

Trustable and Continuous-Improving Question Answering System for Teaching and Learning

Zhengyuan Wei , Albert T.L. Lee, Victor C. S. Lee , *Member, IEEE*, King Hang Lam , *Member, IEEE*, Vincent W.L. Tam , *Member, IEEE*, and S.M. Yiu

Abstract—Generative artificial intelligence has been shown to be effective in question-answer tasks, contributing to a significant reduction in response time and greater working time compared to humans. However, the quality of generated responses might be overlooked, and unreliable responses would harm students in education. This work aims to fill this gap by introducing an AI-facilitated question-answering system, namely CertiQA, equipped with our methodology to provide trustable answers to students. This objective is achieved through the integration of AI capabilities and an authoritative knowledge base, fostering a mechanism conducive to automatic certification and continuous enhancement of knowledge through student engagement. The AI chatbot and communication-based portals enable swift actions, uninhibited inquiries, and shared responses in groups, which helps to alleviate the instructors' workload.

This work examines the system in a discipline-specialized and project-based course for third-year engineering students. The course, attended by an aggregate of 78 students, consists mainly of individuals with various backgrounds. The digital resources employed during the course's development, deployment, and evaluation are presented in detail. This work contributes to demonstrating the efficacy of this approach, which showcases promising results in terms of effectiveness and satisfaction. The positive feedback derived from the evaluation undergirds the proposition for extending the cooperative methodology to a broader spectrum of engineering curricular offerings.

Index Terms—Academic support, learning technology, artificial intelligence, technology application, higher education, pilot study.

I. INTRODUCTION

LETTING more people have access to higher education is a widely discussed and supported point of view, which has become a consensus in the field of education and social development and has gained broad support and promotion throughout the world. The number of students with higher educational degrees is growing each year; however, on the contrary, the number of instructors qualified to teach this level of degree does not match the growing of students [1]. Consequently, the demand for instructions and feedback increases proportionately with the growing student-teacher ratio. Many instructors work overtime (e.g., by exceeding

lecture sessions or office hours) to support students on this demand [2]. Therefore, many governments and institutions are working to offer resources to develop novel methodologies and innovative technologies to address this problem [3], [4]. To ease the workload of instructors, there should be convenient agents able to respond to the students so that they can seek preliminary answers anytime during their learning.

The use of artificial intelligence (AI) would be the most straightforward solution to this problem [5]. However, simply adopting generative AI (GenAI) tools, such as ChatGPT¹, in this matter could easily encounter more problems. Many research works show that the responses from GenAI tools could be unreliable [6] in terms of incorrectness [7], inconsistency [8], and factualness [9]. Moreover, a recent article [10] reports that inconsistent answers from GenAI tools influence user judgment, implying that if students rely on the vanilla GenAI tools to ask the same questions, but get different answers, they may come across conflicts in the discussions among themselves. Thus, it is in demand to regulate the use of GenAI tools so students can have a clear criterion of whether to rely on the responses from AI.

By analyzing the scenario, we find that the questions raised by students in engineering courses have two special properties. First, many questions share significantly high similarities, which are common and generally put repeatedly by different students in the same or different class sessions. Moreover, in the university environment, courses are usually taught by years, which means that the questions inherently would be recursive over the years. In addition, many of these engineering courses require students to carry out a course project. Numerous questions are raised when students are working on the course project outside of lecture time. In this scenario, to be able to solve their issues or doubts in time, it would be helpful to provide the students with an extra opportunity outside the lecture sessions so that they can continue to work on the project smoothly without blocking issues.

From the students' perspective, in the abovementioned project scenario of the engineering courses, students are usually required to group up to take a course project. We notice that the answers acquired by a student after consulting the instructor would generally be shared with his/her group mates. On the other hand, we also notice that the students' characteristics play an important role in their willingness to ask questions. More specifically, some students are too shy to ask questions when facing the instructor. These two insights

Manuscript received April 19, 2021; revised August 16, 2021. This work was supported by the University Grant Committee (UGC) and the University of Hong Kong (HKU) under the Funding for Innovative Technology-in-Education (FITE) 2023. (*Corresponding author: Albert T.L. Lee*)

The work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee (HREC) under Application No. EA240427.

The authors are with the Faculty of Engineering, the University of Hong Kong, Pofulam, Hong Kong (e-mail: zywei4@hku.hk; {tlalee, csvlee, khlam, vtam}@eee.hku.hk; smyiu@cs.hku.hk)

¹<https://chat.openai.com>

motivate us to consider developing a communication portal that enables students' uninhibited inquiry and eases the sharing action.

The paper presents an AI-facilitated conversation-based system, namely CertiQA, equipped with our methodology to assist both students and instructors in questioning and answering. CertiQA involves a large language model (LLM) [11] as an instructor agent to respond to students' questions just in time. The generation of the large language model would be augmented by a golden knowledge base whose quality is assured by the instructor. In this way, the success of the augmentation serves as an oracle² to determine whether to trust the AI response in this generation. Moreover, for unsuccessful augmentation during the generation of responses, CertiQA provides a mechanism forwarding the context (including the question and response) to the instructor for further accurate question answering. To follow up, the instructor would inspect the forwarded context from students and decide whether to add the question to the knowledge base. We developed the prototype of the system and tested it with students in an engineering course, where they are required to conduct a course project. This work aims to answer the following research questions:

- 1) How does the proposed oracle perform regarding the reliability of the generated responses?
- 2) What is the efficacy of adopting the CertiQA system in the class in terms of overall performance and satisfaction?
- 3) What is the effectiveness of LLM tools in alleviating instructors' workload of question-answering in engineering courses?

