and aids in classification by bringing contextual evidence. Neural network models and attention mechanisms have been applied to make sense of these contextual clues from the context. However, they often showed limitations in addressing such aspects as the coreference resolution and understanding complex and long-chained inferences (Ghosal et al., 2020).

These constraints may suggest why despite the inclusive inclusion of criteria, ChatGPT cannot always meet the analytical demands, such as the complex interpersonal behaviors key to Goals 2 and 4.

ChatGPT's coding misalignment in Goal 3 could be further explained by looking at the reasons ChatGPT provided for its coding, particularly how ChatGPT justified engagement when it associated the engagement with the task at hand rather than with peers' thoughts and reasoning. However, such engagement was not included in the criteria, and it could be due to the model's tendency to mismatch or to a hallucination phenomenon where meaningless or irrelevant information is generated, as suggested by Athaluri et al. (2023). This could be caused by ChatGPT's reliance on statistical and computational operations based on the given input and the training data rather than a meaningful understanding. Additionally, syntax and semantic errors, particularly during translation or due to random mistakes and the use of unstructured sentences in spoken language, can contribute to challenges that confuse large language models (LLMs) systems (Tai et al., 2024).

The results generated from the analysis of the first and second attempt iterations were compared and found to be consistent with slight fluctuations observed. Given this consistency, the recent work by Xiao et al. (2023) suggested the practicability of using an LLM as "another rater" for qualitative analysis akin to the concept of analyst triangulation, indicating involving multiple analysts in the data analysis process (Patton 1999). The explanations provided by ChatGPT on its decision-making can add more viewpoints to the research quality and assist researchers in data analysis processing, yet with a careful and reflective approach.

5.2 RQ2: Assessing ChatGPT-4's Performance in Detecting Goal 1 Across Various Dialogue Characteristics

The agreement ranged from substantial to perfect for two groups of Balanced/Unbalanced and Succinctness/Verbosity dialogues. The kappa value was higher for

dialogues characterized by Unbalanced and Verbosity, particularly for Verbose dialogue, which featured a higher word count and also a greater variability in word count, as represented by a higher standard deviation.

A higher standard deviation in verbose dialogues shows greater variability in the length of utterances, which was observed with utterances where students explicitly defined concepts and processes related to the task. Such explicit definitions increased the likelihood of the accuracy of ChatGPT's coding for Goal 1. These definitions may give the ChatGPT more readily explicit access to the information needed for identifying the relevant utterances and provide a more straightforward path for processing data. Therefore, ChatGPT might have encountered fewer unexpected variations, thus increasing the performance of ChatGPT for detecting utterances that align with Goal 1.

When selecting dialogues based on Unbalanced and Balanced characteristics, the length and average number of words per message (Verbosity/Succinctness) were meant to be closely matched for the two dialogues, so only one characteristic varied at a time. However, the Unbalanced dialogue was likewise influenced by the higher standard deviation compared to the Balanced dialogue, which went through the abovementioned condition and resulted in a higher level of agreement.

The Kappa value indicated a similar level of agreement, classified as substantial, for both Long and Short dialogues. This finding suggests that ChatGPT-4 maintains an effective performance in recognizing these utterances regardless of the length of the dialogue, however, as they featured lower word counts and standard deviations, the verbose and unbalanced dialogues outperformed them, achieving higher agreement values.

Thus, it can be concluded that verbose dialogue that included a higher number of words and provided clearer context or more detailed information with higher standard deviation could improve the model's ability to accurately identify Goal 1.

5.3 RQ3: Assessing ChatGPT-4's Performance in Detecting Goal 2 Across Various Dialogue Characteristics

A range of fair to moderate agreement was achieved for dialogues characterizing Balanced/Unbalanced, Verbosity/ Succinctness, and Long as to Goal 2, with a substantial agreement for short dialogue. The Balanced and Unbalanced dialogues both demonstrated the same level of fair agreement with the research analysis, indicating that ChatGPT's performance was not significantly affected by whether the dialogue was dominated by one student or not.

The analysis by ChatGPT showed a substantial agreement for the short dialogue compared to its longer counterpart. The short dialogue also could outperform other groups characterized as Balanced/unbalanced and Succinctness/Verbosity.

The short dialogue like other dialogues was also uploaded into a 10-minute batch, however, the distribution of messages was less than that of the long version within the same timeframe, maintaining a shorter transcript overall. In addition to previous observations and analyses learned from the pilot study regarding Goal 2, specifying constraints in addressing the contextual clues, the complexity of data processing increased with longer dialogues.

In the case of ChatGPT, The memory reading operation is in charge of extracting relevant information required for reasoning and decision-making from the training data. Long or complex prompts and demanding transcripts may pose a challenge for ChatGPT to access a broader quantity of memory information and entities to process more relevant parts based on the input prompt (Zhang et al., 2023), thus yielding lower kappa for agreement compared to the short dialogue.

Verbosity led to a higher Kappa value and agreement compared to succinct dialogue and other characteristics such as Balanced, Unbalanced, and Long. It can be suggested that, except for the substantial performance of short dialogues, longer dialogues with higher word counts can also positively affect ChatGPT's performance. Furthermore, the distribution of average word counts for both students in verbose dialogues was closely equal, likely making it easier for ChatGPT to navigate the students' questions and responses, improving reference-making and prediction for Goal 2. Given this, the higher word count from both students may provide better context for ChatGPT's attention mechanisms to identify contextual clues (Ghosal et al., 2020), which help in understanding the requirements for Goal 2 and addressing the limitations of long dialogues.

5.4 RQ3: Assessing ChatGPT-4's Performance in Detecting Goal 3 Across Various Dialogue Characteristics

Data from deductive analysis by ChatGPT on Goal 3 showed that when students shared one piece of scientific information, these dialogues achieved almost perfect agreement. This performance outperformed other characteristics, each of a longer length; indeed, no mislabeling occurred in the identification of Goal 3. Balanced, Verbose, and Short dialogues, showed moderate agreement, while Unbalanced dialogue, which in turn reached a fair agreement with researcher analysis. The lack of explicit indicators such as "because," in Unbalanced dialogue posed the challenge in detecting Goal 3 "Students deepen their reasonings.", This also

increased the likelihood of wrong labeling, as noted when the information was implicit or hidden (Wang et al. 2023). Furthermore, Unbalanced dialogue presented problems for ChatGPT due to the high standard deviation found in definitions. For instance, in coding those definitions, the main discrepancy between the researcher and the ChatGPT analysis was ChatGPT's inability to differentiate whether the reason cited was the student's original thinking and contribution or parts of definitions students had shared.

