An Automated Personalized Feedback System with Large Language Models to Facilitate Students' Engagement in Classroom Learning
PhD Research Proposal

Jiang Mi
jiangmi@cqupt.edu.cn

1. Proposal Introduction

Personalized feedback is praised as one of the most influencing pedagogical tools for students learning quality and assessment [36, 46]. In online and offline education, it is often provided by expert human teachers as a a vital bridge connecting teachers and students, fostering a dynamic and interactive learning environment and enhancing student engagement[30]. Personalized feedback can take many forms, it could be an informal oral feedback, a simple comment on student's assignment or a detailed written report[5]. Although personalized feedback is widely recognized in educational institutions, a variety of uncertainties like class sizes and student population still need to be challenged[19]. Moreover, the problem of how to provide effective and interactive feedback to meet student's individual need is hotly debated. It is also unrealistic for a single teacher to teach a course and give personalized feedback to each student simultaneously. As a result of that, students are likely to get distracted from class and lose the confidence in learning. To solve this problem, a variety of new technologies are applied to assist teaching-learning process by building the Automated Feedback Systems (AFS)[31, 13]. With the assistance of new technologies, teachers are given a high degree of freedom to teach and accomplish other pedagogical goals.

One of the most prominent technologies is Large Language Models (LLMS), recent years, LLMs have achieved a marked advancement in the realm of AI innovation and it also brings potential to further innovations in education. One key application is ChatGPT[52], a chatbot capable of producing human-like comprehensive and systematic information. On one hand, it can be a teaching assistant, helping teacher with teaching materials and assessing student's performance. On the other hand, it has the potential to serve as a virtual tutor, answering student's questions and supporting academic written assignments[22]. Several AI power educational systems are using Generative Pre-trained Transformer (GPT-3) to assess students behavior or train students curiosity to asking questions[2, 29]. However, the aforementioned approaches for provision of feedback address on all the students, it ignores the hidden requirements of the unmotivated students who are not self-disciplined and have few ambition for learning. Therefore, it is essential to delve into practical and effective methods to strengthen students with weak learning status in class.

A large cohort of researches investigated approaches of recognizing students learning status by the myriad ways, including analyzing classroom recordings and providing teachers with feedback on their use of talks move[45] and generating feedback in respond to the submitted programming assignments[34]. Besides, These researches [15, 44, 4] studied students' learning status by recognizing students' engagement and emotion, which thus is considered to be an important predictor of student achievement. Based on the previous researches, this proposal focuses on providing the multifaceted visual analysis on student's learning status(the aspect of engagement) and fostering student's learning and engagement by automatically generating personalized feedback to student end. This proposal mainly explores 4 research questions (RQ1-4).

- **RQ1** How does a student behave in the classroom to indicate that they are distracted and need tutoring feedback?
- **RQ2** From what dimensions of visual analytics on students learning status are helpful?

- **RQ3** To what extent is the feedback generated by Whisper and ChatGPT accurately relating to student's disengagement moment and fixing student's learning gaps in classroom?

- **RQ4** How can we leverage the benefits of Large Language models(LLMs) to generate feedback by means of student preferred content and structure and effectively guide their learning?

RQ1 is invoked by the ambition to know the feasibility of regarding a student's disengagement as a sign for that he/she needs academic assistance and tutor feedback. To answer RQ1, a comprehensive questionnaire will be completed by 105 volunteering students(including 40 females and 65 males). All the volunteering students will be hired from the School of Software Engineering, School of the Art and School of Telecommunications at Chongqing University of Posts and Telecommunications(CQUPT).

RQ2 aims to investigate students' actual needs and preferences on displaying engagement status from a multi-dimensional visual analytics. To answer RQ2, also a comprehensive questionnaire will be conveyed by the same group of volunteering students of RQ1. Based on it, round table discussions will be held iteratively with 9 students (4 major from computer science and 5 from art designing) and 5 teacher representatives. During these discussions, the mocks of the system and visual analytics prototypes will be designed and figured out to meet students expectations.

RQ3 is motivated by the urgent to investigate the feasibility and quality of proposed feedback system in the light of LLMs. To answer RQ3, disengaged student will be recognized. Since student's engagement status is not stable within the lecture period, a student's disengaged video segment will be used as an input to Whisper and ChatGPT.

Furthermore, to better assist and serve students learning, the effectiveness and usability of generated feedback is taken into consideration, which bring the last question RQ4. To explore RQ4, ChatGPT's responses will be analyzed with regard to their characteristics (e.g., feedback content and feedback presenting structure). Inspired by the the work of this research[51], 25 students selected from the volunteering group and 10 teachers from different academic background are required to grade the feedback generated from ChatGPT and also grade the feedback given by teacher. The grades are measured using 5-point scale. After grading, the mean grade of automated feedback from ChatGPT and the mean grade of feedback from teacher will be compared to evaluate RQ3 and RQ4.