The rest of the paper is organized as follows. Section II introduces the background of the work. We present the methodology and design behind our system in Section III, followed by its evaluation in Section IV. We discuss the recommendations and limitations of our methodology in Section V. Lastly, Section VI concludes the paper.

II. BACKGROUND

A. Large Language Model

A large language model is a type of artificial neural network built on deep learning [12]. It generally contains billions of trainable parameters and requires pre-training on large-scale sets of data. LLM is famous for its superior performance in natural language understanding and processing capabilities. For example, many commercial LLMs were trained in massive amounts of data, covering different subjects, types of media, and domains. They achieved state-of-the-art performance in many benchmark datasets [13]. Therefore, question-and-answering looks suitable for LLMs in nature, thanks to the pretraining efforts on various data sources.

However, one of the biggest issues for LLMs is that the model will always generate responses for arbitrary questions, no matter whether the model has been trained on relevant data [14]. More specifically, LLMs would produce plausible

prose using false, misleading, or made-up "facts". The word *Hallucinate* has been nominated and chosen as the Cambridge Dictionary word of the year 2023 [15]. Thus, the reliability issue of AI becomes a bottleneck for its larger application.

With AI being popular and coming to education, students may be misled by AI's hallucinated responses. Detecting hallucinated responses is a non-trivial work, cause it generally requires performing a lot of fact-checking work. Our work aims to explore a solution from the opposite side, trying to find if any information can provide hints about the non-hallucinated so that we could utilize it for developing the criterion for detecting non-hallucinated responses.

B. Retrieval-Augmented Generation

Although pre-trained LLMs achieve state-of-the-art performance on many benchmark datasets, they are heavily dependent on the knowledge encoded in their parameters. However, the encoded knowledge in a single LLM could not cover all possible domain knowledge. In many knowledge-intensive tasks (e.g., question and answer in university courses), their ability to access and precisely manipulate knowledge is still limited.

Meta AI proposes retrieval-augmented generation (RAG) [16] to address the abovementioned problem. RAG introduces two sources of knowledge for the generation: the internal knowledge that LLMs store in their parameters and the external knowledge that is retrieved from a provided knowledge base. It leverages external knowledge to augment generation with internal knowledge to produce more precise answers. RAG has been demonstrated to be highly effective in generative AI applications, achieving significantly improved accuracy and reliability of LLM responses.

One of the most important advantages of RAG is that it removes the barrier of introducing new knowledge into the LLM generation without finetuning the model. Generally, finetuning LLMs also requires huge computational resources as much as (pre)training them and a significant amount of domain-specific data. RAG relieves the requirements for finetuning and the data magnitude.

C. Question Answering Task

Question answering is a computer science discipline within the fields of information retrieval and natural language processing (NLP) that is concerned with building systems that automatically answer questions in a natural language [17]. The questions are posed by students and responded to by programs in a question-answering system. The system enables students to acquire automatic feedback for their studies. Shute's work [18] summarizes five types of feedback that happen in learning scenarios. One of the most crucial types of feedback related to the question-answering system is providing certain certifications, in which students are informed about the correctness of the responses from the system.

To deliver the certification, instructors may transfer their knowledge to an AI tool that will be able to assist with certifying the responses. This work explores the solution to bridge this gap, trying to find a solution to certify the responses

²In software testing, an oracle is a mechanism that determines the correctness of the output.

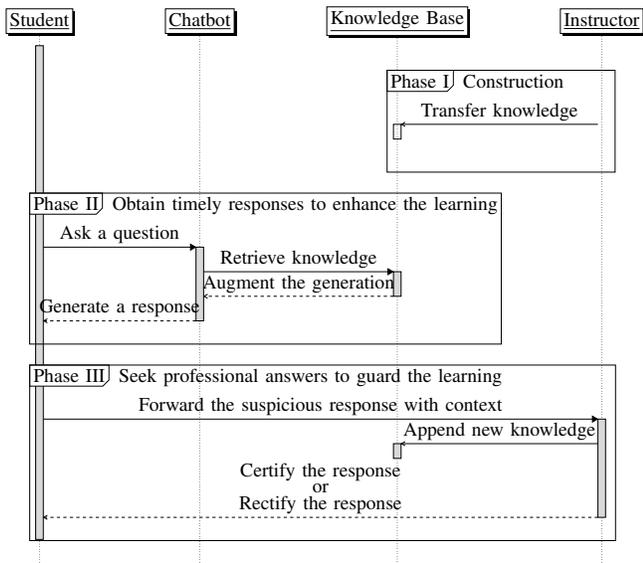


Fig. 1. The overview of the proposed methodology in the question-answering system.

from AI for the users. Consequently, students can receive signals to judge whether to trust the responses.

III. METHODOLOGY

A. Overview

We aim to build a question-answering system to enhance students' learning experience and relieve instructors' answering workload. The main goal of the system is to be able to provide students with timely responses to their questions without the physical presence of an instructor. Moreover, the system is also to be capable of indicating the correctness of the responses to some extent so that students can rely on the indicators to judge whether the response is trustworthy. In the negative case, there is also a mechanism in the system to ensure that students have a way to seek professional answers to avoid misleading by AI-generated inaccurate responses. Also, the system is to be improved if the response is checked to be erroneous. We develop a methodology to support the construction of such a question-answering system named CertiQA.

The developed methodology has three phases involving four roles: 1) the student, who will post questions in the system and forward the suspicious responses to the instructor; 2) the chatbot, which is empowered by a GenAI model and gives instant responses to student questions; 3) the knowledge base, which is constructed by the instructor beforehand to augment the generation of responses; 4) the instructor, to handle the challenging questions from students that are out of the capabilities of the adopted GenAI model. The sequence model is depicted in Fig. 1.

B. Course-Specific Knowledge Base

Phase I is the construction of the knowledge base. The main purpose of the knowledge base is two-fold: to augment the generation of GenAI models and assist in certifying the responses in the later phase.

