Towards a large-language-model-based chatbot system to automatically monitor student goal setting and planning in online learning

Author(s): Khe Foon Hew, Weijiao Huang, Sikai Wang, Xinyi Luo and Donn Emmanuel Gonda

Source: *Educational Technology & Society*, July 2025, Vol. 28, No. 3 (July 2025), pp. 112-132

Published by: International Forum of Educational Technology & Society, National Taiwan Normal University, Taiwan

Stable URL: https://www.jstor.org/stable/10.2307/48827953

Hew, K. F., Huang, W., Wang, S., Luo, X., Gonda, D. E. (2025). Towards a large-language-model-based chatbot system to automatically monitor student goal setting and planning in online learning. *Educational Technology & Society*, *28*(3), 112-132. https://doi.org/10.30191/ETS.202507_28(3).SP08

# Towards a large-language-model-based chatbot system to automatically monitor student goal setting and planning in online learning

**Khe Foon Hew***, **Weijiao Huang, Sikai Wang, Xinyi Luo and Donn Emmanuel Gonda**
The University of Hong Kong, Hong Kong SAR // kfhew@hku.hk // wjhuang1@connect.hku.hk // sikaiw@connect.hku.hk // xinyiluo@hku.hk // dgonda@hku.hk
*Corresponding author

**ABSTRACT:** Despite the prevalence of online learning, the lack of student self-regulated learning (SRL) skills continues to be persistent issue. To support students' SRL, teachers can prompt with SRL-related questions and provide timely, personalized feedback. Providing timely, personalized feedback to each student in large classes, however, can be labor-intensive for teachers. This 2-stage study offers a novel contribution by developing a Large Language Model (LLM)-based chatbot system that can automatically monitor students' goal setting and planning in online learning. Goal setting and planning are two important skills that can occur in all SRL phases. In stage 1, we developed the Goal-And-Plan-Mentor ChatGPT system (*GoalPlanMentor*) by creating an SRL knowledge base with goal and plan indicators, using Memory-Augmented-Prompts to automatically detect student goals and plans, and providing personalized feedback. In stage 2, we compared the accuracy of *GoalPlanMentor*'s detection (coding) of students' goals and plans with human coders, examined the quality of *GoalPlanMentor*'s feedback, and students' perceptions about the usefulness of *GoalPlanMentor*. Results show substantial to near perfect agreement between *GoalPlanMentor*'s and human's coding, and high quality of *GoalPlanMentor*'s feedback in terms of providing clear directions for improvement. Overall, students perceived *GoalPlanMentor* to be useful in setting their goals and plans, with average values being significantly higher than the midpoint of the scale. Students who highly perceived the system's usefulness for goal-setting exhibited significantly greater learning achievements compared to those with a low perception of its usefulness. Implications for future research are discussed.

**Keywords:** Generative artificial intelligence, Chatbot, Self-regulated learning, Online learning, Large language models

## 1. Introduction

The COVID-19 pandemic has dramatically shifted online learning into the mainstream of education. Even after the pandemic has ended, online learning will remain an option for students, offering flexibility and convenience (Dos Santos, 2022). Online learning is also an essential component of blended learning which combines face-to-face and online activities. The use of online learning will expand since blended learning will become more important in a post-COVID world (Singh et al., 2021)

Although online learning has benefits, the lack of student self-regulated learning (SRL) skills is concerning. In online or blended courses, resources are placed in learning management systems for on-demand use (Pedrotti & Nistor, 2019). With students determining the 'when,' 'how,' and 'where' of learning, self-regulation becomes crucial (Pedrotti & Nistor, 2019; Rasheed et al., 2020). Poor self-regulation can lead to unrelated online activities (Rasheed et al., 2020), negatively impacting learning.

SRL involves individuals intentionally planning, monitoring, reflecting, and adapting their learning progress to achieve learning goals (Pintrich, 2000; Zimmerman, 2002). Students with strong SRL tend to perform better (Cheng et al., 2023). Training students in SRL is thus crucial for enhancing online learning. Scholars identify multiple SRL phases (Panadero, 2017), with Wolters and Brady (2021) proposing three main phases based on prominent models: forethought, performance, and post-performance.

In the first phase, forethought, students prepare for learning (Zimmerman, 2002). Key processes in this stage include goal setting and planning. Setting goals is vital for student self-regulation and successful completion of courses (Handoko et al., 2019). According to Latham and Locke (1991), "goal setting facilitates self-regulation in that the goal defines for the person what constitutes an acceptable level of performance" (p. 234). During goal setting, a student determines the specific goals of their learning effort as specific goals promote more consistent student performance than vague goals (Latham & Locke, 1991). The learning goals should also be achievable rather than infeasible. Many students with lower self-regulation skills have unrealistic goals, which may disappoint the learners after their learning (Shih et al., 2010). Goal setting and planning are closely related to each other as students first define their learning goals and then propose how they plan to reach their goals (Nussbaumer et al., 2011).

In the second phase, performance, students engage in academic work, such as completing writing tasks on the learning management system. Monitoring, a key process in this phase, involves students' ongoing awareness of their actions and adaptation during academic work (Usher & Schunk, 2017). It helps students plan and set goals for future efforts more effectively (Zimmerman & Paulsen, 1995). For example, monitoring writing behavior, during the early stages of a semester, can help students plan subsequent writing sessions, evaluate the effectiveness of writing strategies, and decide on future writing goals.

In the third phase, post-performance, students engage in self-regulation by assessing and reflecting on their learning performance (Usher & Schunk, 2017). Described as self-judgment (Zimmerman, 2002), this phase involves people evaluating the outcomes of their efforts in light of previously established goals (Wolters & Brady, 2021). Depending on whether their performance met their goals, people may need to revise their future goals and plans in order to better achieve their desired outcomes.

Given the importance of SRL, scholars have explored various strategies to promote students' SRL in online learning. One strategy is to use reflective writing exercise to foster students' awareness of SRL. For example, students completed a weekly planning exercise by stating their learning goals of the week, the time spent on the previous week's course activities, and things they learned in the previous week (Pérez-Sanagustín et al., 2021). Butzler (2016) employed a document, called *exam wrapper,* asking students to answer questions such as how they prepared for the exam.

Another strategy is to embed SRL-related prompts into video lectures. Moos and Bonde (2016) embedded planning prompts (e.g., What do you already know about [this topic]?) at the start of the video, monitoring prompts (e.g., Do you need to adjust how you are learning?) halfway through the video, and reflection prompts (e.g., Do you need to go back in the video and fill any gaps in understanding) at the end of the video. Students verbalized their replies to these prompts. Van Alten et al. (2020a, 2020b) also embedded SRL prompts (e.g., What are your goals when learning from this video?) into video lectures. Students answered the questions in order to continue the video. In about half the SRL prompts, hints would pop up after a student's answer to show an example of that SRL behavior.

These strategies have two drawbacks. First, they often lack timely feedback after students complete reflective writing activities. Hattie and Timperley (2007) define feedback as providing information about students' performance regarding what they attempted. Without timely feedback, less proficient self-regulating students may struggle to know their next steps. Feedback should be within 48 hours to prevent context loss (Barboza & da Silva, 2016; McCarthy, 2016). For Generation Z, raised in an instant-reaction world (Gabrielova & Buchko, 2021), immediate feedback is crucial, as they dislike delays (Eckleberry-Hunt et al., 2018).

Teacher feedback on students' reflections may not be timely due to the laborious process of reading individual responses and tracking his/her goals and plans. For example, in Butzler's (2016) study, students completed exam wrappers, and the teacher provided individual student feedback on SRL strategies. This feedback was not immediate due to the time needed to read each student's answers. Scaling up such feedback for larger student numbers is challenging. The second drawback is the lack of personalized feedback, as seen in van Alten et al. (2020a, 2020b), where hints were general to all students (e.g., make notes, rewind video) and not tailored to each student's unique response.

This study aims to address the drawbacks by exploring whether a Large Language Model (LLM)-based chatbot can detect and monitor students' goal setting and planning in online learning. LLMs are deep learning models generating outputs based on prompts (Floridi & Chiriatti, 2020), using transformer models with self-attention mechanisms (Vaswani et al., 2017) for better context understanding (Tam, 2023). OpenAI's GPT series, specifically ChatGPT, demonstrates impressive linguistic abilities and has been used for various educational-related tasks such as creating course materials (Topsakal & Topsakal, 2022), translating languages (Baidoo-Anu & Owusu, 2023), and generating answers to assessment questions (Fergus et al., 2023).

## 2. Purpose of the present study

This article is similarly concerned with using ChatGPT in education but explores the problem from a novel angle. It develops and evaluates a prototype Goal-Plan-Mentor Chatbot System (*GoalPlanMentor*) to automatically monitor students' goal setting and planning activities in online learning. Although goals and planning most often occur in the forethought phase, they can also occur in the performance phase (monitoring goal progress and plans), and post-performance phase (evaluating whether one has achieved one's goal and adjusting future plans).

In this two-stage study, we first addressed a key major LLM limitation – the inability to store new experiences in long-term memory during a dialog (Sejnowski, 2023). LLMs are "amnesics, like humans who have lost their hippocampus and are unable to remember new experiences for more than a few minutes, unable to create long-term memories" (Sejnowski, 2023, p. 327). We developed an SRL knowledge base to store the user chat records, along with relevant Memory-Augmented-Prompts that enabled *GoalPlanMentor* to retrieve students' goals and plans and provide personalized feedback. Next, we compared *GoalPlanMentor*'s detection (coding) accuracy of students' goals and plans with human coders, assessed the quality of *GoalPlanMentor*'s feedback on students' goals and plans, and student perceptions on the usefulness of the feedback. We addressed the following research questions:

**Research question 1:** Can *GoalPlanMentor* accurately detect students' goals and plans, and in particular, what is the comparative accuracy between *GoalPlanMentor* system and human evaluators?