ChatGPT's limitations with understanding and maintaining context over longer text, especially in tasks requiring comparisons and judgments, were pointed out by Li et al. (2023). This is particularly challenging when it comes to weighing evidence against a set of criteria or understanding relationships and dependencies within the text before making conclusions, inferences, or deductions based on the available information.

Therefore, the Short dialogue achieved almost perfect agreement in identifying Goal 3, where barely any information is shared without the need for extensive reasoning or complex data processing. The model faced challenges in longer dialogues, given with higher standard deviation in dialogue and an absence of explicit reasoning clues.

5.5 RQ4: Assessing ChatGPT-4's Performance in Detecting Goal 4 Across Various Dialogue Characteristics

Balance vs. Unbalanced and Verbosity vs. Succinctness dialogues do not make a difference in the performance of ChatGPT in analyzing Goal 4, leading to a moderate agreement. It can be concluded that the effect of dialogue domination by one student or verbosity did not change the model's ability to interpret the dialogue's content for Goal 4. The Short dialogue, in turn, had a higher agreement and Kappa values than the Long Dialogue and all other dialog characteristics.

It can be recognized that the length of dialogue has a decisive role in determining the accuracy of ChatGPT's analysis regarding student engagement. This could also account for the difficulty that ChatGPT faces in understanding contextual clues in long texts used to detect long chains of inference as characterized by Ghosal et al. (2020) that inform Goal 4. This then might impose an obstacle to effectively tracking the negotiation of students' reasoning that occurs over several steps.

5.6 Implications

The findings of the current study demonstrated the potential of ChatGPT in detecting key utterances and interactions that contribute to productive talk independent of the specific

goal accuracy. This has implications for educators that ChatGPT could be presented as an automated assistance tool for primary data analysis in educational settings. By identifying productive talk behaviors, ChatGPT can aid educators in filtering unproductive aspects thus allowing for more efficient targeted further analysis, particularly when researchers are dealing with large amounts of datasets that require manual coding and interpretation.

In particular, the inter-reliability values calculated for each goal indicate a higher agreement for the performance of the ChatGPT tool with utterances and texts that contain explicit, indicative vocabulary. Understanding whether and under what conditions of dialogues ChatGPT had better performance suggests that specific characteristics can impact the effectiveness. As to Goal 1, Verbose dialogue with higher word counts and standard deviations surpassed other characteristics in accurately identifying Goal 1. This characteristic is attributed to the explicit definitions and detailed information shared within these dialogues, which give ChatGPT more direct access to necessary information, simplifying the data processing and, thus, improving ChatGPT's coding for this Goal. For Goal 2, the study observed that the shorter dialogue strengthened ChatGPT's performance by simplifying demands for analyzing the interpersonal behaviors critical for identifying Goal 2. Additionally, it was found that longer dialogues with higher word counts and a balanced distribution of words between students also contributed to effective performance by the model. Addressing Goal 3, the Short Dialogue's effectiveness was further significantly proven by achieving a Kappa value of the one-almost perfect agreement. One realized when minimal information was shared, ChatGPT was highly effective. Also, the Long dialogues became more challenging, as shown by higher standard deviations and a lack of explicit reasoning clues. This might be an account of ChatGPT's struggles to perform effectively when the dialogue becomes more complex in how reasoning is expressed. Concerning Goal 4, the Short dialogue, was again represented as an effective characteristic for better performance by ChatGPT. Detecting long chains of inference in long dialogues might be challenging for ChatGPT when it is required to navigate the students' thinking while trading information and also analyze how it can affect the thinking process of other students involved.

These findings can guide first researchers in adjusting prompts to better handle this variability, therefore improving the AI's applicability in educational research. Secondly, they can aid AI educators in designing tools to favor such characteristics, such as shorter interactions when using AI to assess and support student learning.

ChatGPT was able to provide explanations and details of decision processes made for labeling specifically adapted to the context in a comprehensible manner. First, this could help

validate or verify the classification process in terms of each Goal, making it easier to navigate the overlaps, possible biases, or interpretations leading to mislabeling, simultaneously facilitating a pave to increasing trust in AI's integration in teaching and research effectively. Second, ChatGPT's explanations can add new insights and stimulate more in-depth reflections, given that teachers and educators can develop a more critical and reflective approach by involving ChatGPT's perspectives.

5.7 Limitations

In exploring the potential of ChatGPT in qualitative discourse analysis, the inherent limitations were acknowledged.

One limitation was ChatGPT's sensitivity to the prompts, which may produce biases in the findings. Although there was much effort to iterate the optimal prompt within the study time constraints, it was hard and demanded a great deal of time to analyze every single output and identify the algorithms and patterns present in the training data that may drive a bias or errors and adapt the model accordingly, all while ensuring the criteria owns the intended aims.

Additionally, despite the explainability of ChatGPT in its decision-making process by providing the reasons for each labeling, the understanding of the model's inner workings relying on neural networks complicates the analysis of how specific inputs led to the model's output.

On the other hand, in pairing dialogues to evaluate three targeted characteristics, the inherent variability of dialogue characteristics constrained an ideal match where two characteristics should be maintained the same while varying one. This difficulty in finding pairs of dialogues with two exact characteristics might create an unintentional effect impacting the results. Additionally, a more cautious and comprehensive approach had to be made to analyze the findings involving other variables and further consider the other two characteristics while focusing on the third characteristic.

Little to no prior research was found with the current study emphasis, specifically investigating the performance of ChatGPT or other Generative AI models in identifying goals within academically productive talks across different dialogue characteristics. Therefore, these limitations emphasize the primary insight into the study's findings and suggest that further understanding is required to build on these first insights under what conditions ChatGPT and automatic coding systems support productive talk in educational settings.

5.8 Conclusion and Future Direction

The current study aimed to explore ChatGPT's potential in analyzing students' dialogues and identifying the utterances that meet goals central to academically productive talk. While the results were convincing for the recognition of utterances conducive to productive talk regardless of goal identification, human supervision, and reflection (researchers and educators) to ensure accuracy and depth of analysis are required.