TABLE I
AN EXAMPLE OF THE KNOWLEDGE BASE.

Index	Question & Answer
1	Q: What is the password for ELEC3848 wifi? A: "password"
2	Q: What are the use of essential libraries imported? A: 'jetson.inference' contains pre-trained neural network models for transfer learning and inference. 'jetson.utils' contains utility functions to handle tasks like video/image processing.
3	Q: My OLED screen is not displaying anything / not updating. A: Check your code and make sure that you have cleared your screen ('display.clearDisplay()') and reset the cursor ('display.setCursor(0, 0)') every time before displaying a new string and run 'display.display()' at the end. Also, make sure that the cables are correctly connected to the IIC ports.

More pairs of questions and answers are omitted here.

The augmentation feature is adopted from the RAG model [16]. The knowledge base used to augment the generation is constructed by the instructor manually to ensure the correctness. More specifically, the knowledge base consists of a set of pairs of questions and answers, where the pairs of questions and answers are stored in plain text format and with labels (i.e., "Q:" and "A:") to indicate the text fields. Table I presents an example of the constructed knowledge base. Except for a well-prepared Q&A document, it can also be lecture slides, supplementary course materials, etc. The documents in the knowledge base are related to a course the instructor is teaching.

The pairs of questions and answers in the knowledge base are preprocessed to be vectorized using the embedding model of the corresponding adopted GenAI model. During the generation process, the chatbot will query the knowledge base with the question's embedded vector and augment the generation with the retrieved knowledge.

We note that each segment of embedded vectors represents the knowledge from the instructor, where the knowledge is deemed to be always correct and precise toward answering students' questions. By leveraging the RAG model, the instructor transfers his/her knowledge into the GenAI model without finetuning the model.

C. Timely Responses with Certification

Phase II is the main loop of the proposed methodology. In this phase, students can ask the chatbot any questions related to the course. Since the chatbot is empowered by the GenAI model, it will provide instant responses to students' questions. Therefore, students can gain timely responses to clear the obstacles in their learning.

However, as aforementioned, the greatest challenge lies in that GenAI models may occasionally produce incorrect responses without any indication. In this work, we propose an oracle to address this challenge. Reviewing that in phase II, the chatbot will query the knowledge base for each coming question. As a result, there will be two possible situations on the query: a) the chatbot successfully retrieves the proper knowledge from the knowledge base to augment the generation; b) the chatbot fails to find any relevant information, resulting in no augmentation during generation. Our oracle is

STUDENT: Why is my VNC not working?
 CHATBOT: There could be a few reasons why your VNC is not working. Here are some possible solutions to try:

- Check the IP address: Make sure you have entered the correct IP address for the device you are trying to connect to [1][5].
- Verify network connection: Ensure that both devices, the Jetson Nano and the computer, are connected to the same Wi-Fi network [5].
- Password: Double-check that you are using the correct password for keyboard/mouse/VNC access, which is “nvidia” [1].

If these steps do not solve the issue, please provide more specific information or error messages, and consider reaching out to <the instructor> (the@email.address) for assistance with car assembling and sensor operations [3]. They will be able to guide you further in troubleshooting the VNC issue.

Remember to include details such as what you need to do, screenshots of any error messages, and any other relevant information to help them better understand and resolve the problem [6].

Please let me know if you need further assistance.

Fig. 2. An example of the certification in the conversation. Note that the citations here refer to the knowledge base, and the instructor information is edited out in this figure.

defined upon the *success* of the knowledge retrieval. In other words, since the instructor already guarantees the knowledge base to be correct and precise, it implies that the generated responses based on the retrieved knowledge also have these properties.

The CertiQA system implements this oracle as a citation mark on the responses. Fig. 2 presents an example of the responses that contain the citation mark. In the conversation, the chatbot successfully retrieves proper knowledge from the knowledge base. Thus, it explicitly refers to the source of the statements in its response. We note that the citation marks in the figure refer to the knowledge base rather than the reference section in this paper. Students can also click on the mark notations to refer to the original documents embedded in the knowledge base.

From the system perspective, it delivers the certification for assertive AI responses, which are visible to the users. Accordingly, students are able to capture the mark notations in the responses and then make decisions to trust these responses.

D. Professional Answers with Continuous Improvement

Phase III describes the cases where the responses do not contain citation marks. In such a situation, CertiQA also provides a mechanism to enable students to seek correct answers. Specifically, when students suspect the reliability of the received responses, they can forward the context (including

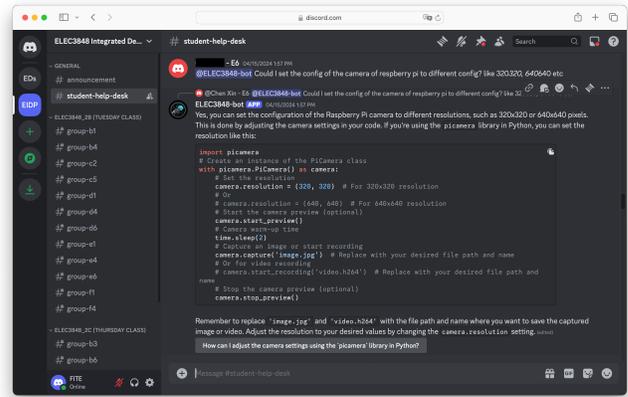


Fig. 3. GUI of the CertiQA system. Note that the student name in the figure is edited out.

the question and the AI response) to the instructor. Then, the instructor is responsible for inspecting the AI response and verifying its correctness for students. In this way, students can always have a way to obtain professional answers from the instructor.