**Research question 2:** What is the quality of the feedback provided by *GoalPlanMentor*?

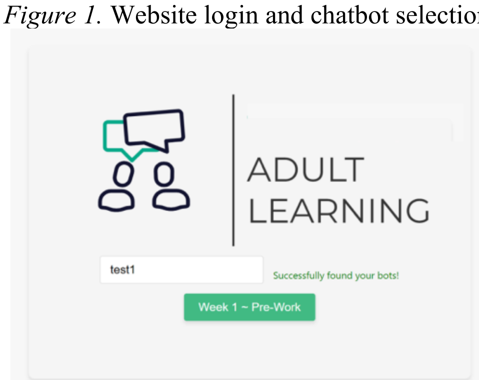
**Research question 3:** What are students' perceived usefulness of the *GoalPlanMentor*?

## 3. Designing the goal-plan-mentor chatbot system (*GoalPlanMentor*)

### 3.1. Overview of *GoalPlanMentor*

*GoalPlanMentor* is a web-based system designed to facilitate the goal setting and planning processes of students' online learning. We utilized Vue.js, a JavaScript framework, for the frontend development of the system. The frontend webpage allows students to interact with the chatbots (see Figures 1 and 2). Initially, students input their usernames before selecting a chatbot by clicking on the green button (see Figure 1) to start a conversation. Upon concluding the conversation, students can submit their current assignments to teacher or access previous submission records through the system (see Figure 2).

*Figure 1.* Website login and chatbot selection




For the backend, we chose Python and Django, a framework for swift, secure, and maintainable web development, providing a set of application programming interfaces (APIs) that the frontend uses to interact with the system (Christie et al., 2020). They enable essential operations such as user authentication, conversation management, and assignment submission. The core of the backend comprises the agent module (Figure 3), which integrates OpenAI's latest generative AI for processing and understanding user inputs, providing appropriate responses, and maintaining conversational context. We leveraged algorithms related to databases, prompt engineering, and finite state machines (FSM) to develop *GoalPlanMentor* with a dual-agent architecture. The system comprises two agents: (1) GoalPlanDetectAgent (GPDA) identifying students' goals and plans based on their chat logs, and (2) GoalPlanAwareTeachingAgent (GPATA) assisting students in reflecting on and revising their previous goals and plans throughout course instruction. By combining the GPDA and the GPATA agents, our system creates an environment that both recognizes students' goals and plans, and actively assists students in achieving them.

*Figure 2.* User interaction webpage: (1) chat message input box, (2) homework upload button, (3) query button for submitted homework

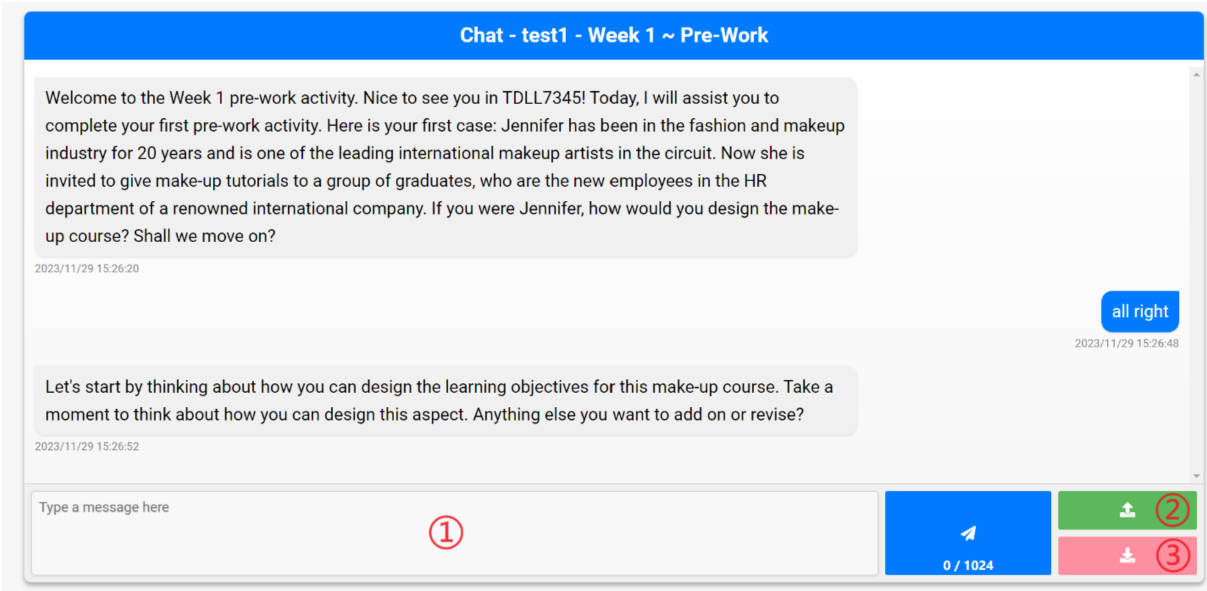
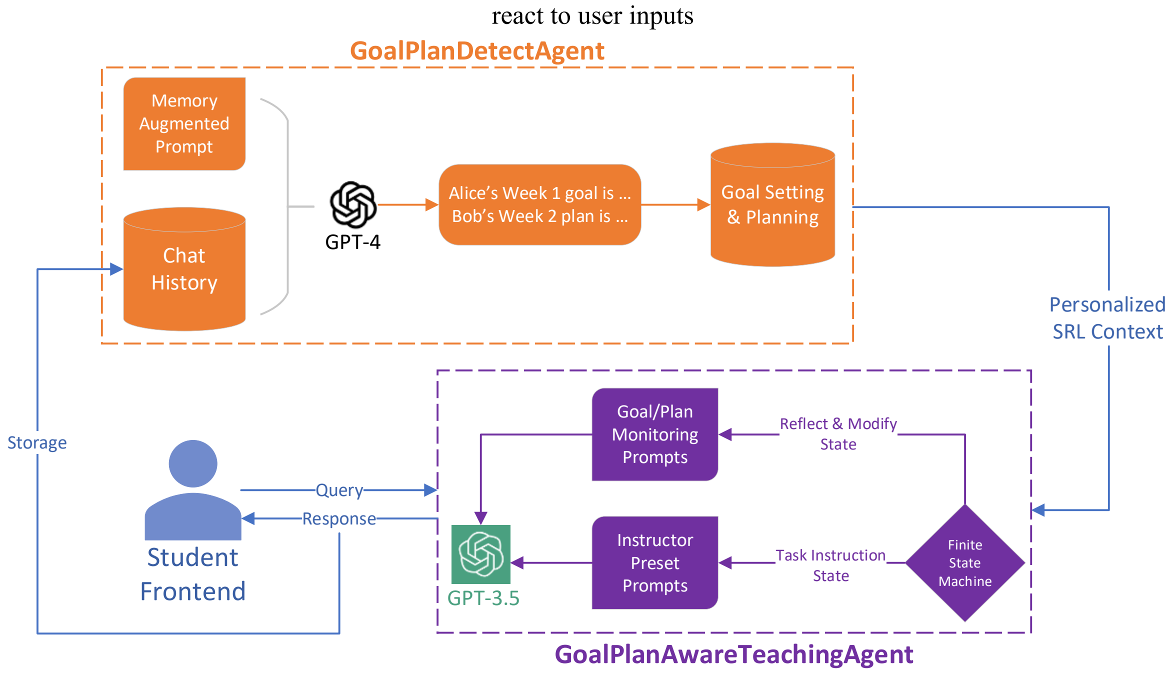



*Figure 3.* The *Agent* module in *GoalPlanMentor*, illustrating a suite of algorithms designed to comprehend and react to user inputs




### 3.2. GoalPlanDetectAgent (GPDA)

Key SRL indicators such as goal-setting and planning facilitate insights into students' learning strategies, enabling pattern recognition and progress tracking. This allows for personalized feedback that cultivates SRL behaviors and ultimately improves learning outcomes (Raković et al., 2022). Thus, we intended the GPDA agent to recognize patterns and phrases indicating a student's goals and plans from chat logs. Once identified, they are stored in a database for future reference and tracking.

#### *3.2.1 Database design*

Our database system employs SQLite, a lightweight and reliable relational database engine (Gaffney et al., 2022). We utilized Object-Relational Mapper (ORM) technology to execute operations on persistently stored objects in the relational database using Python. GPDA comprises two databases: the Chat History Database (CHDb) and the Goal setting and Planning Database (GPDb).

The CHDb stores the chat records of each student with each chatbot, as well as the progress of each conversation (e.g., welcome, ongoing, and finished). It can provide accurate context for each dialogue, aiding the chatbot in remembering and understanding past discussions, thus reducing the generation of hallucinatory responses by supporting the LLM's external memory. The GPDb is responsible for storing the weekly goals and plans set by students that are extracted from dialogues.

Relational databases simplify the storage, query, and analysis of the aforementioned personalized goal and plan data. The structured storage approach allows data attributes such as goal content and planning time frame to be independently managed, thereby enhancing data consistency and validity and improving data readability and operability.