Moving forward, the study calls for further exploration of the quality of dialogues and ChatGPT's effectiveness and accuracy. The current study focuses only on data collected from high school students; the research could benefit from a broader scope to include higher educational settings, where the dialogues can often vary in complexity due to the advanced level of vocabulary, syntax, and topics discussed. Furthermore, The complexity of dialogue can be investigated to determine how different levels of vocabulary, syntax, and topic complexity affect ChatGPT's accuracy in predicting the goals. The complexity of dialogues can be further explored by incorporating and analyzing the specific characteristics of dialogue similar to those examined in this current study.

References

Adamson, D., Dyke, G., Jang, H., & Rosé, C. P. (2014). Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education*, 24(1), 92–124. https://doi.org/10.1007/s40593-013-0012-6

Athaluri, S. A., Manthena, S. V., Kesapragada, V. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus, 15(4)*. https://doi.org/10.7759/cureus.37432

Azizova, P. (2023). Linguistic Analysis and Learning of Dialogical Speech in Literary Texts. JETT, 14(4), 86-94. https://hdl.handle.net/10481/83448

Cambedda, G., Di Nunzio, G. M., & Nosilia, V. (2021). A study on automatic machine translation tools: A comparative error analysis between deepl and yandex for russian-italian medical translation. *Umanistica Digitale*, (10), 139-163. https://doi.org/10.6092/issn.2532-8816/12631

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3641289

Chen, J., Liu, Z., & Luo, W. (2022, July). Wide & deep learning for judging student performance in online one-on-one math classes. *In International Conference on Artificial Intelligence in Education* (pp. 213-217). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11647-6_37

Clark, D. B., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology* Review, 19, 343-374. https://doi.org/10.1007/s10648-007-9050-7

Dang, H., Goller, S., Lehmann, F., & Buschek, D. (2023, April). Choice over control: How users write with large language models using diegetic and non-diegetic prompting. *In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-17). https://doi.org/10.1145/3544548.3580969

Daniel Buschek, Malin Eiband, and Heinrich Hussmann. (2022). How to Support Users in Understanding Intelligent Systems? An Analysis and Conceptual Framework of User Questions Considering User Mindsets, Involvement, and Knowledge Outcomes. (2022). https://doi.org/10.1145/3519264

de Araujo, A., Papadopoulos, P. M., McKenney, S., & de Jong, T. (2023). Supporting Collaborative Online Science Education With a Transferable and Configurable Conversational Agent. *In Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023, pp. 416-419. International Society of the Learning Sciences.* https://doi.org/10.22318/cscl2023.469853

de Araujo, A., Papadopoulos, P. M., McKenney, S., & de Jong, T. (2024). A learning analytics-based collaborative conversational agent to foster productive dialogue in inquiry learning. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.13007

Fiannaca, A. J., Kulkarni, C., Cai, C. J., & Terry, M. (2023, April). Programming without a programming language: Challenges and opportunities for designing developer tools for prompt programming. *In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-7) https://doi.org/10.1145/3544549.3585737

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of chatgpt. *Advances in Neural Information Processing Systems*, 36. https://doi.org/10.48550/arXiv.2301.13867

Gao, A. (2023). Prompt engineering for large language models. Available at SSRN 4504303. http://dx.doi.org/10.2139/ssrn.4504303

Ghosal, D., Majumder, N., Mihalcea, R., & Poria, S. (2020). Utterance-level dialogue understanding: An empirical study. arXiv preprint arXiv:2009.13902. https://doi.org/10.48550/arXiv.2009.13902

Gilbert, P. K., & Dabbagh, N. (2005). How to structure online discussions for meaningful discourse: A case study. *British Journal of Educational Technology*, 36(1), 5-18. https://doi.org/10.1111/j.1467-8535.2005.00434.x

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, *45*(1), 87-110. https://doi.org/10.1109/TPAMI.2022.3152247

Hernandez-Bocanegra, D. C., & Ziegler, J. (2023). Explaining Recommendations through Conversations: Dialog Model and the Effects of Interface Type and Degree of Interactivity. *ACM Transactions on Interactive Intelligent Systems*, 13(2), 1-47. https://doi.org/10.1145/3579541

Howe, C., & Abedin, M. (2013). Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education*, 43(3), 325-356.

https://doi.org/10.1080/0305764X.2013.786024

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15*(9), 1277-1288. https://doi.org/10.1177/1049732305276687

Huang, F., Kwak, H., & An, J. (2023, April). Is chatgpt better than human annotators? Potential and limitations of chatgpt in explaining implicit hate speech. *In Companion Proceedings of the ACM Web Conference* 2023 (pp. 294-297). https://doi.org/10.1145/3543873.3587368

Jiao, W., Wang, W., Huang, J. T., Wang, X., Shi, S., & Tu, Z. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745. https://doi.org/10.48550/arXiv.2301.08745

Kaliisa, R., Rienties, B., Mørch, A. I., & Kluge, A. (2022). Social learning analytics in computer-supported collaborative learning environments: A systematic review of empirical studies. *Computers and Education Open*, 3, 100073. https://doi.org/10.1016/j.caeo.2022.100073

Kent, C., & Rechavi, A. (2020). Deconstructing online social learning: network analysis of the creation, consumption and organization types of interactions. *International Journal of Research & Method in Education*, 43(1), 16-37. https://doi.org/10.1080/1743727X.2018.1524867

Kiemer, K., Gröschner, A., Pehmer, A. K., & Seidel, T. (2015). Effects of a classroom discourse intervention on teachers' practice and students' motivation to learn mathematics and science. *Learning and Instruction*, 35, 94-103. https://doi.org/10.1016/j.learninstruc.2014.10.003

Knight, S., & Littleton, K. (2015). Discourse-centric learning analytics: mapping the terrain. *Journal of Learning Analytics*, 2(1), 185-209. https://doi.org/10.18608/jla.2015.21.9

Knight, S., Shum, S. B., & Littleton, K. (2014). Epistemology, assessment, pedagogy: Where learning meets analytics in the middle space. *Journal of Learning Analytics*, 1(2), 23-47. https://doi.org/10.18608/jla.2014.12.3

Kuhn, D. (2018). A role for reasoning in a dialogic approach to critical thinking. *Topoi*, 37, 121-128. https://doi.org/10.1007/s11245-016-9373-4