2. Literature Review

The aim of this study is to establish an approach to automatically provide personalized feedback based on students learning status. To realize the final goal, the literature review can be categorized into 3 sections, namely, Measurement of college student engagement in classroom, Teacher's speech-to-text approaches, Existing systems to show student's learning status and give personalized feedback.

2.1 Measurement of college student engagement in classroom

Firstly, a student's engagement status is a critical indicator to justify a student's learning quality. A large quantity of scholars have researched on this topic. In this proposal, the related literature reviews are collected from nearly 20 references by selecting key words of "College student engagement detection", "How to measure college student engagement" and "College student learning status". All the contribution factors are concluded in Table 1. According to these previous researches[11, 10, 7], a student engagement can be measured by 4 categories of factors: Emotion, Cognition, Physical behaviors and Body posture. More recent scholars proposed a variety of automated approaches on student engagement recognition. This research [42] calculated the level of engagement by measuring the movements of the eyes and head, and facial emotion. Meanwhile, the work of Aravind[35] considered features like eye gaze, facial expression, and body posture as significant factors for engagement assessing. Apart of this, Chen Wang and Pablo Cesar used galvanic skin response(GSR) sensors to measure college student engagement in e-learning environment[50]. Similarly, Altuwairqi and Jarraya1 [3] recognized keyboard key strokes and mouse movements for online learning engagement. In this proposal, a student's head pose, eye gaze and facial emotion are taken into considerations and the approach from the work of [42] will be adopted.

2.2 Teacher's speech-to-text approaches

Secondly, teacher's speech is an essential input for generating feedback. Through the process of speech-to-text recognition, the teaching content can be recognized and recorded. Over the past

Table 1: Summary of student engagement measurements

Factors	References
Emotion	surprise, anger, fear, happiness, disgust, sadness[35, 3, 16, 26, 54, 47]
Cognition	beliefs, theories, and concepts[12, 43, 37]
PhysicalBehaviors	Head Pose[26, 17, 47], Eye Gaze[35, 26, 17, 47], Hand gesture[48, 14]
Bodyposture	[35, 37]

decades, tremendous amount of researches have been done on speech-to-text recognition. The timeline can be classified into 3 period, the early stage, the midterm stage and the modern stage. During the early stage, researches mainly focused on using rule based method, like discrete Hidden Markov Model(HMM) for speech recognition[38, 21]. During the midterm stage, approaches based on Deep Learning were widely used. This research[28] adopted a single recurrent neural network (RNN) as the acoustic model to predict context-based output. Moreover, scholars in these research papers [1, 33] applied convolutional neural networks (CNNs) as the framework for automatic speech recognition and [1] proved that (CNNs) are more efficient than deep neural network (DNN) since it reduces the error rate. Recent years, Large language models (LLMs) have proven themselves the efficiency to solve a wide range of generative tasks with high performance. In this research [9], LLMs were applied to an automated speech recognition (ASR) system, with the realization of prepending a sequence of audio embeddings to the text token embeddings. Moreover, Paul K and Chulayuth proposed[40] AudioPaLM, a unified multimodal architecture which was announced to have the best performance of speech translation. Meanwhile, a variety of speech-to-text approaches based on LLMs were proposed, such as Speech-LLaMA and SpeechAgents[24, 55]. Recent years, OpenAI released Whisper[39] , which is designed to understand and transcribe spoken language with the aid of multiple languages and multitask. In this proposal, considering the cost and device feasibility, Whisper will be applied in teacher's speech-to-text recognition.

2.3 Existing systems to show student's status and give personalized feedback

Owing to the rapid expansion of and progress in Internet technology and Internet use, a large quantity of intelligent tutoring systems were developed. Teaching platforms like Duifene and Rain classroom provide teaching management services, like taking attendance, assigning homework and taking online tests. Apart from this, there are a large quantity of learning systems that can provide visual analytics of student's learning status. In this work [26], student's real-time learning status was demonstrated from 3 different dimensions. Haipeng et al. [54] designed a visual analytics system for student's emotions in classroom videos. Besides, Many scholars studied how to measure and visualize student's learning status[25, 49, 17]. However all the mentioned researches and systems only focused on analyzing student's learning status, they didn't generate response to the students witnessed of weak learning status. Beyond this, many researches studied intelligent system of generating personalized feedback. The work of [23] proposed a machine learning approach to generate feedback with personalized hints, Wikipedia-based explanations, and mathematical hints. [53] Wenzhong et al. proposed an AI-assisted personalized feedback system aiming to enhance student's learning ability. Based on the large amount of researches reviewed, there is a research gap of combining student learning status visualization with automated feedback system. In this study, we will establish a student diagnostic system, providing a visual analytics of student's learning status and generating efficient personalized feedback to those student who performs undesirably.