During the verification process, the instructor would make the following two decisions: (a) to certify the response, or (b) to rectify the response. The instructor will certify the response as the correct answer to the student if he/she finds out that the response is correct and precise. On the other hand, if the instructor identifies the incorrect points of the response, he/she will rectify the response into a correct answer. In either decision, the students can obtain the correct answer.

Whenever the instructor decides to rectify the response, it is also a crucial moment to improve the knowledge base; otherwise, the instructor would still suffer from the repeated workload of inspecting the responses to those questions similar to that of the uncorrected response. In our proposed methodology, the instructor is capable of maintaining the knowledge base. More specifically, when the instructor rectifies the response, a correct answer to the question is produced. The instructor can then commit a new pair of question and answer to the knowledge base. For subsequent similar questions to the added question, the chatbot will be able to produce certified responses with the associated answer. In this manner, the knowledge base is being continuously improved. Students benefit from obtaining timely responses with certification, and the instructor benefits from relieving their repeated workload.

E. System Architecture

The developed CertiQA system is composed of two important components, namely a communication portal and a cloud computing platform. The communication portal conveniently supports the interactive actions required in our proposed methodology, while the cloud computing platform enables students to access the chatbot at any time. These two components are integrated to serve as a web service for students to use. Fig. 3 presents the user interface of CertiQA.

The communication portal is necessary for the question-answering system in our scenario. On the one hand, our

proposed methodology includes a forward action in Phase III, which a communication portal generally provides. However, as mentioned in the Introduction section, there is a strong demand from students to share the obtained responses with their group mates. Most communication portals support creating subgroups or subchannels for a subset of the users. We adopted the Discord platform³ as the communication portal in our system.

The cloud computing platform acts as a server to provide 7x24 hours of service for the students and the instructor. It should support the inference of large language models and the retrieval-augmented generation model. Currently, thanks to the popularity of LLMs, many cloud computing platforms satisfy the abovementioned requirement. Moreover, it needs to export the web interface for integration. We adopted the Coze platform⁴ as the cloud computing platform in our system.

We integrated the two platforms adopted to construct our system CertiQA. We note that the choice of platform vendors can be varied according to individual concerns as long as the abovementioned designed in the methodology are satisfied. We chose the two platforms for the consideration of ease of use in our curricula.

IV. PILOT STUDY

This section reports on a pilot study of our proposed methodology. The primary goal is to answer the aforementioned research questions. In addition, we also collected some feedback from students and gained experience in introducing AI techniques into teaching. This study enables us to lay the foundations of the system for extending it with more courses in future academic years.

A. Research Context

CertiQA has been deployed for the course *ELEC3848 Integrated Design Project* (IDP) of the Department of Electrical and Electronic Engineering at the University of Hong Kong (HKU) in Semester B of the 2023/24 academic year. The course is run for third-year students pursuing a bachelor's degree in engineering. There were 78 students enrolled in the course, and the course lasted for 13 weeks.

The IDP course requires students to conduct a group project to study and perform hands-on practice integration development between software and hardware. The topic of the project is about coordinated autonomous vehicle systems. In this project, students would accomplish some development tasks, such as identifying a particular location, traversing a pre-defined path, performing wireless charging, etc. Besides, they must apply innovation in developing the group project to demonstrate innovative ideas that further advance existing applications. During the study, students apply theoretical knowledge to develop real systems to solve real-life problems.

The biggest challenge in this course is that the course involves students of different programs (e.g., electrical and electronic, computer science, and mechanical engineering).

The different backgrounds of students implicitly reflect the different difficulties for them to perform the same tasks. As a result, they may need to refer to external documents to gain enough preliminaries or consult the instructor for more advice.

Specifically, the course requires students to perform fundamental tasks such as hardware component selection, sensor data acquisition and processing, software business logic development, and hardware-specific software testing. Students with electrical and electronic backgrounds may excel in the task of manipulating hardware components, while they would likely have difficulties with the programming tasks.

To facilitate students conducting their group projects, a lab is open for students to discuss and work on the tasks. However, since the lab opening hours are out of the lecture hours, there is no guaranteed instructor in the lab. Therefore, we developed a trustable chatbot-based system for addressing their problems. As we introduced at the beginning, one of the goals of this study is to enrich the student's learning resources and eliminate their blocking issues during development.

B. Research Design

As we presented in Section III, this work also had different phases. In this section, we introduce the work in different phases.

The first phase consisted of building the knowledge base for applying our proposed methodology. Since the IDP course has been run for many years, we gathered the questions and answers that the teaching assistants (TAs) of the IDP courses made during the previous academic years. Besides, we also collected the lecture slides, the course materials for the required development tasks, and supplementary programming materials for the current academic year (i.e., Semester B of 2023/24). We put all the above documents into the knowledge base.

Afterward, we utilized the course-specific knowledge base to construct a chatbot. The documents in the knowledge base are processed according to the RAG model [16]. We adopted GPT-4 [19] as the base GenAI model for CertiQA. The system prompt in CertiQA is constructed and tuned following the guidance of Anthropic⁵. More specifically, we generated a system prompt using their provided prompt generator by pre-defining the role of the AI model and polishing the generated prompt to be clear and direct. Finally, the chatbot is deployed into the Coze platform.