Kuhn, D. (2015). Thinking together and alone. *Educational researcher*, 44(1), 46-53. https://doi.org/10.3102/0013189X15569530

Lapadat, J. (2007). Discourse devices used to establish community, increase coherence, and negotiate agreement in an online university course. *International Journal of E-Learning & Distance Education/Revue Internationale du E-Learning et la Formation à Distance*, 21(3), 59-92. https://files.eric.ed.gov/fulltext/EJ805057.pdf

Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: *A systematic review. British Journal of Educational Technology*, 50(5), 2594–2618. https://doi.org/10.1111/bjet.12720

Li, H., Hao, Y., Zhai, Y., & Qian, Z. (2023, November). Assisting static analysis with large language models: A chatgpt experiment. *In Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 2107-2111). https://doi.org/10.1145/3611643.3613078

Littleton, K., & Whitelock, D. (2005). The negotiation and co-construction of meaning and understanding within a postgraduate online learning community. *Learning, Media and Technology*, 30(2), 147-164. https://doi.org/10.1080/17439880500093612

Luginbühl, M., Mundwiler, V., Kreuz, J., Müller-Feldmeth, D., & Hauser, S. (2021). Quantitative and qualitative approaches in conversation analysis: Methodological reflections on a study of argumentative group discussions. *Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion*, 22, 179-236. https://doi.org/10.1007/s10212-024-00840-7

Mercer, N. (1996). The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction*, 6(4), 359-377. https://doi.org/10.1016/S0959-4752(96)00021-7

Mercer, N. (2004). Sociocultural discourse analysis. *Journal of Applied Linguistics*, 1(2), 137-168. https://doi.org/10.1558/japl.2004.1.2.137

Michaels, S., & O'Connor, C. (2012). Talk science primer. https://inquiryproject.terc.edu/shared/pd/TalkScience_Primer.pdf

Michaels, S., O'Connor, M. C., Hall, M. W., & Resnick, L. B. (2010). Accountable talk® sourcebook. Pittsburgh, PA: *Institute for Learning, University of Pittsburgh*. Murphy, PK, Wilkinson, IAG, Soter, AO, Hennessey, MN, & Alexander, JF. https://doi.org/10.1007/s11217-007-9071-1

Min, W., Park, K., Wiggins, J., Mott, B., Wiebe, E., Boyer, K. E., & Lester, J. (2019). Predicting dialogue breakdown in conversational pedagogical agents with multimodal LSTMs. *In Artificial Intelligence in Education*: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20 (pp. 195-200). Springer International Publishing.

https://doi.org/10.1007/978-3-030-23207-8_37

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems* (TiiS), 11(3-4), 1-45. https://doi.org/10.1145/3387166

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. https://www.diva-portal.org/smash/get/diva2:1042586/FULLTEXT01.pdf

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744. https://doi.org/10.48550/arXiv.2203.02155

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health services research*, *34*(5 Pt 2), 1189. https://pubmed.ncbi.nlm.nih.gov/10591279/

Phillipson, N., & Wegerif, R. (2019). The thinking together approach to dialogic teaching. *In Deeper learning, dialogic learning, and critical thinking* (pp. 32-47). Routledge. https://doi.org/10.4324/9780429323058

Pilny, A., McAninch, K., Slone, A., & Moore, K. (2019). Using supervised machine learning in automated content analysis: An example using relational uncertainty. *Communication Methods and Measures*, 13(4), 287-304. https://doi.org/10.1080/19312458.2019.1650166

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., & Morency, L. P. (2017, July). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 873-883). https://doi.org/10.18653/v1/P17-1081

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9. https://doi.org/10.1080/19312458.2019.1650166

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Computer Science*, 2(3), 173. https://doi.org/10.1007/s42979-021-00557-0

Schmidt, D. C., Spencer-Smith, J., Fu, Q., & White, J. (2024). Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *ACM SIGAda Ada Letters*, 43(2), 43-51. https://doi.org/10.1145/3672359.3672364

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, 47, 761-773. https://doi.org/10.1007/s11135-011-9545-7

Scharkow, M. (2017). Content analysis, automatic. *The international encyclopedia of communication research methods*, 1-14. https://doi.org/10.1002/9781118901731.iecrm0043

Schrire, S. (2004). Interaction and cognition in asynchronous computer conferencing. *Instructional Science*, 32(6), 475-502. https://doi.org/10.1007/s11251-004-2518-7

Samei, H. Li, F. Keshtkar, V. Rus, and A. C. Graesser. Context-based speech act classification in intelligent tutoring systems. In *International conference on intelligent tutoring systems*, pages 236–241. Springer, 2014. https://doi.org/10.1007/978-3-319-07221-0_28

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27(2), 4-13. https://doi.org/10.3102/0013189X027002004

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., ... & Baker, R. S. (2011). Open Learning Analytics: an integrated & modularized platform Proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques. *Society for Learning Analytics Research*, 1-19. https://doi.org/10.18608/jla.2018.51.7

Singh, R., Miller, T., Lyons, H., Sonenberg, L., Velloso, E., Vetere, F., ... & Dourish, P. (2023). Directive explanations for actionable explainability in machine learning applications. *ACM Transactions on Interactive Intelligent Systems*, 13(4), 1-26. https://doi.org/10.48550/arXiv.2102.02671

Siiman, L. A., Rannastu-Avalos, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023, August). Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In *International Conference on Innovative Technologies and Learning* (pp. 87-96). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40113-8_9

Spinner, T., Kehlbeck, R., Sevastjanova, R., Stähle, T., Keim, D. A., Deussen, O., & El-Assady, M. (2024). -generAItor: Tree-in-the-loop Text Generation for Language Model Explainability and Adaptation. *ACM Transactions on Interactive Intelligent Systems*, 14(2), 1-32. https://doi.org/10.1145/3652028

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33, 3008-3021. https://doi.org/10.48550/arXiv.2009.01325

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23, 16094069241231168. https://doi.org/10.1177/16094069241231168

Turobov, A., Coyle, D., & Harding, V. (2024). Using ChatGPT for thematic analysis. *arXiv preprint arXiv*:2405.08828. https://doi.org/10.48550/arXiv.2405.08828

Ubani, S., & Nielsen, R. (2022, July). Classifying different types of talk during collaboration. *In International Conference on Artificial Intelligence in Education* (pp. 227-230). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-11647-6_40

Vainio-Pekka, H., Agbese, M. O. O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The role of explainable AI in the research field of AI ethics. *ACM Transactions on Interactive Intelligent Systems*, 13(4), 1-39. https://doi.org/10.1145/3599974

Velez, G., Taylor, L. K., & Power, S. A. (2022). Developing in a dynamic world. *International Perspectives in Psychology.* https://doi.org/10.1027/2157-3891/a000038

Vieira, C., Parsons, P., & Byrd, V. (2018). Visual learning analytics of educational data: A systematic literature review and research agenda. *Computers & Education*, 122, 119-135. https://doi.org/10.1016/j.compedu.2018.03.018

Vygotsky, L. S., & Cole, M. (1978). Mind in society: *Development of higher psychological processes*. Harvard university press.

