# A Step Towards Adaptive Online Learning: Exploring the Role of GPT as Virtual Teaching Assistants in Online Education

**Xiner Liu[1], Maciej Pankiewicz[1], Tanvi Gupta[1], Zhongtian Huang[1] & Ryan S. Baker[1]**
[a]*University of Pennsylvania, U.S.A*
xiner@upenn.edu

**Abstract:** With student learning becoming more continuous and ubiquitous, online courses are increasingly challenged to provide timely support to learners. Human TAs, constrained by limited availability, often delay in addressing student inquiries occurring on weekends or at night. This paper presents JeepyTA, a Virtual Teaching Assistant (VTA) built on GPT model designed to provide round-the-clock assistance by leveraging OpenAI's text embeddings and generative language models. JeepyTA provides responses that mimic typical discourse in discussion forums and, although still limited in scope, addresses logistic, conceptual, and programming questions tailored to specific courses much quicker than human TAs can. In this paper, we outline our development process, analyze JeepyTA's response accuracy and compare its availability to human instructors, investigate student attitudes, and discuss the implications of integrating large language models like JeepyTA into educational settings. This work contributes to understanding how LLMs could improve the timeliness and availability of student support, offering on-the-spot assistance, and delivering personalized feedback.

**Keywords:** Virtual Teaching Assistant, Large Language Model, Discussion Forum, GPT, Embedding-based Search, Retrieval Augmented Generation

## 1. Introduction

As the landscape of higher education increasingly shifts toward online learning, the demand for timely and personalized support has grown significantly (Walsh et al., 2024). In traditional in-person classrooms, students typically have direct access to instructors and teaching assistants (Larson et al., 2023), whether through scheduled office hours, informal interactions after class, or quick one-on-one consultations. These in-person opportunities provide students with immediate feedback and tailored guidance. However, in the context of online education, these interactions are often more fragmented or delayed (Hodge & Chenelle, 2018), particularly in large-scale courses where individual support from human TAs may be limited or difficult to manage (Kearns, 2021; Hew et al., 2021). This challenge becomes even more pronounced in asynchronous online learning environments, where students may ask questions or requests for help and then face long waiting times for a response (Wang & Woo, 2007). In some cases, feedback may not arrive until after a crucial deadline has passed, which may significantly hinder the learning process and lead to frustration.

This growing gap between student need and available support has prompted the need to explore technological solutions that can offer real-time, on-demand assistance. As a response to these growing challenges, artificial intelligence (AI) technologies have emerged as potential solutions. AI-driven tools have been prominent in education for years (Roll & Wylie, 2016; Chen et al., 2020), but most extant systems have involved narrow interaction (e.g. Anderson et al., 1995) or limited scope of content (e.g. Nye et al. 2014). Contemporary large language models, by contrast, can be used in a broader range of contexts, with full natural-language interaction (Brown et al. 2020). Generative Pre-trained Transformer (GPT), a series of large language models which everyday users can interact with through a chatbot, has quickly gained a large user base. Its language processing capabilities allow it to behave as if it comprehends the context and meaning of words in user queries and provide accurate answers based on its extensive knowledge base. Its adaptability and fine-tuning capabilities along with API access make it a versatile solution for various applications in education (e.g., Tsai et al., 2021; Lagakis et al, 2023; Pankiewicz & Baker, 2023; Doughty et al., 2024; Bernal et al., 2024).

In this paper, we discuss our efforts to embed the GPT engine into a university-level course as a virtual teaching assistant, JeepyTA. JeepyTA leverages GPT's capabilities to address several needs in contemporary university courses. Firstly, JeepyTA can review and respond to student discussion posts on online forums involving questions about the course. This reduces the workload for educators who traditionally spend substantial time addressing forum queries, allowing them to allocate more time to other aspects of the course, such as working one-on-one with students. By acting as a first point of contact for student inquiries, JeepyTA's goal is to improve the efficiency of administrative tasks, freeing up human educators to focus on more complex aspects of teaching and learning. Its round-the-clock availability is another significant advantage as students can receive near-immediate responses to their inquiries, even during odd hours or outside of teaching assistants' or professors' working hours. Beyond this, JeepyTA's ability to analyze and classify the content of discussion posts, identify key points, and generate relevant responses has the potential to improve the quality of interactions in online learning environments. Its capacity to automatically generate prompts and questions for classroom discussions may play a useful role in encouraging students to think critically and engage in meaningful conversations. JeepyTA is also able to assist students in debugging their code. This application of GPT not only has the potential to enhance the overall student experience but also ensures immediate access to academic support right when it is needed.

## 2. Related Work

### *2.1 Online Discussion Forums in Education*

The development of forum-based support for teaching has emerged as a potent strategy for facilitating discourse and fostering proactive student engagement (Zhang et al. 2018; Daher et al. 2021). In the virtual realm, online forums become a "third space" (Bhabha, 1990) which promotes faculty-student interactions within an open and collaborative environment. They improve students' learning engagement and motivation, while also reducing procrastination (Kang et al., 2023).