In the next phase, we registered the chatbot to the Discord platform to enable the students to access our deployed chatbot. We wrote down the instructions on the landing page of our setup Discord server, especially for the usage of the chatbot, the certification of the responses, and the forwarding action they could perform to seek professional answers. Students can ask not only software and hardware-related questions, but also course-related questions (e.g., lab opening hours, due date of the assignments). They can also ask questions with images and links thanks to the multimodal capabilities of the AI models, which significantly enriches their learning resources. We also created different group channels for students to share their

³<https://discord.com/>

⁴<https://www.coze.com/>

⁵<https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering>

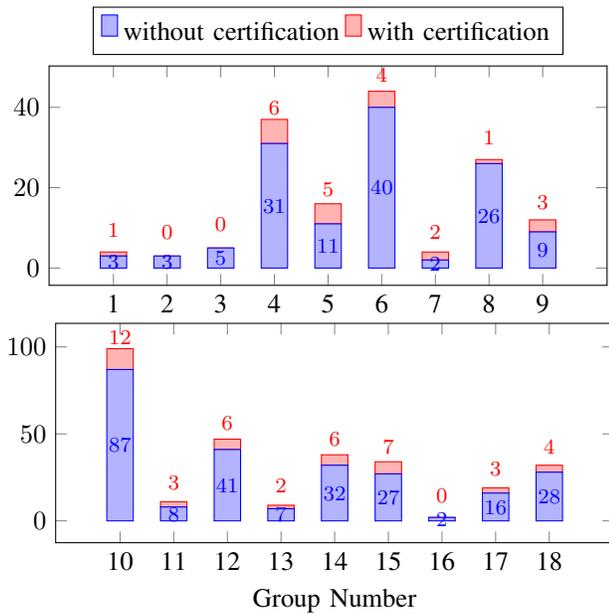


Fig. 4. Distribution of the responses from the chatbot in the group channels.

queries with their group mates. Fig. 3 shows this setting. Students can choose to ask questions to the chatbot individually or within the group channel to protect their privacy from other students, which is helpful in promoting their willingness to ask questions.

The fourth phase consisted of maintaining the system. The system is open 7x24 hours to the students, during which the instructor would receive the forwarded suspicious responses with context from the students. As our methodology designs, we appended the new rectified responses into the knowledge base during the maintenance time each week. This operation was maintained throughout the course until the final exam in the 2023/24 academic year.

At the end of the course, we conducted a satisfaction questionnaire to evaluate our system as a proxy to assess our methodology. The 78 students were requested to fill out the questionnaire, and we collected their responses. The questionnaire was also designed to collect feedback from the students to improve CertiQA. The questionnaire was completely anonymous. On the other hand, we also manually inspected the AI-generated responses with certifications to assess the effectiveness of our proposed oracle. Lastly, we also compared the number of questions each individual in the teaching team received in CertiQA with the number of questions each received by email in the past year. This comparison was to empirically evaluate the effectiveness of AI applications in alleviating instructors' workload of question-answering.

Finally, we answer the research questions according to the data analysis of our collected feedback. We discuss the findings of our study in Section V.

C. Oracle Performance

Recap that since some students may inquire about the chatbot privately, we could not access the AI responses in their private chat. Thus, this evaluation was conducted on the responses in the group channels only. 78 students in this course were split into 18 groups. The distribution of the AI responses is presented in Fig. 4. The x-axis indicates the group number and the y-axis shows the number of responses generated by the chatbot. The number of responses with and without certification is demonstrated in different colors, where the ones with certification are in red while the remaining ones are in blue.

One of the key findings is that there was a relatively small proportion of responses containing certification. More specifically, we collected 443 AI-generated responses from the 18 group channels, of which 65 responses contained the certification (i.e., the citation marks presented in Section III-C). The resulting certified ratio was 14.67%. This was due to the fact that many student-asked questions were not covered by our current knowledge base. The existing coverage is reasonable, as we are still in the first version of the question-answering system. Also, we would like to restate that the purpose of the knowledge base was not to include all the questions from students, rather it was to enhance the AI generation process that may produce imprecise or incorrect responses. The performance of the oracle that we build upon to produce the certification is more important.

In order to assess the performance of the proposed oracle, we manually inspect the 65 certified responses to measure the certification accuracy. The certification accuracy is measured as follows. We read each certified response with its context (i.e., the question asked by the student and the associated knowledge retrieved from the knowledge base). If the certified response answers the question correctly with the associated knowledge, this response is deemed to be successfully certified. Otherwise, it is deemed as unsuccessful. After inspection, we found that *all* the certified responses correctly answered the student's questions. In other words, the empirical certification accuracy was 100% in our study, indicating that the oracle is flawless in verifying the reliability of AI responses.

Answering RQ1, the proposed oracle performs effectively in certifying the reliability of certain AI-generated responses.

D. Satisfaction Questionnaire

The questionnaire was composed of nine questions in total. Among the questions, six of them (Q1, Q3 – Q7) were to be answered with a single item based on a Likert scale [20], where a Likert item is simply a satisfaction statement that the students were asked to evaluate by giving it a quantitative option on their objective dimension, with levels of agreement/disagreement being the dimension most commonly used. A multiple choice question (Q2) was asked for the reasons that attracted the students to use our developed system, with an open-ended option to allow students to provide their additional reasons. A binary choice question (Q8) was asked to evaluate their recommendations for applying our system in other courses. Last, the questionnaire also included an optional

TABLE II
QUESTIONS OF THE QUESTIONNAIRE

ID	Question
Q1	How often did you use the class chatbot during the semester?
Q2	What are the reasons attracting you to use the class chatbot?
Q3	Use of the chatbot increases my willingness to ask questions in class or after class.
Q4	Did the chatbot help you save time in getting answers to your questions compared to other resources (e.g., instructor’s office hours, online forums)?
Q5	Rate your overall experience with the class chatbot.
Q6	How helpful was the chatbot in answering your questions related to the project work?
Q7	How would you rate the accuracy and reliability of the information provided by the chatbot?
Q8	Would you recommend the use of the teaching chatbot in other courses?
Q9	Please provide any additional comments or suggestions for improving the class chatbot.