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2024). Prompts, Pearls, Imperfections: Comparing ChatGPT and a Human Researcher in Qualitative Data Analysis. *Qualitative Health Research*, 10497323241244669. https://doi.org/10.1177/10497323241244669

Wang, S., & Jin, P. (2023). A brief summary of prompting in using gpt models. Qeios. https://doi.org/10.32388/IMZI2Q

Wang, X., & Yin, M. (2022). Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems, 12(4), 1-36.* https://doi.org/10.1145/3519266

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. *Computers & Education*, 78, 115-123. https://doi.org/10.1016/j.compedu.2014.05.010

Webb, N. M., Franke, M. L., Ing, M., Turrou, A. C., Johnson, N. C., & Zimmerman, J. (2019). Teacher practices that promote productive dialogue and learning in mathematics classrooms. *International Journal of Educational Research*, 97, 176-186. https://doi.org/10.1016/j.ijer.2017.07.009

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

West, C. G. (2023). AI and the FCI: Can ChatGPT project an understanding of introductory physics?. *arXiv preprint arXiv*:2303.01067. https://doi.org/10.48550/arXiv.2303.01067

Wu, T., Terry, M., & Cai, C. J. (2022, April). Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (pp. 1-22). https://doi.org/10.48550/arXiv.2110.01691

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. (2023, March). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 75-78). https://doi.org/10.1145/3581754.3584136

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, 3(4). https://doi.org/10.1016/j.patter.2022.100455

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2023). Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. arXiv preprint *arXiv:2309.10771*. https://doi.org/10.48550/arXiv.2309.10771

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning* (pp. 12697-12706). PMLR. https://arxiv.org/pdf/2102.09690

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023, April). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-21). https://doi.org/10.1145/3544548.3581388

Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students*, 29(3), 42-46. https://doi.org/10.1145/3589649

Appendix A. Codebook

Goal	Criteria	Example
Goal 1: Individual students share, expand, and clarify their own thoughts.	➢ Students share their ideas, knowledge, and observations about the digestion concepts and it related concepts with their peers without explaining the underlying reasons.	Student-1 2022-12-05 08:36:57 but it doesn't have to be good it's what we think Student-1 2022-12-05 08:37:17 so what do you think
	➢ Students exchange the answers to the questions on task about the digestion concepts and it related concepts with peers without mentioning or deepening the reason.	Student-1 2022-12-05 08:36:57 but it doesn't have to be good it's what we think Student-1 2022-12-05 08:37:17 so what do you think
	➢ Students explain unknown words or terms.	Student-2 2022-12-12 09:31:53 we have to do those hypotheses Student-1 2022-12-12 09:32:00 yes Student-2 2022-12-12 09:32:37 we should make sentences with those words
Goal 2: Students listen carefully to one another	➢ Students give solutions or reactions to another student's query. ➢ Students answer to the questions asked by their peers. ➢ Students comment on peers' thoughts by saying words such as "yes" or "good", I think so, I don't think so, that's right, I don't know if it is right.	Student-2 2022-12-05 08:28:48 you should vgm go to the digestive system and then down there it says no assignment Student-2 2022-12-05 08:28:53 One Student-1 2022-12-05 08:29:09 Oh yes see it

Goal	Criteria	Example
	➢ Student summarize peers' statements by repeating and rephrasing the main points said by their partners. ➢ Students apply another student's solutions and confirms its success.	Student-2 2022-12-12 08:34:55 where my arrow is you don't have to fill in anything Student-2 2022-12-12 08:35:14 only here below is what you do have to fill out Student-3 2022-12-12 08:35:38 that's the way I think it is right
Goal 3: Students deepen their reasoning	➢ Students deepen their understanding as they proceed with the task and reflect on their thinking about the concept of digestion and related concepts. ➢ Students explain the reasoning behind and justify sharing ideas related to digestion. ➢ Students explain how they arrived at such a conclusion. ➢ Students think about alternative conditions and explain how they might change the outcomes or understandings.	Student-1 2022-12-05 08:36:28 me neither... Student-2 2022-12-05 08:36:41 ow Student-1 2022-12-05 08:36:57 but it doesn't have to be good it's what we think Student-1 2022-12-05 08:37:17 so what do you think Student-2 2022-12-05 08:37:19 yes Student-1 2022-12-05 08:38:23 "I think then it's easier to digest and it also fits through the intestines and so on"
Goal 4: Students Engage with Others' Reasoning	➢ Student asks questions about their classmates' ideas about the task. ➢ Student uses a peer's idea to solve a problem or justify their own claims about digestion and its related concepts. ➢ Student clearly expresses agreement or disagreement with their peer's statements and provides reasons for their perspectives about	Student-2 2022-12-05 09:46:46 because otherwise you have a big bump in your stomach and then suddenly nothing which makes you suddenly very hungry right Student-1 2022-12-05 09:47:24 no it is because otherwise the nutrients cannot pass through your blood

Goal	Criteria	Example
	digestion and its related concepts. ➢ Expresses lack of understanding, asking for more clarification and further peer discussion about digestion and its related concepts. ➢ Students reflect on what their peers say by adding more depth and bringing up new perspectives to their opinions about digestion and its related concepts. ➢ Students compare and contrast their ideas with their peers to identify similarities and differences. ➢ Students take a suggestion provided by a peer and extend it.	Student-2 2022-12-12 08:21:03 do you get this Student-1 2022-12-12 08:21:24 think you should make a sentence Student-2 2022-12-12 08:21:27 yes a hypothesis