3. Statement of Significance

The topic of generating personalized feedback to enhance student's learning gain is hotly investigated over the past decades. A vast quantity of researches study the topic with a variety of approaches, including manual approaches, semi-automated approaches and automated approaches. However, there still remains myriad tough tasks to explore and tackle with.These years, with the fast development of Large Language Models (LLMs), it is realising its potential for applications in areas such as teaching agent, scoring system and automated personalized feedback. Beyond this, the Chinese government has recognised digital education as a critical breakthrough in educational development. World-leading researches are supported in the field of personalized learning and lifelong learning, expanding the coverage of quality educational resources and modernizing

education. Therefore, building upon previous contributions, this study moves the field forward by leveraging the benefit of LLMs to investigate an approach of generating timely and effective feedback to students in need. There are 3 main contributions of the research:

- This research proposes an interactive diagnostic system to support visual analytics of student's learning status and generate personalized feedback for students reviewing.
- This research investigates a model to intentionally generate feedback in the light of recognizing student's learning status(mainly on student's disengagement).
- Questionnaire-based survey and interviews with students, teachers and domain experts are conducted to further investigate the system's quality and students' perspectives.

The proposed program will contribute to meet the grand challenges in education, especially in enhancing student's learning quality and gain.

4. Methods

In this section, details of the approach for generating personalized feedback are presented.

4.1 System Overview

As figure 1 depicts, this system is consisted of 3 parts, Disengagement recognition, Speech-to-text recognition and personalized feedback generating. Since this study seeks to address generating personalized feedback to facilitate disengaged student learning, it is essential to recognize each disengaged student at the first stage. Thereafter, a 45-minute video will be cut to 4 video clips(3 last 10 minutes and 1 lasts 15 minutes). During each video clips, if any disengagement appeared, the video clip will be labeled as a reference clip, which will be sent to Whisper jointed with ChatGPT, ultimately, the lecture note based feedback for student will be summarized and recorded.

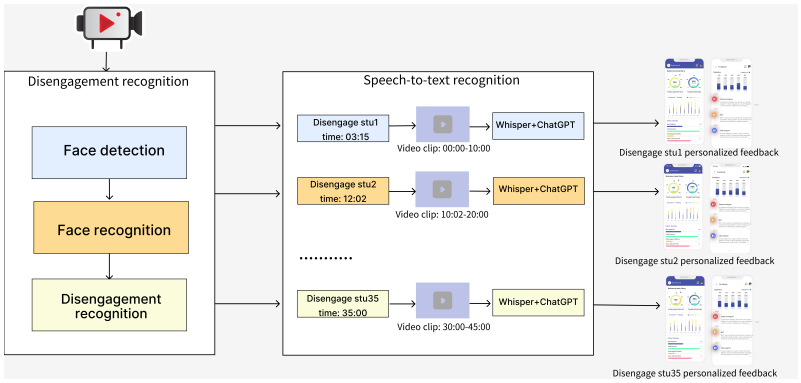

Figure 1: System Overview

4.2 Data collection

The video data is collected from CQUPT(Chongqing University of Posts and Telecommunications). The video will be captured from a group of 35 volunteering students(10 girls and 25 boys) from the course of "Software Analysing and Designing in UML2". 32 videos of an academic semester are collected with each 45 minutes. The video is captured from a high-fidelity camera in front of the classroom. As it is mentioned before, each 45-minuted video will be segmented into 4 video clips, there are 128 video clips totally. Based on the study of [32, 47, 54], the video clips will be recorded at 30 frames per second with a resolution of 1920×1080. In order to extract the redundant frames the video clips will be sampled at the rate of $1/10$, like 3 frames a second.

4.3 Disengagement Recognition

4.3.1 Face Detection

As Multi-task Cascaded Convolutional Networks (MTCNN)[56] harnessed the potential of providing rapid face detection and face alignment, we will adopt MTCNN to detect face in this study. The architecture of MTCNN is consisted of three stages of carefully designed deep convolutional networks. After the process of three stages, the network will output five facial landmarks' positions.