The relational database also facilitates efficient and flexible data queries and analysis. We can obtain the goals and plans of specific students or count the number and proportion of student goals achieved over a period by writing SQL queries. Additionally, by leveraging the database's data correlation, we can gain a deeper understanding of students' learning behaviors and progress. Analyzing students' goals, plans, and their completion status can reveal learning patterns (e.g., whether students tend to set long- or short-term goals, their pace in achieving those goals, etc.). This information forms a significant reference in our design and provision of personalized learning support.

### *3.2.2 Goal and plan detection*

The GPDA's main task is to extract specific content relating to students' goal-setting and planning behavior from chat records. As these goals and plans often gradually emerge over multiple rounds of dialogue, stronger Natural Language Understanding (NLU) capabilities are needed to recognize intent and entity in context. Transformer-based LLMs such as GPT, unlike humans, inherently lack long-term memory beyond their immediate context window due to their auto-regressive nature, which predicts the next token purely based on preceding context (Vaswani et al., 2017).

To address this, we designed a Memory-Augmented Prompt (MAP) to drive GPT-4 (via the OpenAI API, model name *gpt-4-0613*) (OpenAI, 2023), which is among the most advanced current LLMs. MAP incorporates past interactions into current context, simulating the model's "memory" and overcoming its inherent inability to remember past interactions. Accordingly, the dialogues stored in the CHDb are made available for GPT-4 to query and analyze. By combining past interactions, MAP stimulates GPT-4 to perform more accurate intent and entity recognition across multiple dialogue turns.

*Table 1.* Memory-augmented prompt (MAP) for goal/plan detection

| MAP Component | Prompt Text | Description |
|---|---|---|
| Chain of Thought Nodes | You are skilled at extracting key information from conversations, and you can analyze the intent and entity of each sentence.<br>Your task is to analyze the whole conversation history and determine:<br>1. What personal learning goal has the user established?<br>2. What plan has the user formulated to meet this goal? | Role Configuration |
| | Here are the steps to identify the user's goal and plan:<br>1. Examine the assistant's discussions about the user's goals and plans.<br>2. Derive from the user's replies their specific goals and plans for the course. Prioritize the user's actual learning objectives for the course over the assistant's hypothetical situations or proposed cases.<br>3. Locate user plan: Users may use the SMART framework to make plans for achieving goals, or they might also directly say their own plans.<br>4. Addressing undefined goals and plans: In situations where the user's goals remain undefined, it is reasonable to infer that plans associated with these goals are also absent. | Dialogue History Analysis |
| | Here is a corner case to help you better understand and analyze user objectives and strategies:<br>EXAMPLE #1: | One-Shot Corner Case Demonstration |

| | | |
|---|---|---|
| | Conversation History:<br>“““<br>Assistant: What are the desired learning objectives for Jennifer’s makeup course? What specific behaviors should students demonstrate upon completion? If you were Jennifer, how would you design the make-up course?<br>User: I think I should demonstrate confidence and proficiency in makeup application across different situations.<br>“““<br>For this example, your response should be:<br>{<br>“goal”: “<null>,”<br>“plan”: “<null>,”<br>}<br>Explain: From the given conversation history, it can be observed that the user hasn’t explicitly stated their personal learning objectives or strategies. Instead, the user and the assistant are discussing objectives and strategies for a hypothetical situation, where “Jennifer” is conducting a makeup course. Therefore, the user’s personal goal and plan are not defined in this conversation. | |
| Response Formatter | Please provide your response in JSON format (can be recognized by python json.loads):<br>{<br>“goal”: “Place the identified goal here. If the user did not mention it or uncertain, insert ‘<null>‘.,”<br>“plan”: “Place the identified plan here. If the user did not mention it or uncertain, insert ‘<null>‘.,”<br>} | Render the Model Output into JSON Format |
| Memory Injection | Now, I will provide you with a history of a human user and an AI assistant talking to each other:<br>{CHAT_HISTORY} | Retrieve from CHDb |

Table 1 shows how our MAP integrates the Chain of Thought (CoT), Response Formatter, and Memory Injection components to enhance the GPT-4 model’s understanding of these “memories.” Fundamentally, CoT aims to provide LLMs with illustrative examples that clarify an underlying reasoning which the LLM should mirror within its responses, enhancing the precision of generated text (Wei et al., 2022). Our CoT method delineates the essential steps involved in extracting goals and plans, providing the model with a structured approach to processing and understanding information retrieved from CHDb.

The first CoT node is *Role Configuration*, which involves assigning the model a specific role and task. This pivotal step equips the model to better grasp the actions it is anticipated to perform.

The following node is *Dialogue History Analysis*. Here, the MAP directs the model to critically analyze previous dialogues the assistant has participated in, focusing on the user’s goals and plans. CoT then shifts the model’s attention to deriving these goals and plans. This reasoning process emphasizes the user’s actual learning objectives for a particular course, giving precedence to these over any theoretical scenarios (e.g., course design cases) suggested by the assistant. Once identified, the CoT transitions to fine-tuning and validating the extracted goal and plan; it instructs the model to determine if the user has utilized a specific framework, such as SMART, to achieve their goals, or if they have explicitly outlined their plans. This process exemplifies the reasoning involved in understanding the user’s strategies for achieving their objectives. In this way, the model gains a comprehensive understanding of the conversational context and nuances; thus, it learns a reasoning process facilitating accurate analysis and information extraction.

In the subsequent *One-Shot Corner Case Demonstration* node, the MAP addresses situations where the user’s goals are vaguely defined. The model is guided to deduce that associated plans are likely null in these instances. This design aims to strike a trade-off between providing structured guidance and preserving the model’s generative potential. Controlling the breadth and depth of examples in prompts presents a significant challenge: overly detailed examples may make the model’s generative direction overly singular, while a lack of detail can exacerbate hallucinations (Sun et al., 2023). Accordingly, we emphasize using negative examples in the MAP. This helps the model understand unclear user goals and plans and promotes its capacity to generate concise summaries of user goals and plans from multiple-round dialogues.

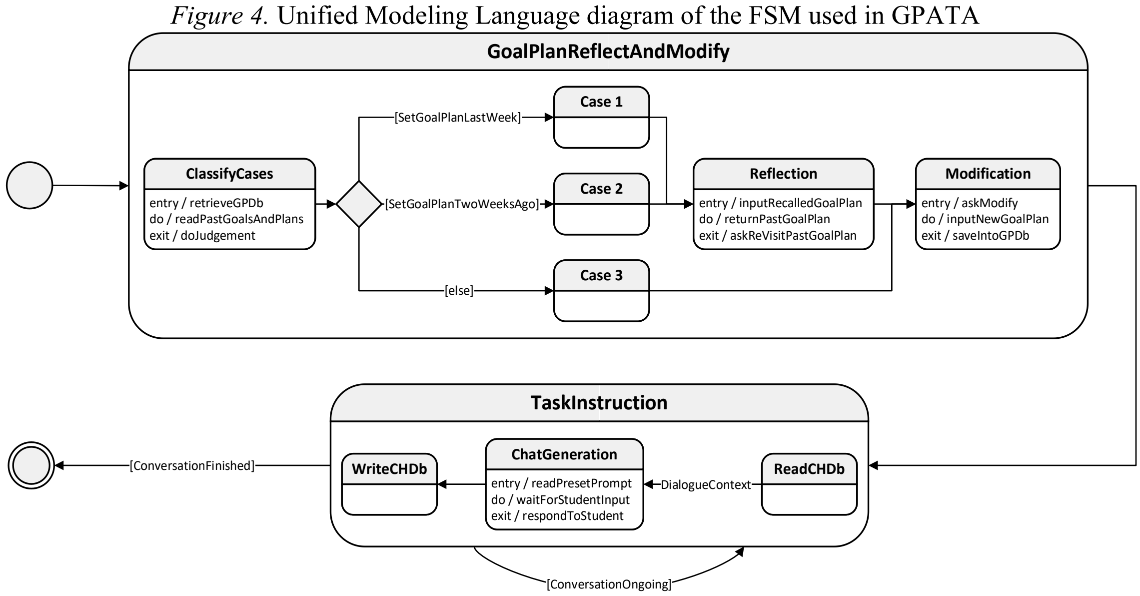
Finally, the MAP asks the model to render its response in a structured JSON format, which is compatible with Python's "json.loads" function. This enhances the programming accessibility of model output, which in turn optimizes subsequent ORM code execution and facilitates the storage of goal and plan data in the GPDb.

### 3.3. GoalPlanAwareTeachingAgent (GPATA)

The GPATA chatbot is specifically designed to interact directly with students. It leverages the capabilities of prompted GPT3.5 (via the OpenAI API, model name *gpt-3.5-turbo-16k-0613*) and an FSM to enable smooth conversation and effective dialogue-state management. Its core functionality lies in retrieving individual student-related information from the GPDA. It focuses mainly on the learning goals and plans set by the student in the weeks leading up to the current interaction and chat records within the same timeframe. Therefore, the GPATA can integrate the information of goals and plans to form a personalized conversational context for each student. This context not only reflects a history of the student's interactions with the GPATA to complete their learning tasks, but also serves as a reference for understanding the student's learning situation. With this context, the GPATA obtains an accurate awareness of each student's goal setting and planning behaviors during current chat interactions and provides personalized recommendation. We explained the overall design of the FSM and the prompts for GPT3.5 in GPATA below.