Appendix B. Prompt Model

Goal 1

Instruction:
Your role is to be an academic expert in Qualitative deductive Analysis, aiming to help teachers. You will assist in detecting students' moves and interpreting results. Your analysis should focus on accuracy, relevance, and depth while avoiding giving personal opinions.
You will follow step-by-step instructions to respond to user inputs:
Step 1: Read the following text explaining the context in which the dialogues occur.
Context: Each two students is automatically paired. They work together and consult with each other using chat. They work on the science with the topic of digestion assignment in an online environment. The assignment includes seven sections. The upcoming lessons will cover digestion within the human digestive system, focusing on how nutrients from food are absorbed into the blood. Simple nutrients like glucose, minerals, water, and vitamins can directly pass through the intestinal wall into the bloodstream. However, complex nutrients such as proteins, fats, and most carbohydrates need to be broken down into simpler forms that can be absorbed. This breakdown is achieved through chemical digestion, where digestive juices convert substances into absorbable molecules. This process relies on the key-lock principle, with enzymes designed to target specific molecules. For instance, the enzyme in saliva specifically breaks down starch following this principle, and enzymes in digestive juices overall expedite the digestion process.
The students will also explore what conditions this process works best we will find out in these two lessons.
The students first are shown the image of the human body where they are tasked to drop and drag the key terms involved in the digestive system such as Stomach juice, bile, pancreatic juice, and intestinal juice. Next in the following task titled "chewing food", they should discuss and think together why dividing food into small pieces is good for digestion.
Next in the task named "the action of salvia", they are assigned to complete 10 sentences and fill in the blanks using the words " enzyme, water, chewing, blood, digestion products, key lock principle, fats, nutrient, proteins, enzyme, nutrient. The answer is the following: 1. Enzyme, 2. Nutrients, Digestion product, 3. Enzyme, enzyme, nutrients, 4. Blood, 5. Fats, protein, 6. Water, 7. Chewing, 8. Key-lock principle. In the following task titled " What do you think", the students should make a hypothesis, dragging the words provided such as IF, Then, breaks off, is high, is low, doesn't break off, is around 37 degrees, the, a, an, starch, salvia, temperature, water.
The next task was " your own research" The students should test their hypothesis by doing four experiments designed in the digital lab.
The experiment requires the students to complete the table, writing down the temperature and the starch leftovers under each temperature, and then in the next section titled "What we learned" answer four questions: 1. Describe three examples of how enzymes play an important role in our digestion. 2. Saliva contains an enzyme that breaks down starch. Why can this enzyme only break down starch? 3. What happens to the way enzymes work when the temperature is higher than the optimal temperature? 4. What happens to enzyme activity when the temperature is lower than 37 degrees Celsius? Can you think of an explanation for this?
After uploading the dialogue, follow step 2:
Code Identification
Step 2. Process:
1. Read and comprehend the user's uploaded dialogue
2. Analyze the entire dialogue and identify

Where

- Students share their ideas, knowledge, and observations about the digestion concepts and it related concepts with their peers without explaining the underlying reasons.
- Students exchange the answers to the questions on the task about the digestion concepts and it related concepts with peers without mentioning or deepening the reason.
- Students explain unknown words or terms.

3. Code “Goal 1” if one of the above goals is met or “Not labeled” if not.
4. Provide a clear reason for why the utterance has been identified in a certain way.
5. Read the following examples.

Example 1:	Student-1:	2022-12-05 08:36:57 but it doesn't have to be good it's what we think
	Student-1:	2022-12-05 08:37:17 so what do you think
Label: Goal 1		
Reason:	shares the understanding of what she got from the procedure of the task	
Example 2:	Student-2:	2022-12-12 09:53:36 when does starch not break down?
	Student-1:	2022-12-12 09:55:37 when the temperature is below 37 degrees
Label: Goal 1		
Reason:	Student 2 shares their understanding of the experiment.	
Example 3:	Student-2:	2022-12-12 09:31:53 we have to do those hypotheses
	Student-1	2022-12-12 09:32:00 yes
	Student-2	2022-12-12 09:32:37 we should make sentences with those words
Label: Goal 1		
Reason:	Student 2 shares the idea and clarifies the term.	

Output:	Present a table with the following columns:
Column 1:	Username
Column 2:	Timestamp
Column 3:	Utterance
Column 4:	Criteria
Column 5:	Reason

Goal 2

Instruction:
Your role is to be an academic expert in Qualitative deductive Analysis, aiming to help teachers. You will assist in qualitative text analysis, coding data, detecting students' moves, and interpreting results. You should focus on accuracy, relevance, and depth in your analysis while avoiding giving personal opinions.
You will follow step-by-step instructions to respond to user inputs: Step 1: Read the following text explaining the context in which the dialogues occur. Context: Each two students is automatically paired. They work together and consult with each other using chat. They work on the science with the topic of digestion assignment in an online environment. The assignment includes seven sections. The upcoming lessons will cover digestion within the human digestive system, focusing on how nutrients from food are absorbed into the blood. Simple nutrients like glucose, minerals, water, and vitamins can directly pass through the intestinal wall into the bloodstream. However, complex nutrients such as proteins, fats, and most carbohydrates need to be broken down into simpler forms that can be absorbed. This breakdown is achieved through chemical digestion, where digestive juices convert substances into absorbable molecules. This process relies on the key-lock principle, with enzymes designed to target specific molecules. For instance, the enzyme in saliva specifically breaks down starch following this principle, and enzymes in digestive juices overall expedite the digestion process.
The students will also explore what conditions this process works best we will find out in these two lessons.
The students first are shown the image of the human body where they are tasked to drop and drag the key terms involved in the digestive system such as Stomach juice, bile, pancreatic juice, and intestinal juice. Next in the following task titled "chewing food", they should discuss and think together why dividing food into small pieces is good for digestion.
Next in the task named "the action of salvia", they are assigned to complete 10 sentences and fill in the blanks using the words " enzyme, water, chewing, blood, digestion products, key lock principle, fats, nutrient, proteins, enzyme, nutrient. The answer is the following: 1. Enzyme, 2. Nutrients, Digestion product, 3. Enzyme, enzyme, nutrients, 4. Blood, 5. Fats, protein, 6. Water, 7. Chewing, 8. Key-lock principle. In the following task titled " What do you think", the students should make a hypothesis, dragging the words provided such as IF, Then, breaks off, is high, is low, doesn't break off, is around 37 degrees, the, a, an, starch, salvia, temperature, water. The next task tilted " your own research" The students should test their hypothesis by doing four experiments designed in the digital lab. The experiment requires the students to complete the table, writing down the temperature and the starch leftovers under each temperature and then in the next section titled "What we learned" answer four questions: 1. Describe three examples of how enzymes play an important role in our digestion. 2. Saliva contains an enzyme that breaks down starch. Why can this enzyme only break down starch? 3. What happens to the way enzymes work when the temperature is higher than the optimal temperature? 4. What happens to enzyme activity when the temperature is lower than 37 degrees Celsius? Can you think of an explanation for this?
After uploading the dialogue, follow step 2:
Code Identification
Step 2: Process:
1. Read and comprehend the user's uploaded dialogue
2. Analyze the entire dialogue and identify
Where

- The student gives solutions or reactions to another student's query.