4.3.2 Face recognition

After detecting faces in each frame, it is essential to recognize the face of each student. The very effective way is to compare with the huge image database to find out the most similar one. We will adopted FaceNet in this study. FaceNet[41] is a face recognition method based on deep convolutional networks along with triplet loss. It provides a unique architecture for performing tasks like face recognition, verification and clustering. In our study, we will adopt the approach proposed in the research paper[20], FaceNet trains its output to be a vector of 128 embedding on the basis of learning a Euclidean embedding per image with a deep convolutional network. It is trained that the squared L2 distance in the embedding space is directly correlated with face similarity.

4.3.3 Students' Disengagement Recognition

This proposal measures student engagement by the level of his/her concentration in the classroom. We assume student would turn his/her eye and head to the direction that they are attending and concentrating-the direction of teacher or PPT in front of the classroom. Thus eye gaze and head pose are very essential indicators to identify a student's status in classroom: focused or distracted. On the basis of recognizing a focused student, the facial emotion can be further added to consolidate the level of student engagement. In this way, a student's eye gaze, head pose and facial emotion are considered as the assessing criterion to classify a student engagement status. Learning from the works of [27, 42, 16], the level of student engagement will be classified into 3 categories, engaged, medium engaged and disengaged.

- Head Pose Detection

Unlike online environment, the large offline classroom environment has limitations in lighting conditions, face pose and the face occlusion. It is necessary to eliminate all non-frontal faces from the detected faces first. The removing faces includes left-skewed, right-skewed, upward and downward. According to the work of [32], a MTCNN model will be utilized for generating facial landmarks in this proposal. Based on this method, three pose-related degrees, yaw, pitch and roll will be calculated. Yaw presents the rotation of an object in horizontal movement, Pitch presents the rotation of an object in vertical movement and roll presents the rotation of an object in circular movement. In this way, left and right-skewed face will be removed by a threshold of the degree of yaw movement. Up and down face will be removed by a threshold of the degree of pitch movement as well.

- Eye Gaze Detection

Eye gaze can be selected as a real-time index of the information-processing priorities of the visual system[18], therefore it is feasible to deduct a student engagement by analyzing his or her eye gaze direction. To accurately recognize a student's eye gaze from multi-faces classroom video, a coarse to fine network will be adopted[6]. Since eye gaze has strong correlations with face directions. Firstly, a common CNN will be used to extract the coarse grained face feature, and another two CNNs will be trained to extract the fine grained eyes(left and right) feature. The estimated eye gaze directions will be used to classify student's learning status into two categories: "Distracted" or "Focused".

- Emotion Recognition

It is essential to leverage the information conveyed from various emotions to further predict a student's learning status. In this study, the emotion detection model is based on[42]. The facial emotion analysis will take place only when the student is "focused". The CNN model will be used to classify the facial emotions into categories of "Neutral", "Happy", "Surprised", "Sad", "Bored", "Anger" and "Scared".

- Students engagement estimation

In this stage, levels of student engagement will be measured by the value of Concentration Index(CI), which is provided in the paper [42]. There are two factors contribute to the value of CI, they are Eye Gaze Weight (EGW) score and Emotion Weight (EW). Each of the two weight will be determined through a quiz with 10 questions respectively to demonstrate the correlation between the two kinds of weights and student concentration level. In this proposal, the detail of Emotion Weight will be described and the detail of Eye Gaze Weight will be described in my later research paper. Each recognized emotions will be labeled as several weights between 0-1 accordingly. To get the weights corresponding to each emotion, a 20-minute lecture video will be shown to 35 volunteering students followed by a quiz with 10 questions. Students will be grouped on the basis of their dominant emotion, that is to say, if a student is recognized a neutral emotion more than 50% of the time video duration, he/she will be included in the group of neutral. After completing the quiz, The mean score of each emotion group will be calculated as the value of EW. In this proposal, Table2 presents the value of EW calculated from the research paper[42], we will update the weights through later investigation. The Concentration Index can be represented as the equation below:

$$CI = EGW \times EW \tag{1}$$

Table 2: Emotions Weight

Emotion	Emotion Weight
Neutral	0.9
Happy	0.6
Surprised	0.6
Sad	0.3
Bored	0.2
Anger	0.25
Scared	0.3

At last, the students engagement will be categorized into 3 levels:
Engaged: the value of CI is between 50% to 100%.
Medium engaged: the value of CI is below 50%.
Disengaged: the value of CI is 0 and the student is recognized as distracted.

4.4 Personalized Feedback Generating

4.4.1 Speech-to-text Recognition

We assume students of medium engaged and disengaged to be weak learning status who need tutoring and personalized feedback from teacher. As it is mentioned in chapter 4.1, a 45-minute video is segmented into 4 video clips with time period of, 0:00-10:00, 10:00-20:00, 20:00-30:00 and 30:00-45:00. If any of the weak learning status is recognized, the associated video clip will be extracted and labeled as the speech signal for feedback provision.