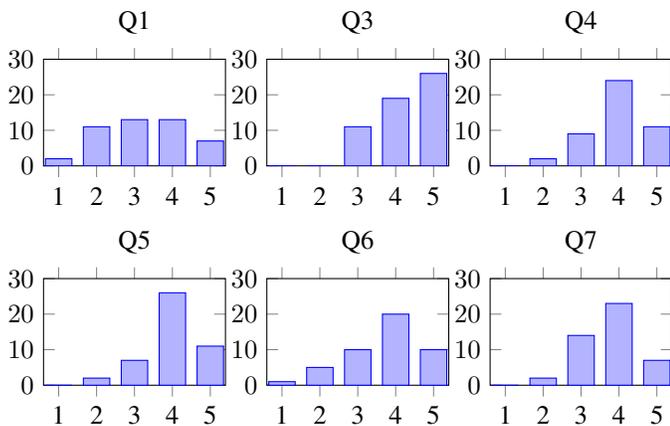


Fig. 5. Distribution of the answers to the Likert scale-based questions.

open question (Q9) to collect comments or suggestions from students for improving our system. The questions are presented in Table II.

We received 46 responses from 78 students on our questionnaire. To measure the reliability of the questionnaire, we conducted the Cronbach’s alpha test [21] on the questionnaire results. The results are also presented in Fig. 5. The resultant Cronbach’s-alpha coefficient was 0.82, indicating that the acceptable level of reliability of the questionnaire is good according to the interpretation table from Eugene and Moayad [22].

We investigated the usage of CertiQA and the reasons that attract students to use CertiQA in the questionnaire. The students showed a strong interest in giving CertiQA a try according to the results of Q1. The results are shown in Fig. 6. We note that this question inquired about the usage including the private chats from students to the chatbot in CertiQA so that we could not simply count the responses we collected. 95.65% of the investigated students answered that they had used CertiQA in their study. In addition, 43.48% of the students stated that they used CertiQA frequently or very frequently.

Regarding attraction reasons, we provided five preset options and an open-ended response field to state their opinions

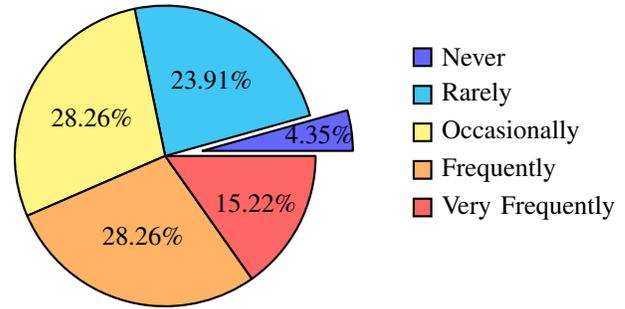


Fig. 6. Usage Frequency including private chats.

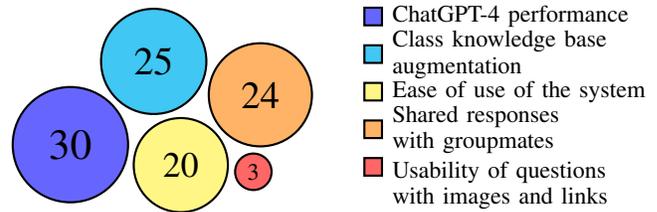


Fig. 7. Reasons attracting students to use our system with voting counts.

in Q2. Fig. 7 shows the voting results of the students. The most attractive reason for CertiQA is that we integrated the state-of-the-art AI model (i.e., ChatGPT-4 at the period of this work) into our system to provide responses to the questions from the students. Following this reason, many students voted for the options of *Class knowledge base augmentation*, *Ease of use of the system*, and *Shared responses with groupmates* with close voting counts. In contrast, the voting counts for the *Usability of questions with images and links* was significantly less than the others. This may be because our course was about programming with hardware modules, so students do not have many questions with multimedia associates.

In the open-ended responses of Q2, we collected 13 effective reactions from students. The responses without meaningful opinions were filtered out, e.g., “No” and “N.A.”. Besides the opinions similar to the preset options, three responses provided additional reasons. 1) *Code related matters*. The integrated AI model with superior performance related to the course attracts the students. 2) *TAs can also see the questions and help*. Our methodology of seeking professional answers to guard the learning plays an important role for the students. 3) *It is accurate*. The satisfactory accuracy of the responses is a key point for the students. Further discussion in this regard can be found in Section V.

In general, CertiQA increased students’ willingness to ask questions in their study according to Q3. Fig. 8 presents the results. More than 70% of the surveyed students expressed an increase in willingness to some extent, i.e., a slight increase with 41.31% and a strong increase with 34.78%. As we analyze in Section I, the increase may be due to the factor of no embarrassment that students have no such concerns when they interact with a software system. The remaining students put a neutral attitude toward the promotion of their willingness to ask questions. In the answers to Q3, we did not receive any

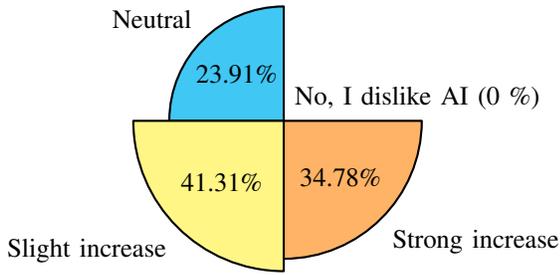


Fig. 8. Increase of willingness to ask questions.

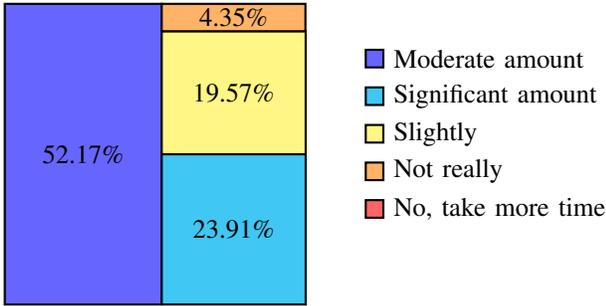


Fig. 9. Save time in getting answers to the questions.

reports from the students that they had negative aspects of the use of AI that harmed their willingness to ask a question.