The FSM is a mathematical model used to simulate sequential logic and procedural control. It is an abstract machine that can be in one of a finite number of states, transitioning between them in response to external inputs (Lee et al., 2018). By managing conversation states clearly, the FSM ensures organized and systematic dialogue progression. It also bolsters maintainability and scalability, permitting state transitions without system disruption. Moreover, FSM is beneficial in error handling, effectively managing unexpected user inputs and steering the conversation towards a stable state.

*Figure 4.* Unified Modeling Language diagram of the FSM used in GPATA




The Unified Modeling Language (UML) diagram in Figure 4 lists the various states and transitions defined in the proposed FSM, systematically demonstrating the dialogue flow progression. Student–GPATA dialogues begin by entering the *GoalPlanReflectAndModify* state, in which GPATA retrieves the student's goals and plans established over the past two weeks from the GPDb. It then employs a rubric to categorize the student into one of three cases:

- The student established goals and plans in the previous week.
- The student only set goals and plans two weeks prior and not in the preceding week.
- The student did not set any goals or plans over the past two weeks.

Once the GPATA identify students' status, it can guide individual students to reflect and modify learning goals and plans and guiding them to complete weekly online learning tasks. For each stage, the GPATA utilize prompts to stimulate GPT3.5 to generate dialogues. Prompt engineering—strategically crafting input prompts to obtain desired model outputs (Liu et al., 2023)—is crucial when working with LLMs. Appendix A shows the prompts design in each stage of the GPATA.

### *3.3.1 Reflection stage*

For cases 1 and 2, GPATA guides the student to *reflect*. It initially prompts the student to recall and input the goals and plans they set. Regular reflection allows students to evaluate whether their current goals are realistic and attainable. Next, GPATA returns the stored "ground truth" from GPDb to the student, encouraging them to revisit their previous goals and plans. Figure 5 demonstrates how the GPATA support a student to recall and revisit her goals and plans. The dialogue then transitions into the *Modification* stage, where GPATA queries whether the student needs to modify their current goals and plans. The student can inform GPATA of newly established goals and plans (if any), which GPATA subsequently stores in the GPDb. For case 3, GPATA directly enters the *Modification* stage.

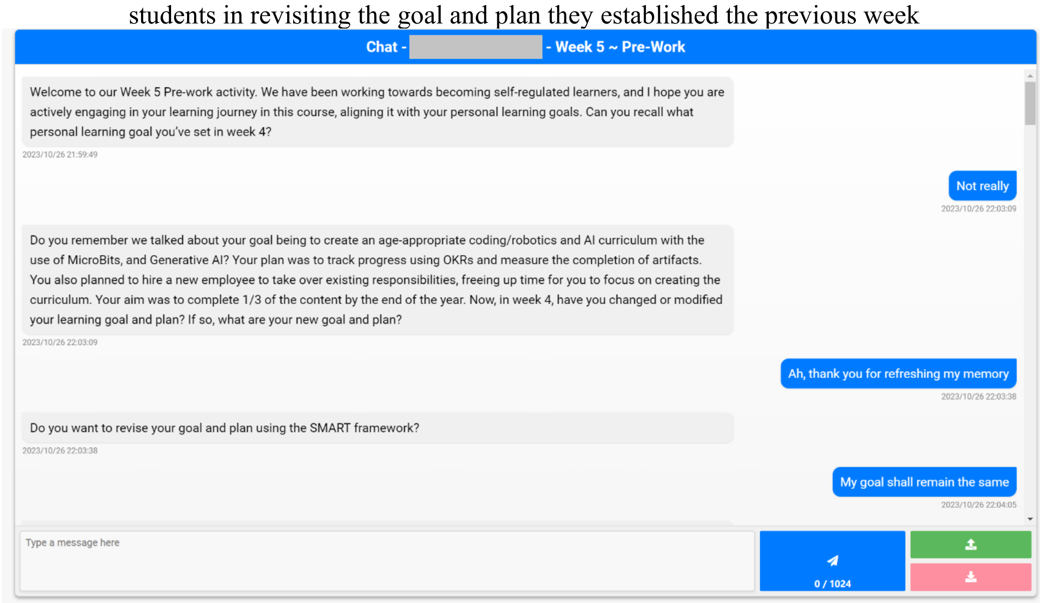
### *3.3.2. Modification stage*

The modification process is initiated by a prompt that encourages students to reconsider and potentially revise their goals and plans. The SMART framework, encompassing Specific, Measurable, Attainable, Relevant, and Time-bound elements, guided the improvement of learning objectives. The SMART framework is largely underpinned by good practices drawn from goal-setting theory – which according to Locke and Latham (2002), goals should be specific, relevant or important to the individual, achievable for the individual, and time-bound (deadlines, e.g., Locke & Latham, 1990). In addition, goals should be measurable so that the degree of accomplishment can be accurately assessed (MacLeod, 2012). Students have the flexibility to modify their current goals and plans based on their needs. Upon choosing to revise their goals, students engage in a systematic exploration of each SMART element. This modification process is designed to be personalized, with students provided suggestions on their revised goals and plans and encouraged to adapt them to their unique aspirations. The final step in this process involves summarizing the refined goals and plans.

### *3.3.3 TaskInstruction state*

The dialogue subsequently transitions into the *TaskInstruction* state, where students are asked to complete the usual weekly learning tasks (e.g., to design an adult education course). The GPATA can discuss the course design with students step by step. In each dialogue turn, GPATA initially reads from the CHDb to obtain chat context. It then generates personalized responses based on the student's input and a preset prompt. The chat records between the student and GPATA are then stored in the CHDb. This state repeats until the dialogue concludes as per the preset prompt, at which point it exits.

*Figure 5.* Screenshot of *GoalPlanMentor* interacting with a student. *GoalPlanMentor* effectively facilitates students in revisiting the goal and plan they established the previous week



## 4. Method

### 4.1. Participants and context

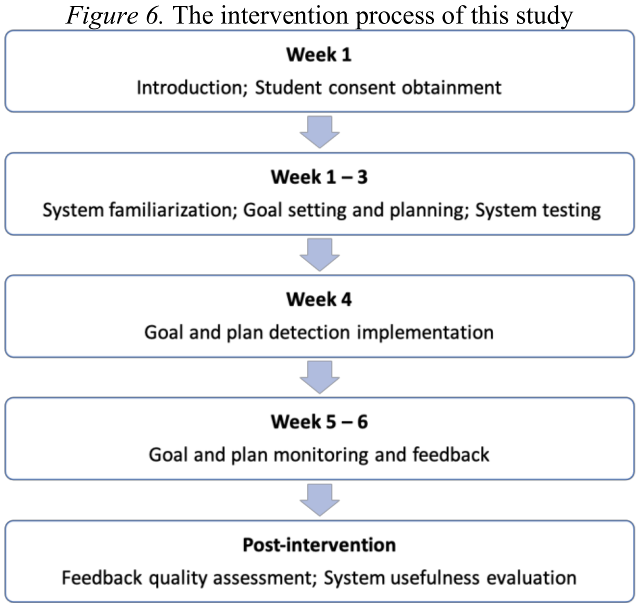
A total of 25 Asian students (20 female, 5 male) participated in an eight-week Education course. Ethical approval was obtained from the University Ethics Review Board. The course consisted of six weeks of lectures on adult learning strategies, followed by two weeks in which students demonstrated the application of these strategies in the classroom.

During the six-week lecture period, the *GoalPlanMentor* system was implemented to help students set personal learning goals and plans and monitor students' goals and plans by giving personalized feedback. *GoalPlanMentor* also guided students to complete the weekly online learning tasks. *GoalPlanMentor* was accessible via Moodle, a learning management platform where the teacher had uploaded all learning materials.

In the first week, a researcher introduced the study and *GoalPlanMentor*. From week 1 to week 3, students interacted with *GoalPlanMentor* to familiarize themselves with it. During this time, students were able to set their personal learning goals and plans with *GoalPlanMentor*. At the same time, we fixed any technical issues reported by the students and make the necessary updates. In week 4, we implemented the goal and plan detection feature (i.e., GPDA) in the backend of the system. This feature is used to recognize each student's previously set goals and plans and later provide them with appropriate feedback for their personal learning. Before releasing this feature to students in the frontend, we measured its accuracy in detecting students' goals and plans by comparing it with human coders in week 4 (RQ1). Since the detection accuracy was satisfactory, *GoalPlanMentor* started monitoring students' goals and plans by the GPTAT agent from week 5.

After the intervention, we assessed the quality of the system's feedback in terms of *coherence*, *relevance*, and *positive tone* (RQ2). Students' perceptions of the *GoalPlanMentor*'s feedback on their goals and plans were assessed through a usefulness questionnaire and follow-up semi-structured interviews (RQ3). Figure 6 illustrates the intervention process.

*Figure 6.* The intervention process of this study




### 4.2. Measurements

#### *4.2.1. Accuracy of goal and plan detection*

To answer the first research question, we coded the accuracy of the system's goal and plan detection in weeks 4-6 on a scale of five: 5 for being completely accurate, 4 for being more than half accurate, 3 for being half accurate, 2 for less than half accurate, and 1 for being not at all accurate. Appendix B presents some examples of the goal and plan detection.