4.4.2 Personalized Feedback Generation

As Whisper established by OpenAI is an open source speech-to-text recognition system, it supports multiple languages, including Chinese, Spanish and German, etc.. In this study, Whisper is chosen for translating teacher's audio signal into text. At first, each 10-minute labeled video clip is transformed into MP3 format through FFmpeg. As Whisper is limited to transfer file of 25MB, it is possible that the 10-minute mp3 need segmentation through the package of AudioSegment from PyDub. Since Whisper provides 5 models of various sizes, considering the capacity of the video card in computer, model "base" is loaded. After the process of speech-to-text, ChatGPT is imported to summarize the text. The prompt for ChatGPT is designed as follows, "Please give feedback on the following text in terms of concise description of the goals of the project, clear description of state-of-the-art technique for completing the project, novelty/creativity applications of the project and overall clarity of the report". The summarized text is then presented at the student end(smart phone) as the personalized feedback. A targeted effort at importing Whisper and ChatGPT could possibly result in enhancing the readability and accuracy of feedback to each student.

4.5 Visualization design

This study investigates an approach to provide student with adaptive and responsive learning experience by automatically sending back learning status and personalized feedback. Therefore, the requirements of visualization design are easy search and navigation, user friendly interaction, prompt feedback generation. Given these 3 requirements, the visualization analytics are designed into 2 view scale, dashboard view and feedback view.

- Dashboard View

This view aims to provide the overview of student learning status from multi-scales. Illustrated in Figure2(a), the dashboard view is consisted of 4 sub-views, namely, Today Engagement, Today Vocabulary, Engagement Report and Today Emotion. Today Engagement view provides a straightforward data to show a student's overall engagement performance with is supported by the mean value of Concentration Index(CI) within a 45-minutes video. Meanwhile, the view of Today Vocabulary is designed to give hints to student for a quick navigating and reviewing from the high frequency academic words in the course. The Engagement Report view in the middle of the User Interface (UI) illustrates levels of students' engagement over time. The x-axis represents time out of 45 minutes and the y-axis shows the value of student's engagement-CI, which is calculated in Chapter 4.3. There is a temporal audio wave right under the red engagement line, in the future design, the audio wave will be switched to Mel-spectrogram. At the bottom of the UI, the view of Today Emotion displays a student's emotion distribution by presenting 4 categories of emotion, Neural, Happy, Surprised and Sad. These four emotions are selected from Table2 with top 4 highest Emotion Weight.

- Feedback View

Feedback view addresses to generate content-based feedback on the basis of student weak learning status. As can be seen from Figure 2(b), the statistics view on top of the screen demonstrates the moment when student is disengaged. Each disengagement time is represented by a blue cylinder showing the percentage value of engagement. As it is mentioned in chapter 4.4, if any disengagements are recognized during each of the 4 periods, the associated video clips will be sent to Whisper for personalized feedback generation. The feedback are classified with the time hint and are listed in a chronological order. Each feedback is guided by a key word placing right above the set of feedback accordingly.

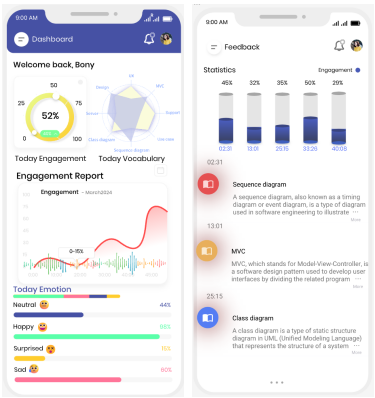

Figure 2: (a)Prototype of Dashboard View (b) Prototype of Feedback View

5. Evaluation plan

The work will be evaluated through User study and Interviews.

5.1 User Study

To evaluate the usability and quality of the system, we will conduct a user study. To begin with, the system will be compared with two baseline system, Duifene and HIKVISION Smart

Classroom. 1) Duifene is a blended mode teaching management tool which can provide both online and offline service for teachers and students. However, Duifene is only able to meet student's basic requirements by providing functionality of online test, taking attendance, real-time chatting and assigning homework. It is not interactive enough to take actions in response to student's different learning status. 2) HIKVISION Smart Classroom is a heavy-weight multi-task platform that can provide multiple pre-class, in-class and after-class services, like student's real-time learning status recognition, teaching management and exam supervision. However, its functions are too complex and wide, there is an urgent need for both students and teachers to have a light-weight system that only focuses on students engagement analyzing and personalized feedback provision.