- The student answers to the questions asked by their peers.
- The student comments on peers' thoughts by saying words such as "yes" or "good", I think so, I don't think so, that's right, I don't know if it is right.
- The student summarizes peers' statements by repeating and rephrasing the main points said by their partners.
- The student applies another student's solutions and confirms its success.

3. Code "Students listen carefully to each other" if one of the above goals is met or "Not labeled" if not.
4. Provide a clear reason for why the utterance has been identified in a certain way.
5. Read the following examples.

Example 1:	Student-2:	2022-12-12 08:33:00 you had to do them one at a time but that doesn't matter now anyway
Label:	Students listen carefully to each other	
Reason:	Student-2 listens carefully by commenting on other peers' thoughts while giving and sharing new insights.	
Example 2:	Student-2:	2022-12-05 08:41:06 yes but don't know if it's any good
Label:	Students listen carefully to each other	
Reason:	Student 2 listens carefully by expressing their doubts.	
Example 3:	Student-2:	2022-12-05 08:28:48 you should vgm go to the digestive system and then down there it says no assignment
	Student-2:	2022-12-05 08:28:53 one
	Student-1:	2022-12-05 08:29:09 oh yes see it
Label:	Students listen carefully to each other	
Reason:	Student 2 is listening carefully by providing relevant answers to the student 1 question.	
Label:	Students listen carefully to each other	
Reason:	Student 1 is listening carefully by applying the Student 2 approach and confirming its success.	

Output:	Present a table with the following columns:
Column 1:	Username
Column 2:	Timestamp
Column 3:	Utterance
Column 4:	Criteria
Column 5:	Reason

Goal 3

Instruction:
Your role is to be an academic expert in Qualitative deductive Analysis, aiming to help teachers. You will assist in qualitative text analysis, coding data, detecting students' moves, and interpreting results. Your analysis should focus on accuracy, relevance, and depth while avoiding giving personal opinions. You will follow step-by-step instructions to respond to user inputs:
Step 1: Read the following text explaining the context in which the dialogues occur. Context: Each two students is automatically paired. They work together and consult with each other using chat. They work on the science with the topic of digestion assignment in an online environment. The assignment includes seven sections. The upcoming lessons will cover digestion within the human digestive system, focusing on how nutrients from food are absorbed into the blood. Simple nutrients like glucose, minerals, water, and vitamins can directly pass through the intestinal wall into the bloodstream. However, complex nutrients such as proteins, fats, and most carbohydrates need to be broken down into simpler forms that can be absorbed. This breakdown is achieved through chemical digestion, where digestive juices convert substances into absorbable molecules. This process relies on the key-lock principle, with enzymes designed to target specific molecules. For instance, the enzyme in saliva specifically breaks down starch following this principle, and enzymes in digestive juices overall expedite the digestion process.
The students will also explore what conditions this process works best we will find out in these two lessons.
The students first are shown the image of the human body where they are tasked to drop and drag the key terms involved in the digestive system such as Stomach juice, bile, pancreatic juice, and intestinal juice. Next in the following task titled "chewing food", they should discuss and think together why dividing food into small pieces is good for digestion.
Next in the task named "the action of salvia", they are assigned to complete 10 sentences and fill in the blanks using the words " enzyme, water, chewing, blood, digestion products, key lock principle, fats, nutrient, proteins, enzyme, nutrient. The answer is the following: 1. Enzyme, 2. Nutrients, Digestion product, 3. Enzyme, enzyme, nutrients, 4. Blood, 5. Fats, protein, 6. Water, 7. Chewing, 8. Key-lock principle. In the following task titled " What do you think", the students should make a hypothesis, dragging the words provided such as IF, Then, breaks off, is high, is low, doesn't break off, is around 37 degrees, the, a, an, starch, salvia, temperature, water. The next task tilted " your own research" The students should test their hypothesis by doing four experiments designed in the digital lab. The experiment requires the students to complete the table, writing down the temperature and the starch leftovers under each temperature and then in the next section titled "What we learned" answer four questions: 1. Describe three examples of how enzymes play an important role in our digestion. 2. Saliva contains an enzyme that breaks down starch. Why can this enzyme only break down starch? 3. What happens to the way enzymes work when the temperature is higher than the optimal temperature? 4. What happens to enzyme activity when the temperature is lower than 37 degrees Celsius? Can you think of an explanation for this?
After uploading the dialogue, follow step 2:
Code Identification
Step 2. Process:
1. Read and comprehend the user's uploaded dialogue
2. Analyze each utterance of the entire dialogue and identify

Where

- Students deepen their understanding as they proceed with the task and reflect on their thinking about the concept of digestion and related concepts.

- Students explain the reasoning behind and justify sharing ideas related to digestion.
- Students explain how they arrived at such a conclusion.
- Students think about alternative conditions and explain how they might change the outcomes or understandings.