### *4.2.2. Quality of feedback*

Drawing from prior studies (e.g., Motz et al., 2021; Munoz et al., 2006; Siekmann et al., 2022), we evaluated the quality of feedback provided by the chatbot system in week 5 and week 6 based on three dimensions: Coherence, Relevance, and Positive Tone (see Appendix C for examples). Our scores for these dimensions are calculated as follows:

- Coherence: Coherence refers to "a form of connecting sentences and ideas aimed at supporting readers' understanding" (Siekmann et al., p. 2). Coherent feedback relates information together so that the reader can follow the thread of ideas (Munoz et al., 2006). We give a score of 3 if the feedback is coherent, 2 if the feedback is somewhat coherent, and 1 if it is incoherent.
- Relevance: Students should be provided with relevant feedback on their work because this will motivate them to improve both the work they have already done and the work they will do in the future (Goodwin & Kirkpatrick, 2023). We give a score of 3 if the feedback relates to the students' goal setting and planning, 2 if the feedback has some digressions and irrelevancies, and 1 if the feedback is not related to the students' goals and plans.
- Positive tone: Feedback should also be affirming and uses a supportive tone. Positive feedback (e.g., praise) can motivate students to complete a task (Motz et al., 2021). We give a score of 3 if the feedback has an overall positive tone, 2 if it is neutral, and 1 if it is negative.

### *4.2.3. Students' perceived usefulness*

Participants were invited to complete a 5-point Likert-scale questionnaire (1 = strongly disagree; 5 = strongly disagree) in week 7. This survey was to assess their perceived usefulness of *GoalPlanMentor* in setting and monitoring their goals and plans. Perceived usefulness refers to the extent to which individuals believe that using a particular system would improve their job performance (Davis, 1989). The perceived usefulness scale was adapted from the original Technology Acceptance Model questionnaire (Davis, 1989, p. 340). The *perceived usefulness* questionnaire in this study consisted of 10 items, including 5 items on the perceived usefulness of the system for goal setting and 5 items on the perceived usefulness for planning. We received an 80% response rate, with 20 participants completing the questionnaire. The reliability of the questionnaire was high in this study, as shown by the Cronbach's α coefficients of 0.939 for perceived usefulness in goal setting and 0.942 for perceived usefulness in planning.

Following the intervention, semi-structured interviews were conducted to further explore students' perspectives on the usefulness of the system, and to gather suggestions for improving *GoalPlanMentor*. The interview questions included examples such as "In what ways do you find the system useful in setting and monitoring your personal goals and plans?" and "How can we improve the system to better support your goal setting and planning process?" Fifteen participants volunteered to take part in the interviews (13 online interviews and 2 face-to-face interviews at their own discretion). Each interview lasted approximately 30 minutes.

A grounded theory approach was used to analyze the interview data. To ensure the reliability of the qualitative data analysis, two independent researchers coded the interview data and identified initial themes. They then compared the initial findings and identified common themes through a consensus process. Inter-rater agreement was calculated (92%). The disagreements were resolved through discussion between the two researchers.

## 5. Results and discussion

### 5.1. Automatic detection of students' goals and plans

Cohen's kappa was computed to determine the agreement between two coders' judgments of whether the chatbot system could accurately identify the goals and plans of 25 students in Week 4.

In terms of students' goal detection, the two coders agreed that the goals of 20 students were 100% correctly identified; the goals of 3 students were half correctly detected. The goals of 2 students were partially recognized. Based on Landis and Koch's (1977) standards for strength of agreement for the kappa coefficient, there was near perfect agreement between the two coders' judgements, $\kappa = 0.84$, 95% CI [0.53, 1.15], $p < .001$.

In terms of students' plan detection, the two coders agreed that the plans of 17 students were 100% correctly identified; the plans of 2 students were almost correctly identified. The plans of 5 students were half correctly detected. For 1 student who did not discuss personal learning plans with the system, it generated plans based on the students' chat history. There was perfect agreement between the two coders' judgements, κ = 1.000, 95% CI [1, 1], $p < .001$.

The Cohen's kappa results show that the observed agreements between the two coders are statistically significant (for both goal and plan detection, $p < .001$), indicating a high level of agreement between the coders. Therefore, we averaged the system detection accuracy levels coded by the two researchers. The average goal detection accuracy was 4.68 ($SD$ = .802), while the average plan detection accuracy was 4.32 ($SD$ = 1.215).

To further verify the accuracy of the automatic detection system, we implemented it during weeks 5 and 6. As there was one missing value for each week, we examined the goals and plans of 24 students. In week 5, the average goal detection accuracy was 4.58 (SD = 1.100), while in week 6, it was 4.54 (SD = .833). Regarding plan detection, the average accuracy for week 5 was 4.21 (SD = 1.179) and for week 6, it was 4.58 (SD = 1.176). We conducted a Friedman test to investigate any significant differences in the system's goal and plan detection accuracy across weeks 4, 5, and 6 (Table 2). This test was chosen due to the violation of the normality assumption in the data for each of the three weeks. The results showed no significant differences in the system's goal detection accuracy over the three weeks, with $\chi^2(2) = 0.950$, p = 0.622. Similarly, the system's plan detection accuracy results were $\chi^2(2) = 3.636$, p = 0.162, indicating no significant differences across the three weeks. Thus, the automatic detection system demonstrated consistent accuracy in detecting goals and plans during weeks 4, 5, and 6.

*Table 2.* Friedman's ANOVA Test on the system's goals and plans detection accuracy

| Dimension | Group | *N* | Mean (*SD*) | Chi-Square | *df* | *Sig.* |
|---|---|---|---|---|---|---|
| Goal Accuracy | Week 4 | 24 | 4.68 (0.082) | .950 | 2 | .622 |
| | Week 5 | 24 | 4.58 (1.100) | | | |
| | Week 6 | 24 | 4.54 (1.179) | | | |
| Plan Accuracy | Week 4 | 24 | 4.32 (1.215) | 3.636 | 2 | .162 |
| | Week 5 | 24 | 4.21 (1.179) | | | |
| | Week 6 | 24 | 4.58 (1.176) | | | |

### 5.2. Quality of feedback of *GoalPlanMentor*

In Week 5, twenty-one students participated in the goal setting and planning activity, and the feedback provided by *GoalPlanMentor* to these students was analyzed. Both coders agreed that 20 students received coherent feedback, except for one student who received partially coherent feedback ("Apologies for any confusion earlier. Let's focus on your goal...") due to this student's brief input ("no"). The agreement between the two coders was substantial as indicated by the Cohen's kappa result, κ = 0.64, 95% CI [0.01, 1.28], $p = .002$. Additionally, there was perfect agreement between both coders that *GoalPlanMentor* provided relevant feedback with a positive tone to all 21 participants. For example, the *GoalPlanMentor* suggested a student specifying his goal: "That sounds like a fantastic goal! To make it more specific, let's clarify the target audience for your curriculum."

In Week 6, twenty-two students completed the goal-setting and planning activity, and the feedback provided by *GoalPlanMentor* to these students was analyzed. In terms of the coherence, the two coders agreed the *GoalPlanMentor* provided 18 participants with coherent feedback. Three students received partially coherent feedback, again due to their short input. Two of them received the response "Apologies... Let's revise your goal..." after responding with a single word "yes"; while the system repeated this goal-setting and planning process after one student deliberately asked the system to continue the conversation but typed "yes." There was near perfect agreement between the two coders' judgments, with, κ = 0.831, 95% CI [0.51, 1.15], $p < .001$. Additionally, there was perfect agreement between both coders that *GoalPlanMentor* was able to provide relevant feedback on all 22 participants' goals and plans with an overall positive tone. For example, after a student shared that he had tried to reach a previously set goal "I have recently written the script for the 2 of the 24 interactive video lessons," *GoalPlanMentor* responded: "That's great to hear that your goal remains the same! It sounds like you've made progress by writing...."

### 5.3. Students' perceived usefulness of GoalPlanMentor

In Table 3, we present the descriptive statistics for students' perceived usefulness of the system in goal-setting and planning. Overall, students perceived the *GoalPlanMentor* system to be moderately useful in setting and monitoring their goals (item 4) and plans (item 10). We conducted *t*-tests to determine if the results are statistically different from the midpoint. Following the practice of previous research (e.g., Beder et al., 2011; Wang et al., 2017), we defined the midpoint as a score of 2.5 (midpoint of rating 1 to 5), which indicates neither a positive nor negative finding. Results show that the overall mean perceived usefulness in goal-setting score ($M$ = 3.40, $SD$ = 0.94), and overall mean perceived usefulness in planning score ($M$ = 3.35, $SD$ = 0.93) were significantly greater than the midpoint score of 2.5, $t(19) = 4.28, p < .001$, and $t(19) = 4.09, p < .001$ respectively. This suggests that participants found the system to be useful in both goal-setting and planning aspects.