Besides, the results of Q4 in Fig. 9 demonstrate that our AI-facilitated question-answering system can help to save students' time in getting answers to their questions in the context of our course. As aforementioned, students were required to conduct a group project that they would work on even out of the lecture time. A 7x24 available system provided them with instant feedback to address their blocking issues. More than 90% of the surveyed students reported that CertiQA saved them time in getting answers, among which 52.17% of them opted for a moderate amount of time savings, followed by a significant amount (23.91%) and a slight amount (19.57%) of time, respectively. 4.35% of the students felt uncertain about saving time, but no student reported that they spent more time to get answers through CertiQA.

In view of the satisfaction questions (Q5 – Q7), we investigated overall experience, helpfulness, and accuracy & reliability from the student's perspective. As shown in Fig. 10, the students gave positive opinions regarding the overall experience of CertiQA (Q5). They felt the system was helpful (Q6) and accurate & reliable (Q7). The satisfaction ratios (accounting for the top two levels in a five-item Likert scale) are 81%, 65%, and 65%, respectively. All three dimensions have surpassed 50%, indicating a favorable satisfaction of CertiQA from students.

Answering RQ2, the CertiQA system demonstrates satisfactory performance in the question-answering task in university courses.

E. Workload Reduction

To assess the alleviation of instructor's workload of question-answering task, we compare the numbers of query

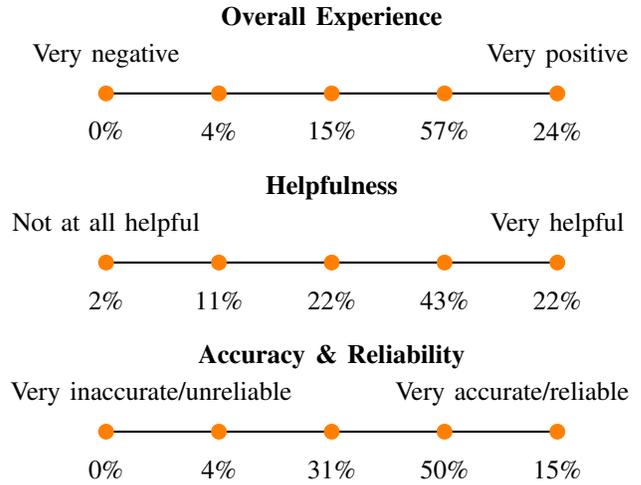


Fig. 10. Satisfaction evaluation on the CertiQA system.

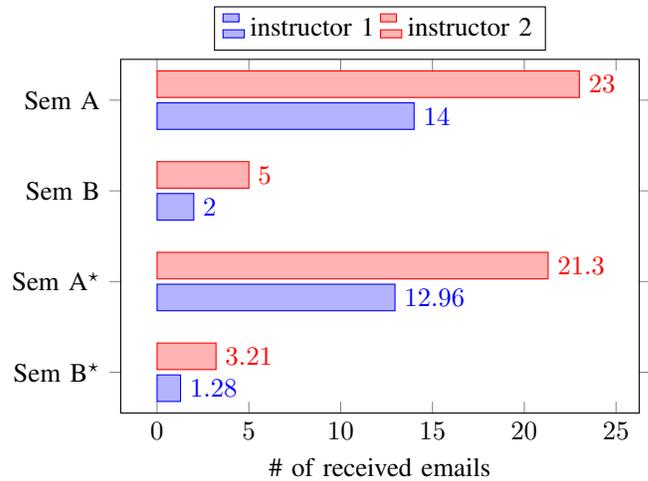


Fig. 11. Workload assessment results with and without CertiQA. The * symbol stands for normalizing the number of students.

emails received by the instructor with and without the present of CertiQA system. The IDP course was taught together by two instructors in the last two semesters, i.e., Semester A and Semester B in the 2023/24 academic year. In Semester A, there were 54 students enrolled in the course and CertiQA was not introduced, while in Semester B, the number of students was 78 and CertiQA was deployed.

Fig. 11 shows the number of query emails received by two instructors. Instructor 1 received 14 emails in Semester A and 2 emails in Semester B. Instructor 2 received 23 and 5 emails in Semester A and B, respectively. To eliminate the impact of different numbers of students in two semesters, we normalize the data to the same number of students (i.e., 50 students), presenting in the bottom two bars. To sum up, in the semester without CertiQA, the number of query emails was 34.26 per 50 students. With the assistant of CertiQA, the email number decreased to 4.49 per 50 students. The reduction rate was 86.89%, which is significant. The reduced number of queries in the email system would be inherently transit to the CertiQA system that continuously improves the course-specific knowledge

base.

Answering RQ3, adopting LLMs helps alleviate the instructor's workload in question-answering tasks.

V. DISCUSSION

Although the above results of the pilot study seem promising, the system, as well as its underlying methodology, has its applicability and limitations. In this section, we discuss the generalization of the methodology and the benefits and limitations of CertiQA. Last, we state our recommendations for applying this methodology in suitable curricula.

A. Generalization

The generalization of our proposed methodology to apply AI in education lies mainly in Phase II and Phase III (Fig. 1). A quality-assured knowledge base is utilized to provide certifications for AI-generated responses in Phase II. Different courses in different fields would have their domain-specific knowledge to guard the answers to students. Our method applies if the domain-specific knowledge can be cast into a text-based knowledge base so that the AI generation can benefit from the certifications.