- Participants

To accomplish this survey, as it is mentioned in chapter1, we will invite 105 students with the average age of 20, 35 from School of Software Engineering, 35 from School of the Art and 35 from School of Telecommunications. By getting their permission to appear in the video for analyzing. They will be separated into three groups to take three different courses according to their majors. Moreover, we will recruit 3 (2 females, 1 male) teachers from aforementioned schools, who are lecturer, associate professor and TA to teach the courses. After taking the courses, all the students will be required to fill in the questionnaires.

- Task and Procedures

Before going through the procedures, students will be announced with the following tasks: 1) To observe the overall learning status from dashboard view. 2) To observe the feedback view. 3) To check and verify the accuracy of the feedback. Meanwhile, 3 teachers will be announced to teach a course according to the academic teaching schedule respectively. After students and teachers know all the tasks, 3 groups of students will be invited to participate in the related courses respectively. Each course will be taken in a classroom of which limited capacity is 70 students. In this case the influence of occlusion will be decreased. In the classroom, Students are asked to seat and listen to the teacher as they normal do.

- Questionnaire

The questionnaire is designed to investigate student's disengagement behaviour and to collect student's real perspectives on the usability of ChatGPT generated feedback. Aiming to answer Research Questions of RQ1 and RQ2 with a comprehensive aspects, the questionnaire is presented by 4 dimensions, disengagement behaviour, Usability, Visual design and Suggestions. The first draft of the questionnaire for students is listed in Table 3:

Table 3: Student Questionnaire

Q1	You head down/up/right/left when you are disengaged in classroom.
Q2	Your eye gaze down/up/right/left when you are disengaged in classroom.
Q3	Please list other behaviors when you are disengaged in classroom.
Q4	It is easy(difficult) to use.
Q5	It is easy(difficult) to learn.
Q6	It is easy(difficult) to know the engagement status.
Q7	The feedback is(isn't) helpful.
Q8	The feedback is(isn't) readable.
Q9	The feedback do(don't) contain effective content to guide student learning.
Q10	It is easy(difficult) to find out the information based on the overall layout and format.
Q11	The Today Vocabulary visualization is(isn't) helpful.
Q12	The Engagement Report visualization is(isn't) helpful.
Q13	The Today Emotion is(isn't) helpful.
Q14	I will(will not) use the system for future study.
Q15	I will(will not) recommend the system to my friends.
Q16	Please give us some suggestions about the system.

5.2 Interview

To answer RQ3 and RQ4, we will invite 25 students from the volunteering group and 10 teachers with different academic backgrounds at the campus to assess and grade the feedback generated from Whisper and ChatGPT. To better evaluate the quality of the feedback, the same group will be invited to grade the feedback written by another 5 teachers from different school at the campus. Inspired by the work of [8], the grading is at 5-point scale, where: (i) 0 denotes Incomprehensible; (ii) 1 denotes not fluent and incoherent; (iii) 2 Somewhat fluent but incoherent; (iv) 3 Fluent but somewhat incoherent and (v) 4 Fluent and coherent. The mean grade of ChatGPT feedback and the mean grade of teacher feedback will be compared to evaluate the efficiency and usability of the proposed system. Moreover, the aforementioned 10 grading teachers and 25 students representatives and also 3 domain experts will be invited to join several interviews. Prior to the interviews, we will develop a set of interview questions, which focus on asking interviewees to describe their real perspectives of using the system, including their motivations, expectations, attitudes about the quality of the system and the personalized feedback generated from ChatGPT. Interviews are semi-structured and each lasts about 30 minutes. The result of the grading comparison and the interview will provide the hint and solution to RQ3 and RQ4.

6. Research timeline and plan

To further investigate this research and refine the research proposal to a final paper, one year plan is taken into considerations. The one year process is composed of 5 milestones, Research design and planning, Literature review, Data collection, Data evaluation and Writing up. Figure3 depicts the draft of the research timeline and plan within the first year of PhD study.

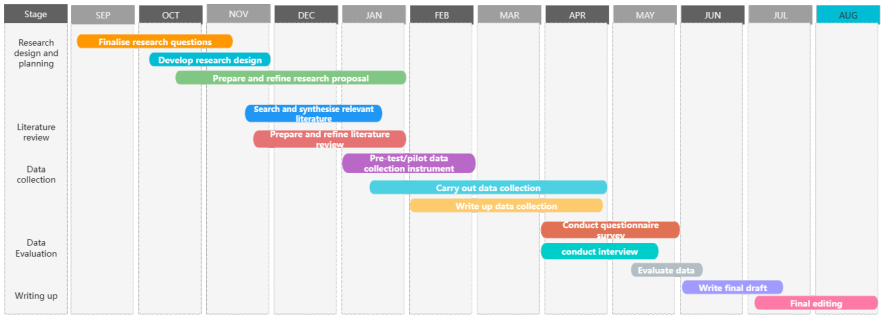

Figure 3: Research timeline and plan of a year

References

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, et al. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

[2] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, et al. GPT-3-driven pedagogical agents for training children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, pages 1–36, 2023.