*Table 3.* Mean and standard deviation for students' perceived usefulness of the system

| | Item | *M* (*SD*) |
|---|---|---|
| Perceived usefulness in goal-setting ($n$ = 20) | 1. Using the system enabled me to reflect on my personal goal-setting. | 3.15 (.81) |
| | 2. Using the system made it easier to comprehend my personal goal-setting. | 3.35 (.88) |
| | 3. My comprehension of the personal goal-setting would be easy to obtain with the system. | 3.20 (.83) |
| | 4. The system enhanced my effectiveness in preparing my personal goal-setting. | 3.30 (.87) |
| | 5. Overall, I found the system was useful in my personal goal-setting. | 3.40 (.94) |
| Perceived usefulness in planning ($n$ = 20) | 6. Using the system enabled me to reflect on my personal plan. | 3.50 (1.00) |
| | 7. Using the system made it easier to comprehend my personal plan. | 3.30 (.92) |
| | 8. My comprehension of the personal plan would be easy to obtain with the system. | 3.35 (.93) |
| | 9. The system enhanced my effectiveness in preparing my personal plan. | 3.30 (.80) |
| | 10. Overall, I found the system was useful in my personal plan. | 3.35 (.93) |

We examined the differences in students' learning achievement on their final assignment between the low- and high-groups in terms of their perceived usefulness. Students were categorized into either a high-perceived usefulness group ($n$ = 5, top 25% of students) or a low-perceived usefulness group ($n$ = 5, bottom 25% of students) based on the ranking of their responses on perceived usefulness in the goal setting and planning scales. The course instructor, with eight years of experience, assessed students' final assignments on designing an adult workshop toolkit, which included a document outlining the workshop's topic, goals, activities and their rationale. Both the instructor and a coder individually rated the assignments and then reconciled any differences in their evaluations.

#### *5.3.1. Difference in learning performance between the low- and high-perceived usefulness groups in terms of goal-setting*

The final scores of the high-perceived group ($M$ = 51.0, $SD$ = 5.15) and the low-perceived group ($M$ = 31.8, $SD$ = 14.26) were normally distributed for both groups (Shapiro-Wilk's test, $p \geq .05$). The variances of the final scores were homogeneous for both groups (Levene's test, $p$ = .119). An independent-samples t-test (Table 4) showed that students who highly perceived *GoalPlanMentor's* usefulness for goal-setting exhibited significantly greater learning achievements compared to those with a low perception of its usefulness, $t(8) = 2.833$, $p$ = 0.022.

#### *5.3.2. Difference in learning achievement between the low- and high- perceived usefulness groups in terms of planning*

The final scores of the high-perceived group ($M$ = 52.0, $SD$ = 5.52) and the low-perceived group ($M$ = 40.6, $SD$ = 14.35) were normally distributed for both groups (Shapiro-Wilk's test, $p \geq .05$). The assumption of homogeneity of variances was violated (Levene's test, $p$ = .048). We, therefore, used the statistics for the t-test under the condition of equal variances not assumed. An independent-samples t-test showed that there was no statistically significant difference in learning performance between the two groups in terms of perceived usefulness of the system for planning, $t(5.16) = 1.658$, $p$ = .156. Overall, the results suggest that higher perceived usefulness of *GoalPlanMentor* for goal setting could significantly improve learning achievement. One possible reason is that *GoalPlanMentor's* emphasis on goal setting may have a more substantial impact on learning achievement, as it helps students establish clear objectives and stay focused on their desired achievements.

*Table 4.* Independent-samples *t*-tests on students' learning performance between the low- and higher-achievers

| Perceived usefulness | Group | *n* | *Mean (SD)* | *p*-value |
|---|---|---|---|---|
| Goal setting | High | 5 | 51 (5.15) | .022 |
| | Low | 5 | 31.8 (14.26) | |
| Planning | High | 5 | 52 (5.52) | .156 |
| | Low | 5 | 40.6 (14.35) | |

We analyzed students' interview data to better understand their perceptions of *GoalPlanMentor's* usefulness and gather suggestions for future improvements. Three key themes emerged from the interview data highlighting the usefulness of *GoalPlanMentor*: clear goals and plans clarification, ongoing learning reflection, and personalized communication.

### *5.3.3. Clear goals and plans clarification*

Setting clear goals and plans is crucial, especially at the beginning of a course. Students ($n$ = 8) reported *GoalPlanMentor* clarified their goals and plans through the SMART framework and provided diverse ideas and suggestions. Students elaborated on the initial simplicity of their goals and how they gradually refined them with the guidance of *GoalPlanMentor*. For example:

At the beginning of our conversation, I stated my goals in a very simple way. However, as it [*GoalPlanMentor*] prompted me to become more specific using the SMART framework, my learning goals became more concrete at the end. (Participant 8)

Providing diverse ideas and suggestions on students' goals "broaden my perspectives and expand my thinking during the goal-setting and planning process" (Participant 12). This highlights students' active engagement with *GoalPlanMentor* for varied perspectives in goal setting. Clear goals and plans at the start of the course encourages student ownership and set the tone for a self-regulated learning experience.

### *5.3.4. Ongoing learning reflection*

In weeks 5 and 6, *GoalPlanMentor* displayed students' previous goals and plans, "providing a time-saving advantage" (Participant 14), enabling students to recall and reflect on their personal goals and plans without the need for extensive memory retrieval. These students ($n$ = 12) highlighted the importance of the weekly reminders in maintaining their learning focus. For example:

In the midst of absorbing new knowledge and completing various learning activities throughout this course, I sometimes forget my learning objectives. The weekly reminders not only help me recall my goals but also allow me to reflect on my learning progress and assess whether I am on track. (Participant 6)

Students also appreciated the flexibility offered by *GoalPlanMentor* to modify their goals and plans as their understanding of the course content deepened in later sessions:

My initial goals were unrelated to the course since I was not familiar with the learning content at the beginning of the semester. As my knowledge of the course material deepened, I tended to revise some of my initial goals. I would ask the system to repeat what we discussed last week. Based on that, I would make modifications. (Participant 4)

The ability to revise goals and plans helped students reflect on their own learning and align their objectives with the course material, ensuring that their goals remain relevant and mirror their developing understanding.

### *5.3.5. Personalized communication*

Some students ($n$ = 3) reported that *GoalPlanMentor* helped them customize learning goals and plans to their individual needs and professional contexts. The lack of human judgment allowed students to freely explore goal-setting questions related to their unique backgrounds. One student share:

It [*GoalPlanMentor*] allows me to make my goals more related to my personal development. I work in a startup and my background is different from the majority of my classmates who are in-service teachers. Sometimes I'm hesitant to ask questions in class because my questions may not be relevant to them. So having this [*GoalPlanMentor*] allows me to set my personal learning goals because I know that it is really just for me, and I don't need to worry about if my goals relate to others. (Participant 7)

Students suggested improvements for the system's utility in goal and plan setting. Currently, the system generates feedback based solely on students' previous messages/utterances. Students ($n$ = 13) stressed the need for the system to provide more specific feedback that is tied to the course content. For example:

It would be helpful if the system could give me feedback like, 'For the first learning goal, you need to pay more attention on the learning contents in week 4.' Perhaps in this way, it can help me think more deeply about the goals I set. (Participant 10)

Other students ($n$ = 6) stressed the importance of teacher involvement in providing feedback and adjusting the course design based on students' goals and plans. This fosters a sense of connection and support, while lack of teacher feedback may decrease students' interest in goal setting and online learning activities:

I also wonder if the goals we set are noticed by the teacher, and if the teacher would adjust the course content based on our goals. In this course, there was no such process. Personally, I might lose enthusiasm for this goal-setting activity as time goes on. (Participant 14)

## 6. Conclusion

Given the importance of self-regulation, this study developed and evaluated a LLM chatbot system (*GoalPlanMentor*) to automatically detect and monitor students' goal setting and planning. The results suggest *GoalPlanMentor* to be a feasible solution to provide timely and personalized feedback regarding students' goals and plans. Results suggest substantial to near perfect agreement between *GoalPlanMentor* and human coding, with clear feedback directions. Students found GoalPlanMentor useful for goal setting and planning, with higher perceptions of usefulness in terms of goal setting related to significantly greater learning achievements. Consequently, this study presents new insights into circumventing the inability of current LLMs to create long-term memories by developing a SRL knowledge base and using appropriate Memory-Augmented Prompt to simulate *GoalPlanMentor*'s "memory" and overcoming its inherent inability to remember past interactions.

Although the results appear promising, it is important to acknowledge certain limitations that may affect their generalizability and applicability. One limitation is the small and homogeneous sample size used in our study. The participants were selected from a specific demographic group, which may not accurately represent the broader population. Consequently, the results may not be readily applicable to other demographic groups or settings with different characteristics. Future research should aim to replicate the study with larger and more diverse sample sizes to enhance the generalizability of the findings. Second, our system employs a combination of technologies, including Vue.js. SQLite, Python, Django, and ChatGPT-4. While Vue.js. SQLite, Python and Django are open source and freely available, ChatGPT-4 is a proprietary language model. The dependency on proprietary software may pose challenges in terms of accessibility, adaptability, and cost. The use of ChatGPT-4 can limit the accessibility of our system to other researchers or practitioners who may not have access to this proprietary technology. Depending on proprietary software can also reduce the system's adaptability, making it more difficult to integrate with other tools or platforms that are not supported by OpenAI. Additionally, access to proprietary software often involves subscription fees or usage-based costs, which can be expensive over time. As usage scales, costs can increase significantly, making it potentially prohibitive for large-scale applications without substantial financial resources. Third, the current study focuses on students' goal setting and planning strategies in their forethought phase of SRL. There is a need to include a broader range of SRL indicators, such as help-seeking and monitoring during the performance and self-reflection phases to develop a more comprehensive SRL recommendation system. Finally, the generative AI-based SRL recommendation system in this study may generate fabricated output in terms of learners' goals and plans. Such phenomenon is known as artificial hallucination (Varghese & Chapiro, 2024). In our future research, we plan to use Retrieval-Augmented Generation which integrates external sources of verifiable facts into the chatbot generative processes, thereby improving response relevance and accuracy.