3. Code “Students deepen their reasoning” if one of the above goals is met or “ Not labeled” if not.
4. Provide a clear reason for why the utterance has been identified in a certain way.
5. Read the following example:

Example 1:	Student-2:	Student-2 2022-12-12 08:40:46 [I do not know]
	Student-3:	2022-12-12 08:42:23 dunno don't think so
	Student-2	2022-12-12 08:42:54 [is it not with the previous one that instead of little starch
Label:	Students deepen their reasoning	
Reason:	The student 2 deepens the reasoning by reflecting, challenging, and explaining an alternative condition and how they might change the outcomes or understandings and could deepen the reasoning from “I do not know” gradually to “ is it not with the previous one that instead of little starch it has no starch” and “because otherwise, it should be a little darker”.	
Example 2:	Student-1:	2022-12-05 08:36:28 me neither...
	Student-2:	2022-12-05 08:36:41 ow
	Student-1:	2022-12-05 08:36:57 but it doesn't have to be good it's what we think
	Student-1:	2022-12-05 08:36:57 but it doesn't have to be good it's what we think
	Student-2:	2022-12-05 08:37:19 yes
	Student-1:	2022-12-05 08:38:23 “I think then it's easier to digest and it also fits through the intestines and so on”
Label:	Students deepen their reasoning.	
Reason:	Student 1 deepens the reasoning by going through the dialogue from “ I don’t know” to “ “I think then it's easier to digest and it also fits through the intestines and so on”	
Example 3:	Student-2	2022-12-05 08:36:57 Dividing pieces is more convenient so that the intestines have to divide less

Lebel:	Students deepen their reasoning.
Reason:	The student 2 explains the reasoning of their understanding

---

Output:	Present a table with the following columns:
Column 1:	Username
Column 2:	Timestamp
Column 3:	Utterance
Column 4:	Criteria
Column 5:	Reason

Goal 4

Instruction:

Your role is to be an academic expert in Qualitative deductive Analysis, aiming to help teachers. You will assist in qualitative text analysis, coding data, detecting students' moves, and interpreting results. You should focus on accuracy, relevance, and depth in your analysis while avoiding giving personal opinions.

You will follow step-by-step instructions to respond to user inputs:

Step 1: Read the following text explaining the context in which the dialogues occur.

Context: Each two students is automatically paired. They work together and consult with each other using chat. They work on the science with the topic of digestion assignment in an online environment. The assignment includes seven sections. The upcoming lessons will cover digestion within the human digestive system, focusing on how nutrients from food are absorbed into the blood. Simple nutrients like glucose, minerals, water, and vitamins can directly pass through the intestinal wall into the bloodstream. However, complex nutrients such as proteins, fats, and most carbohydrates need to be broken down into simpler forms that can be absorbed. This breakdown is achieved through chemical digestion, where digestive juices convert substances into absorbable molecules. This process relies on the key-lock principle, with enzymes designed to target specific molecules. For instance, the enzyme in saliva specifically breaks down starch following this principle, and enzymes in digestive juices overall expedite the digestion process.

The students will also explore what conditions this process works best we will find out in these two lessons.

The students first are shown the image of the human body where they are tasked to drop and drag the key terms involved in the digestive system such as Stomach juice, bile, pancreatic juice, and intestinal juice. Next in the following task titled "chewing food", they should discuss and think together why dividing food into small pieces is good for digestion.

Next in the task named "the action of salvia", they are assigned to complete 10 sentences and fill in the blanks using the words " enzyme, water, chewing, blood, digestion products, key lock principle, fats, nutrient, proteins, enzyme, nutrient. The answer is the following: 1. Enzyme, 2. Nutrients, Digestion product, 3. Enzyme, enzyme, nutrients, 4. Blood, 5. Fats, protein, 6. Water, 7. Chewing, 8. Key-lock principle. In the following task titled " What do you think", the students should make a hypothesis, dragging the words provided such as IF, Then, breaks off, is high, is low, doesn't break off, is around 37 degrees, the, a, an, starch, salvia, temperature, water.

The next task tilted " your own research" The students should test their hypothesis by doing four experiments designed in the digital lab.

The experiment requires the students to complete the table, writing down the temperature and the starch leftovers under each temperature and then in the next section titled "What we learned" answer four questions: 1. Describe three examples of how enzymes play an important role in our digestion. 2. Saliva contains an enzyme that breaks down starch. Why can this enzyme only break down starch? 3. What happens to the way enzymes work when the temperature is higher than the optimal temperature? 4. What happens to enzyme activity when the temperature is lower than 37 degrees Celsius? Can you think of an explanation for this?

After uploading the dialogue, follow step 2:

Code Identification

Step 2. Process:

1. Read and comprehend the user's uploaded dialogue
2. Analyze the entire dialogue and identify

Where

- The student asks questions about their classmates' ideas about the task.

- Student uses a peer's idea to solve a problem or justify their own claims about digestion and its related concepts.
- The student clearly expresses agreement or disagreement with their peer's statements and provides reasons for their perspectives about digestion and its related concepts.
- Expresses lack of understanding, asking for more clarification and further peer discussion about digestion and its related concepts.
- Students reflect on what their peers say by adding more depth and bringing up new perspectives to their opinions about digestion and its related concepts.
- Students compare and contrast their ideas with their peers to identify similarities and differences.
- Students take a suggestion provided by a peer and extend it.

3. Code "Students engage with other peer's reasoning and thoughts" if one of the above goals is met or "Not labeled" if not.
4. Provide a clear reason for why the utterance has been identified in a certain way, directly linking back to the definitions and examples.
5. Read the following examples.

Example 1:	Student-2:	2022-12-05 09:46:46 because otherwise you have a big bump in your stomach and then suddenly nothing which makes you suddenly very hungry right
	Student-1	2022-12-05 09:47:24 no it is because otherwise the nutrients cannot pass
Label:	Students engage with other peer's reasoning and thoughts	
Reason:	The students engaged with each other's reasoning by disagreeing, and explaining the reason challenging ideas." So this utterance can be coded as " Students Engage with Others' Reasoning	
Example 2:	Student-2:	2022-12-12 08:21:03 do you get this
	Student-1:	2022-12-12 08:21:24 think you should make a sentence
	Student-2:	2022-12-12 08:21:27 yes a hypothesis
Label:	Students engage with other peer's reasoning and thoughts	
Reason:	The students engage with other peers' reasoning by asking for information and clarification to understand unfamiliar words or concepts, and by trying to make sense of others' reasoning.	
Example 3:	Student-1	2022-12-12 08:25:20 if the temperature is around 37 degrees then the starch breaks down the saliva

	Student-2	2022-12-12 08:25:28 we can also add another hypothesis to it
	Student-2	2022-12-12 08:25:35 yes
Label:	Students engage with other peer's reasoning and thoughts	
Reason:	The student 2 engages with the other student's reasoning by reflecting on the other's idea.	

Output:	Present a table with the following columns:
Column 1:	Username
Column 2:	Timestamp
Column 3:	Utterance
Column 4:	Criteria
Column 5:	Reason