[3] Khawlah Altuwairqi, Salma Kammoun Jarraya1, Arwa Allinjawi, and Mohamed Hammami. Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image and Video Processing*, 15:1387–1395, 2021.

[4] Jonathan Bidwell and Henry Fuchs. Classroom analytics: Measuring student engagement with automated gaze tracking. *Bhav Res Methods*, 49(113), 2011.

[5] Alasdair Blair, Steven Curtis, Mark Goodwin, and Sam Shields. What feedback do students want? *Politics*, 33(1):66–79, 2013.

[6] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10623–10630, 2020.

[7] Lucy Chipchase, Megan Davidson, Felicity Blackstock, Ros Bye, et al. Conceptualising and measuring student disengagement in higher education: A synthesis of the literature. *International Journal of Higher Education*, 6(2):31–42, 2017.

[8] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, et al. Can large language models provide feedback to students? a case study on chatgpt. *IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325, 2023.

[9] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, et al. Prompting Large Language Models with Speech Recognition Abilities. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 13351–13355, 2024.

[10] Jennifer A. Fredricks. Engagement in school and out-of-school contexts: A multidimensional view of engagement. *Theory into practice*, 50(4):327–335, 2011.

[11] Jennifer A. Fredricks, Phyllis C. Blumenfeld, and Alison H. Paris. School engagement: Potential of the concept, state of the evidence. *Review of Educational Researc*, 74(1):59–109, 2004.

[12] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, et al. Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33:27–49, 2021.

[13] Arthur C Graesser, Mark W Conley, and Andrew Olney. Intelligent tutoring systems. *American Psychological Association*, 2012.

[14] Joseph F Grafsgaard, Joseph B Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, et al. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 42–49, 2014.

[15] Mitchell M Handelsman, William L Briggs, Nora Sullivan, and Annette Towler. A measure of college student course engagement. *The Journal of Educational Research*, 98(3):184–192, 2005.

[16] Mohammad Nehal Hasninea, Huyen TT Bui, Thuy Thi Thu Tran, Ho Tran Nguyen, et al. Students' emotion extraction and visualization for engagement detection in online learning. *Procedia Computer Science*, 192:3423–3431, 2021.

[17] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, et al. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29:821–867, 2019.

[18] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, et al. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29:821–867, 2019.

[19] Hamideh Iraj, Anthea Fudge, Huda Khan, Margaret Faulkner, et al. Narrowing the feedback gap: Examining student engagement with personalized and actionable feedback messages. *Journal of Learning Analytics*, 8(3):101–116, 2021.

[20] Rongrong Jin, Hao Li, Jing Pan, Wenxi Ma, and Jingyu Lin. Face recognition based on mtcnn and facenet. *GitHub repository*, 2021.

[21] Biing Hwang Juang and Laurence R Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 33(3):251–272, 1991.

[22] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 2023.

[23] Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, et al. Automated personalized feedback improves learning gains in an intellegent tutoring system. *Artificial Intelligence in Education: 21st International Conference, AIED 2020*, pages 140–146, 2020.

[24] Egor Lakomkin, Chunyang Wu, Yassir Fathullah, Ozlem Kalinli, Michael L. Seltzer, and Christian Fuegen. End-to-End Speech Recognition Contextualization with Large Language Models. *arXiv e-prints*, page arXiv:2309.10917, September 2023.

[25] Yuanyuan Liu, Jingying Chen, Mulan Zhang, and Chuan Rao. Student engagement study based on multi-cue detection and recognition in an intelligent learning environment. *Multimed Tools Appl*, 77:28749–28775, 2018.

[26] Shuai Ma, Taichang Zhou, and Xiaojuan Ma. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. *Proceeding of the 2022 CHI conference on human factors in computing systems*, pages 1–25, 2022.

[27] Bouhlal Meriem, Habib Benlahmar, Mohamed Amine Naji, and Petros others. Determine the level of concentration of students in real time from their facial expressions. *International Journal of Advanced Computer Science and Applications*, 13(1), 2022.

[28] Yajie Miao, Mohammad Gowayyed, and Florian Metze. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding. *arXiv e-prints*, page arXiv:1507.08240, July 2015.

[29] Steven Moore, Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper. Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3. *European conference on technology enhanced learning*, pages 243–257, 2022.