## Acknowledgment

All authors declare that they have no conflicts of interest. This work was supported by the Research Grants Council of Hong Kong Research Fellow Scheme (Reference no: RFS2223-7H02). Generative AI tool (ChatGPT 4) was used to improve the readability of some writings in this manuscript. After using this tool, the authors reviewed and edited the specific contents and take full responsibility for the accuracy of the contents.

## References

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of Generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62. https://doi.org/10.61969/jai.1337500

Barboza, E.J.S., & da Silva, M.T. (2016). The importance of timely feedback to interactivity in online education. In I. Nääs, O. Vendrametto, J. M. Reis, R. F. Goncalves, M. T. Silva, G. von Cieminski & D. Kiritsis (Eds.), *Advances in Production Management Systems. Initiatives for a Sustainable World* (pp. 307-314). Springer. https://doi.org/10.1007/978-3-319-51133-7_37

Beder, J., Coe, R., & Sommer, D. (2011). Women and men who have served in Afghanistan/Iraq: Coming home. *Social Work in Health Care*, *50*(7), 515-526. https://doi.org/10.1080/00981389.2011.554279

Butzler, K. B. (2016). The synergistic effects of self-regulation tools and the flipped classroom. *Computers in the Schools*, *33*(1), 11–23. https://doi.org/10.1080/07380569.2016.1137179

Cheng, Z., Zhang, Z., Xu, Q. Maeda, Y., & Gu, P. (2023). A meta-analysis addressing the relationship between self-regulated learning strategies and academic performance in online higher education. *Journal of Computing in Higher Education*, 1-30. https://doi.org/10.1007/s12528-023-09390-1

Christie, M., Marru, S., Abeysinghe, E., Upeksha, D., Pamidighantam, S., Adithela, S. P., Mathulla, E., Bisht, A., Rastogi, S., & Pierce, M. (2020). An extensible Django-based web portal for Apache Airavata. In *Practice and Experience in Advanced Research Computing* (pp. 160-167). Association for Computing Machinery. https://doi.org/10.1145/3311790.3396650

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340. https://doi.org/10.2307/249008

Dos Santos, L. M. (2022). Online learning after the COVID-19 pandemic: Learners' motivations. *Frontiers in Education*, *7*, 879091. https://doi.org/10.3389/feduc.2022.879091

Eckleberry-Hunt, J., Lick, D., & Hunt, R. (2018). Is medical education ready for Generation Z? *Journal of Graduate Medical Education*, *10*(4), 378–381. https://doi.org/10.4300/JGME-D-18-00466.1

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education*, *100*(4), 1672-1675. https://doi.org/10.1021/acs.jchemed.3c00087

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1

Gabrielova, K., & Buchko, A. A. (2021). Here comes Generation Z: Millennials as managers. *Business Horizons*, *64*(4), 489-499. https://doi.org/10.1016/j.bushor.2021.02.013

Gaffney, K. P., Prammer, M., Brasfield, L., Hipp, D. R., Kennedy, D., & Patel, J. M. (2022). SQLite: Past, present, and future. *Proceedings of the VLDB Endowment*, *15*(12), 3535-3547. https://doi.org/10.14778/3554821.3554842

Goodwin, R., & Kirkpatrick, R. (2023). Using rubrics to improve writing skills: A study in Kuwait. *Language Testing in Asia*, 13, 17. https://doi.org/10.1186/s40468-023-00224-6

Handoko, E., Gronseth, S. L., McNeil, S. G., Bonk, C. J., & Robin, B. R. (2019). Goal setting and MOOC completion: A study on the role of self-regulated learning in student performance in massive open online courses. *The International Review of Research in Open and Distributed Learning*, *20*(3). https://doi. org/10.19173/irrodl.v20i4.4270

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159 – 174. https://doi.org/10.2307/2529310

Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal setting. *Organizational Behavior and Human Decision Processes*, *50*(2), 212-247. https://doi.org/10.1016/0749-5978(91)90021-K

Lee, K., Lee, Y. S., & Nam, Y. (2018). A model of FSM-based planner and dialogue supporting system for emergency call services. *The Journal of supercomputing*, *74*(9), 4603-4612. https://doi.org/10.1007/s11227-018-2432-4

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, *55*(9), 1-35. https://doi.org/10.1145/3560815

Locke, E.A., & Latham, G.P. (1990). *A theory of goal setting and task performance*. Prentice-Hall Inc.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

MacLeod, L. (2012). Making SMART goals smarter. *Physician Executive, 38*(2), 68-72.

McCarthy, J. (2016). *Timely feedback: Now or never*. Edutopia. George Lucas Educational Foundation. https://www.edutopia.org/blog/timely-feedback-now-or-never-john-mccarthy

Moos, D. C., & Bonde, C. (2016). Flipping the classroom: Embedding self-regulated learning prompts in videos. *Technology, Knowledge and Learning*, *21*(2), 225-242. https://doi.org/10.1007/s10758-015-9269-1

Motz, B., Canning, E., Green, D., Mallon, M., & Quick, J. (2021). The influence of automated praise on behavior and performance. *Technology, Mind, and Behavior*, *2*(3), 1–12. https://doi.org/10.1037/tmb0000042

Munoz, A. P., Mueller, J., Álvarez, M. E., & Gaviria, S. (2006). Developing a coherent system for the assessment of writing abilities: Tasks and tools. *Íkala, revista de lenguaje y cultura*, *11*(1), 265-307. https://doi.org/10.17533/udea.ikala.2791

Nussbaumer, A., Albert, D., & Kirschenmann, U. (2011). Technology-mediated support for self-regulated learning in open responsive learning environments. In 2*011 IEEE global engineering education conference (EDUCON)* (pp. 421–427). IEEE. https://doi.org/10.1109/EDUCON.2011.5773171

OpenAI. (2023). *GPT-4 Technical Report*. https://cdn.openai.com/papers/gpt-4.pdf

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8. https://doi.org/10.3389/ fpsyg.2017.00422

Pedrotti, M., & Nistor, N. (2019). How students fail to self-regulate their online learning experience. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou & J. Schneider (Eds.), *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning* (pp. 377-385). Springer. https://doi.org/10.1007/978-3-030-29736-7_28.

Pérez-Sanagustín, M., Sapunar-Opazo, D., Pérez-Álvarez, R., Hilliger, I., Bey, A., Maldonado-Mahauad, J., & Baier, J. (2021). A MOOC-based flipped experience: Scaffolding SRL strategies improves learners' time management and engagement. *Computer Applications in Engineering Education*, *29*(4), 750-768. https://doi.org/10.1002/cae.22337

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Academic Press. https://doi.org/10.1016/B978-012109890-2/50043-3

Rasheed, R. A., Kamsin, A., & Abdullah, N. A. (2020). Challenges in the online component of blended learning: A systematic review. *Computers & Education*, *144*, 103701. https://doi.org/10.1016/j.compedu.2019.103701

Raković, M., Bernacki, M. L., Greene, J. A., Plumley, R. D., Hogan, K. A., Gates, K. M., & Panter, A. T. (2022). Examining the critical role of evaluation and adaptation in self-regulated learning. *Contemporary educational psychology*, *68*, 102027. https://doi.org/10.1016/j.cedpsych.2021.102027

Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural computation*, *35*(3), 309-342. https://doi.org/10.1162/neco_a_01563

Shih, K. P., Chen, H. C., Chang, C. Y., & Kao, T. C. (2010). The development and implementation of scaffolding-based self-regulated learning system for e/m-learning. *Educational Technology & Society*, *13*(1), 80–93. https://www.j-ets.net/collection/published-issues/13_1

Siekmann, L., Parr, J. M., & Busse, V. (2022). Structure and coherence as challenges in composition: A study of assessing less proficient EFL writers' text quality. *Assessing Writing*, *54*, 100672. https://doi.org/10.1016/j.asw.2022.100672

Singh, J., Steele, K., & Singh, L. (2021). Combining the best of online and face-to-face learning: hybrid and blended learning approach for COVID-19, post vaccine, & post-pandemic world. *Journal of Educational Technology Systems*, *50*(2), 140–171. https://doi.org/10.1177/00472395211047865

Sun, B., Li, Y., Mi, F., Bie, F., Li, Y., & Li, K. (2023). Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Vol. 2, Short Papers, pp. 1741-1750). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.148

Tam, A. (2023). *What are Large Language Models*. Machine Learning Mastery. https://machinelearningmastery.com/what-are-large-language-models/

Topsakal, O., & Topsakal, E. (2022). Framework for a foreign language teaching software for children utilizing AR, Voicebots and ChatGPT (Large Language Models). *The Journal of Cognitive Systems*, *7*(2), 33–38. https://doi.org/10.52876/jcs.1227392

Usher, E. L., & Schunk, D. H. (2017). Social cognitive theoretical perspective of self-regulation. In D. H. Schunk, & J. A. Greene. (Eds.), *Handbook of Self-Regulation of Learning and Performance* (pp. 19-35). Routledge.

van Alten, D. C. D., Phielix, C., Janssen, J., Kester L. (2020a). Effects of self-regulated learning prompts in a flipped history classroom. *Computers in Human Behavior*, *108*, 106318. https://doi.org/10.1016/j.chb.2020.106318

van Alten, D. C. D., Phielix, C., Janssen, J., Kester L. (2020b). Self-regulated learning support in flipped learning videos enhances learning outcomes. *Computers & Education*, 158, 104000. https://doi.org/10.1016/j.compedu.2020.104000