On the other hand, in Phase III, CertiQA is equipped with a mechanism to improve the AI generation continuously. We collected 443 AI-generated responses in our study, part of which would be incorporated into the knowledge base to improve the effectiveness of CertiQA. The mechanism enables continuous improvement, and we believe that the volume of the knowledge base will grow and consolidate in a few more semesters after its launch.

B. Benefits & Limitations

An obvious drawback of applying the methodology is that the instructor needs to construct a high-quality knowledge base. In our practice, we involved three teaching assistants to help in constructing the initial knowledge base. After obtaining the initial version, the instructor inspected every item in the knowledge base to ensure that the content was not incorrect or misleading. The construction process took several weeks for the teaching team to prepare the deployed version.

Although the construction process is non-trivial, the knowledge base does not have to be comprehensive at its first stage. It is well known that large language models inevitably produce hallucinated responses without supporting knowledge in the generation process [23], [24]. Students play an important role in discovering the deficiency of the adopted AI model in Phase III of the methodology, which helps to refine and enrich the knowledge base.

C. Recommendations

The core advantage of CertiQA is that it provides certifications for certain AI-generated responses. The certification is delivered upon the success of the knowledge retrieval from the knowledge base. In other words, the more frequently the in-knowledge-base questions are asked, the more certifications are produced. Therefore, it is highly recommended to apply to courses with either intensive or repeated questions. Courses requiring group projects usually have these properties.

VI. CONCLUSION

This paper presents an AI-facilitated question-answering system CertiQA and its underlying methodology aimed at strengthening teaching and learning. The methodology enables a trustable and continuous-improving system benefiting both students and instructors. CertiQA has been run in a real course with students in different backgrounds and operated for a semester with a follow-up evaluation. The evaluation result for this pilot study is promising, showing the potential to incorporate AI in university courses. We discuss the pros and cons of the methodology and put our recommendations. In the future, we aim to extend the system to more courses and refine the methodology further.

ACKNOWLEDGMENTS

The authors would like to thank all students in the course for providing invaluable feedback that improves the system and the teaching assistants for preparing the course-specific knowledge base.

REFERENCES

- [1] National Academies of Sciences, Engineering, and Medicine, *Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments*. Washington, DC: The National Academies Press, 2018. [Online]. Available: <https://doi.org/10.17226/24926>
- [2] E. D. Steiner, A. Woo, and S. Doan, *All Work and No Pay — Teachers' Perceptions of Their Pay and Hours Worked: Findings from the 2023 State of the American Teacher Survey*. Santa Monica, CA: RAND Corporation, 2023.
- [3] T. I. C. R. and V. O. D., "On the futures of technology in education: Emerging trends and policy implications," Luxembourg (Luxembourg), Anticipation and foresight KJ-NA-31-652-EN-N (online), 2023.
- [4] University Grants Committee, "Ugc establishes fund for innovative technology-in-education," 2023, accessed on May 14, 2024. [Online]. Available: https://www.ugc.edu.hk/eng/ugc/about/press_speech_other/press/2023/pr20230628.html
- [5] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.
- [6] I. Amaro, A. Della Greca, R. Francese, G. Tortora, and C. Tucci, "Ai unreliable answers: A case study on chatgpt," in *Artificial Intelligence in HCI*, H. Degen and S. Ntoa, Eds. Cham: Springer Nature Switzerland, 2023, pp. 23–40.
- [7] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. Huang, "A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 431–469. [Online]. Available: <https://aclanthology.org/2023.findings-acl.29>
- [8] M. Jang and T. Lukasiewicz, "Consistency analysis of ChatGPT," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 970–15 985. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.991>
- [9] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6449–6464. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.397>
- [10] S. Krügel, A. Ostermaier, and M. Uhl, "Chatgpt's inconsistent moral advice influences users' judgment," *Scientific Reports*, vol. 13, no. 1, p. 4569, Apr 2023. [Online]. Available: <https://doi.org/10.1038/s41598-023-31341-0>

- [11] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier *et al.*, “Chatgpt for good? on opportunities and challenges of large language models for education,” *Learning and individual differences*, vol. 103, p. 102274, 2023.
- [12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [13] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac’h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron, L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou, “A framework for few-shot language model evaluation,” 12 2023. [Online]. Available: <https://zenodo.org/records/10256836>
- [14] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier *et al.*, “Chatgpt for good? on opportunities and challenges of large language models for education,” *Learning and individual differences*, vol. 103, p. 102274, 2023.
- [15] Cambridge Dictionary, “‘hallucinate’ is cambridge dictionary’s word of the year 2023,” 2023, accessed on May 24, 2024. [Online]. Available: <https://www.cambridge.org/news-and-insights/hallucinate-is-cambridge-word-of-the-year-2023>
- [16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- [17] P. Cimiano, C. Unger, and J. McCrae, *Ontology-Based Interpretation of Natural Language*, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2014. [Online]. Available: <https://books.google.com.hk/books?id=3GA-AwAAQBAJ>
- [18] V. J. Shute, “Focus on formative feedback,” *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008. [Online]. Available: <https://doi.org/10.3102/0034654307313795>
- [19] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya *et al.*, “Gpt-4 technical report,” 2024. [Online]. Available: <https://arxiv.org/abs/2303.08774>
- [20] R. Likert, “A technique for the measurement of attitudes,” *Archives of psychology*, 1932.
- [21] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.
- [22] E. Allevato and M. Al Marrar, “A project-based approach to examine university teachers attitudes towards visually impaired students,” *Scholarly Journal of Psychology and Behavioral Sciences*, vol. 2, no. 10.32474, 2019.
- [23] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, mar 2023. [Online]. Available: <https://doi.org/10.1145/3571730>
- [24] Z. Xu, S. Jain, and M. Kankanhalli, “Hallucination is inevitable: An innate limitation of large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.11817>