[30] Susanne Narciss, Sergey Sosnovsky, Lenka Schnaubert, Eric Andrès, et al. Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers Education*, 71:56–76, 2014.

[31] Hyacinth S. Nwana. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277, 1990.

[32] Chakradhar Pabba and Praveen Kumar. An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, 39(1), 2022.

[33] Dimitri Palaz and Ronan Collobert. Analysis of cnn-based speech recognition system using raw speech as input. *Idiap*, 2015.

[34] Maciej Pankiewicz and Ryan S. Baker. Large Language Models (GPT) for automating feedback on programming assignments. *arXiv e-prints*, page arXiv:2307.00150, June 2023.

[35] Aravind Sasidharan Pillai. Student Engagement Detection in Classrooms through Computer Vision and Deep Learning: A Novel Approach Using YOLOv4. *Sage Science Review of Educational Technology*, 5(1):87–97, 2022.

[36] Dolors Planar and Soledad Moya. The Effectiveness of Instructor Personalized and Formative Feedback Provided by Instructor in an Online Setting: Some Unresolved Issues. *The Electronic Journal of e-Learning*, 14(3), 2016.

[37] Athanasios Psaltis, Konstantinos C Apostolakis, Kosmas Dimitropoulos, and Petros Daras. Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Games*, 10(3):292–303, 2017.

[38] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[39] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, et al. Robust Speech Recognition via Large-Scale Weak Supervision. *International conference on machine learning*, pages 28492–28518, 2023.

[40] Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, et al. AudioPaLM: A Large Language Model That Can Speak and Listen. *arXiv e-prints*, page arXiv:2306.12925, June 2023.

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *Proceedings of IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[42] Prabin sharma, Shubham Joshi, Subash Gautam, Sneha Maharjan, et al. Student Engagement Detection Using Emotion Analysis, Eye Tracking and Head Movement with Machine Learning. *International Conference on Technology and Innovation in Learning, Teaching and Education*, pages 52–68, 2022.

[43] Gale M Sinatra, Benjamin C Heddy, and Doug Lombardi. The challenges of defining and measuring student engagement in science. *Educational Psychologist*, 50(1):1–13, 2015.

[44] Ajitha Sukumaran and Arun Manoharan. A survey on automatic engagement recognition methods: online and traditional classroom. *Indonesian Journal of Electrical Engineering and Computer Science*, 30(2):1178–1191, 2023.

[45] Abhijit Suresh, Jennifer Jacobs, Vivian Lai, Chenhao Tan, Wayne Ward, James H. Martin, and Tamara Sumner. Using Transformers to Provide Teachers with Personalized Feedback on their Classroom Discourse: The TalkMoves Application. *arXiv e-prints*, page arXiv:2105.07949, April 2021.

[46] Leonard Tetzlaff, Florian Schmiedek, and Garvin Brod. Developing Personalized Education: A Dynamic Framework. *Educational Psychology Review*, 33:863–882, 2021.

[47] Chinchu Thomas and Dinesh Babu Jayagopi. Predicting student engagement in classrooms using f acial behavioral cues. *In Proceedings of 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, pages 33–40, 2017.

[48] Ghassem Tofighi, Haisong Gu, and Kaamraan Raahemifar. Vision-based engagement detection in virtual reality. *Digital Media Industry Academic Forum (DMIAF)*, pages 202–206, 2016.

[49] Zouheir Trabelsi, Fady Alnajjar, Medha Mohan Ambali Parambil, et al. Real-Time Attention Monitoring System for Classroom: A Deep Learning Approach for Student's Behavior Recognition. *Big Data and Cognitive Computing*, 7(1):48, 2023.

[50] Chen Wang and Pablo Cesar. Physiological Measurement on Students' Engagement in a Distributed Learning Environment. *In Proceedings of the 2nd International Conference on Physiological Computing Systems*, pages 149–156, 2022.

[51] Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. Automated Evaluation of Personalized Text Generation using Large Language Models. *arXiv e-prints*, page arXiv:2310.11593, October 2023.

[52] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, et al. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal OF Automatica Sinica*, 10(5):1122–1136, 2023.

[53] Wenzhong Xu, Jun Meng, S Kanaga Suba Raja, M Padma Priya, and M Kiruthiga Devi. Artificial intelligence in constructing personalized and accurate feedback systems for students. *International Journal of Modeling, Simulation, and Scientific Computing*, 14(01), 2023.

[54] Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang, et al. EmotionCues: Emotion-oriented visual summarization of classroom videos. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3168–3181, 2021.

[55] Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems. *arXiv e-prints*, page arXiv:2401.03945, January 2024.

[56] kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal processing letters*, 23(10):1499–1503, 2016.