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Varghese, J., & Chapiro, J. (2024). ChatGPT: The transformative influence of generative AI on science and healthcare. *Journal of hepatology*, *80*(6), 977-980. https://doi.org/10.1016/j.jhep.2023.07.028

Wang, X. Q., Petrini, M. A., & Morisky, D. E. (2017). Predictors of quality of life among Chinese people with schizophrenia. *Nursing & health sciences*, *19*(2), 142-148. https://doi.org/10.1111/nhs.12286

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://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

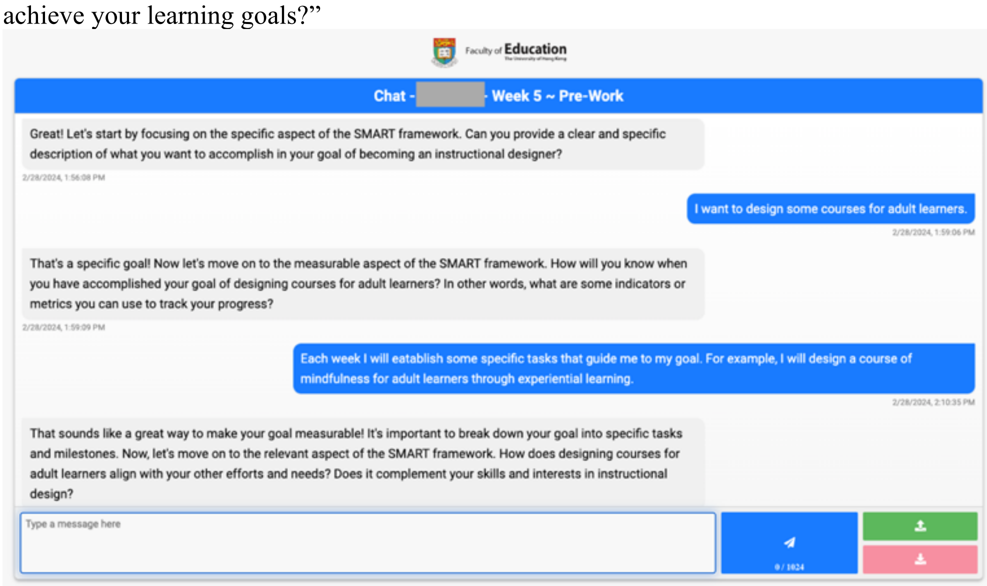
Wolters, C. A., & Brady, A. C. (2021). College students' time management: a self-regulated learning perspective. *Educational Psychology Review*, 33, 1319-1351. https://doi.org/10.1007/s10648-020-09519-z

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*, 64–70. https://doi.org/10.1207/s15430421tip4102_2

Zimmerman, B. J., & Paulsen, A. S. (1995). Self-monitoring during collegiate studying: An invaluable tool for academic self-regulation. *New Directions for Teaching and Learning, 63*, 13-27. https://doi.org/10.1002/tl.37219956305

# Appendix A. Prompts design in GPATA

| Stage | Prompt text |
|---|---|
| Reflection | Taking Case 1 as an example, we prompted the GPT3.5 as follow:<br><br>Your first prompt is: We've been putting in effort to become self-regulated learners, and I really hope you're actively embracing your learning journey in this course, aligning it with your own personal goals for growth and improvement. Can you recall what personal learning goal you've set in week 4?<br><br>You should refine the following sentence using the expression of the second person:<br>I remember we talked about your goal and plan in week 4: "<week 4 goal>,<week 4 plan>." Have you changed/modified your learning goal and plan? If so, what are your goal and plan now? Here you should wait for my answer. |
| Modification | We just finished recalling my own personal learning goal of the course in week 4.<br><br>Your first prompt should be: "Do you want to revise your goal and plan using the SMART framework?." You should wait for my response.<br><br>If I don't want to revise my goal and plan, you should skip the first section, and dive into the second section directly.<br><br>If I want to revise, you should guide me to clarify and plan my new goals based on 5 aspects ("SMART," "Specific," "Measurable," "Attainable," "Relevant," "Time-based") in SMART in order.<br><br>Here are some example questions for SMART Framework you can use in your response to my goal:<br>1. Specific: Define a clear, specific goal (e.g., What do I want to accomplish?)<br>2. Measurable: make sure you can track progress (e.g., How will I know when it is accomplished?)<br>3. Relevant: A relevant goal can answer "yes" to these questions (e.g., Does this match my other efforts/needs?)<br>4. Attainable: (e.g., How can I accomplish this goal?)<br>5. Time-bound/ Timely (reference data for completion): (e.g., When do you expect to achieve your learning goals?" |



| | |
|---|---|
| | #Rules for setting personal learning goals based on SMART framework:<br>1. If I ask for your suggestion, you should provide suggestions and ask me to revise or adapt your suggestion to my own goal.<br>2. Keep your response under 100 words.<br>3. You should give me a concise summary of my personal learning goals. After giving me a summary, you should ask me: Did I summarize your goals and plans accurately? If not, feel free to edit.<br><br>After discussing all 5 aspects (i.e., specific, measurable, relevant, attainable, Time-bound/Timely) of the SMART framework, you can go to the second section below. |
| TaskInstruction | Now, you will act as a guide to help me complete a course design activity. you should say: "Your task is to design a lesson to teach adults on the topic 'Writing an effective resume'. I will guide you step by step. Shall we start?." You should wait for my answer.<br><br>In this section, you should guide me to design the "resume course" for my adult learners from five components: learning objectives, course structure, course content, teaching methods, and assessment. If I skip any of the components, you should remind me, and ask me to complete these missed components. |

## Appendix B. Examples of the accuracy of goal and plan detection

| Accuracy scale | Definition | Student response | System detection result |
|---|---|---|---|
| [1] 0% accuracy | The system generates fabricated goals or plans based on the students' chat history. | Nil (Student did not explicitly mention any plan) | The user plans to focus on specific areas within instructional design such as the proper way of information presentation and facilitation skill (generate by system). |
| [2] < 50% accuracy | The system correctly detected a minor portion of goals or plans based on the students' chat history. | **I plan to use feedback as a way of goal measurement; I plan to achieve my goal via creating an engaging and student-centred learning environment** ... by the end of this semester. | The student plans to achieve these goals by the end of the semester. |
| [3] 50% accuracy | The system half correctly detected goals or plans based on the students' chat history. | I would like to **organize adult learning workers to engage them** and create my own workshop. | The student aims to create own workshop relevant to job duties. |
| [4] > 50% accuracy | The system almost correctly detected goals or plans based on the students' chat history. | **I can also receive feedback to evaluate my progress.** Lecture ppt and related reference should be reviewed and apply to my planned activities. I think that half year for the literature review and framework development is needed and half year for the execution. To learn the related theories, framework, or cases of adult learning so that I can try to apply to my situation. I can apply what I learned in the planning of staff development. | Learn related theories, frameworks, or cases of adult learning and apply them to the user's situation. The user plans to review lecture PowerPoint presentations and related references, and apply this knowledge to planned staff training activities. The user expects to spend half a year on literature review and framework development, and another half year on the execution of planned activities. |
| [5] 100% accuracy | The system 100% correctly detected goals or plans based on the students' chat history. | I want to learn how to engage adult learner.<br>I want to learn skills to encourage learner to join my class. | Learn skills to encourage learners to join the student's class. |

*Note.* The bolded responses were not automatically detected by the system.

## Appendix C. The quality of feedback coding scheme

| Dimension | Scale | Definition | Examples of system response |
|---|---|---|---|
| Coherence | 1 | The feedback is incoherent. e.g., lacking a clear logical flow. e.g., making confusing or vague statements. | [Not found] |
| | 2 | The feedback is somewhat coherent. e.g., does not fully grasp the user's intent but response with clear logical flow. | I apologize for any confusion caused. Let's refocus on your goal of designing courses specifically for adult learners. When considering the relevancy of your goal, it's essential to ensure that the courses you design meet the unique needs and interests of adult learners. |
| | 3 | The feedback is coherent. e.g., demonstrate clear logic and well-structure statement, e.g., without any ambiguity or confusion. | That's a great way to measure the effectiveness of your course! By conducting pre-course and post-course surveys to gauge participants' perception of their stress levels and their knowledge and understanding of stress reduction strategies, you can track their progress. A positive shift in their self-reported stress levels and increased knowledge would indicate that your goal has been accomplished. |
| Relevant | 1 | The feedback is not related to the students' goals and plans. | [Not found] |
| | 2 | The feedback has some digressions and irrelevancies in terms of studetns' goals and plans. | [Not found] |
| | 3 | The feedback relates to the students' goal setting and planning. | I understand that you may feel pressed for time, but revising your goal and plan using the SMART framework can greatly benefit your learning journey. It will help ensure that your goals are well-defined and achievable. Taking the time to go through each aspect of the framework can provide you with a clear direction and roadmap for your online program for adult learners. Are you open to giving it a try? |
| Positive Tone | 1 | The feedback lacks supportive and encouraging elements and may highlight criticism. | [Not found] |
| | 2 | The feedback is neutral, without expressions of support or criticism. | [Not found] |
| | 3 | The feedback is positive, providing supportive and encouraging words. | It's great to hear that you have enough time to prepare for the training and that you can communicate with other on-the-job teachers to understand the problems students may face in different schools. This will definitely help you in tailoring your online training to address those specific challenges. |