Online discussion forums present opportunities for interactive learning, inquiry-based learning, and effective communication among students and instructors. To make the most of their impact, it is crucial to have substantial participation from both students and instructors (Onyema et al., 2019, Andres et al. 2018). Empirical evidence indicates that active participation in online discussions correlates with better academic performance (Lindblom-Ylänne et al., 2003), emphasizing the potential value of instructional interventions to enhance engagement (Chen, 2024).

Furthermore, a good discussion forum can help mitigate the fact that many students are unable to meet with teaching assistants and faculty during office hours due to factors such as work schedule conflicts (Abdul-Wahab et al., 2019). Students frequently need assistance during unconventional hours which highlights the limitations of conventional support systems (Mounsey et al., 2013). Educators also face the difficulty of responding to questions promptly, particularly after lecture hours and during peak exam preparation periods (Knobloch et al., 2018). This absence of immediate support can negatively impact student satisfaction (Després-Bedward et al. 2018). While teaching assistants serve as valuable resources, their availability, similarly to the instructors', may be constrained by their own commitments. As such, there may be benefits from creating more readily accessible forms of student assistance (Mirzajani et al., 2016; Knobloch et al., 2018).

Given limited time, lecturers cannot distribute their attention equally to all students. While not all posts require immediate instructor attention, other posts may be critical. If critical posts are not responded to in a timely fashion, it may negatively impact students' motivation and engagement (Després-Bedward et al. 2018). In one analysis performed on the data originating from educational discussion forums, as many as 20% of posts were urgent (Khodeir, 2021).

Some work has attempted to focus instructors’ time by automatically detecting which forum posts are most urgent (Khodeir, 2021; Svabensky et al., 2023).

### *2.2 LLM-Powered Virtual Teaching Assistants and Educational Tools*

Automated question-answering methods have thus far required manual mapping of potential questions and teaching context to be able to respond to queries on course content (e.g., Knobloch et al., 2018; Saleh et al., 2022). Virtual teaching assistants focused on addressing frequently-asked logistics questions and content-related factual questions can be helpful in reducing the workload for instructors and TAs by automating the routine part of instructor-student interaction (Zylich et al., 2020). For example, the AI-augmented intelligent educational assistance framework developed by Sajja et al. (2023) leverages fine-tuned GPT-3 (Davinci) to automatically generate virtual assistants given a course syllabus. This tool can answer questions related to curriculum, logistics and course policies and customize responses based on the sentiment of students’ questions. However, while this system helps overcome communication barriers between students and instructors, it still struggles to correctly respond to course/logistics questions when this information is not clearly provided to it (Sajja et al., 2023).

Taneja et al. (2024) developed Jill Watson, a VTA powered by GPT-3.5, which provides instant responses to course-related queries using materials like slides, notes, and syllabi. While it performs well in generating accurate, relevant responses, Jill Watson struggles with tasks requiring understanding of longer text, such as summarizing entire chapters, unless explicit summaries are provided in the text. Similarly, Dong et al. (2023) proposed an AI tutor using GPT API and Retrieval-Augmented Generation to address student queries by retrieving and referencing course-specific materials. Their AI Tutor showed strong performance in providing accurate and contextually relevant responses to qualitative queries (where answers are more conceptual, descriptive, or open-ended) and included citations to validate sources. However, it faced limitations with summarization tasks, quantitative problems in complex calculations, and information hallucination.

Beyond answering questions, LLM-based tools are also being explored for other educational applications. Mehta et al. (2023) explored ChatGPT’s role in providing constructive feedback on programming assignments and its ability to auto-grade programs. They found that while ChatGPT is good at identifying areas for improvement and suggesting refinements in code structure and logic, it struggles with reliably grading either the correctness or the quality of code. Pankiewicz and Baker (2023) implemented GPT model for automated generation of feedback for programming assignments on an educational platform. They observed increased performance in task solving among students receiving the GPT feedback, but also noted a drop when the GPT feedback was blended out which they attributed to the over-reliance on AI support. Chen et al. (2024) designed an intelligent tutoring system, ChatTutor, powered by chained LLMs, which engaged in real-time dialogues with the learner, adjusting teaching strategies (e.g., modifying lesson pacing, content depth, or quiz difficulty) based on the learner’s progress and preferences. However, limitations included occasional hallucinations in the generated content, delays in response times, and challenges in ensuring content validity and objectivity. Lastly, Sajja et al. (2024) explored the development of an VirtualTA to answer student inquiries, generate quizzes and flashcards, offer personalized learning pathways, and provide support in course-related topics. Their findings showed that the system successfully provided easy access to information. However, the paper noted that challenges exist in handling unstructured input data, particularly from scanned PDF files, due to imprecise content parsing.

Hence, projects to support learning with LLMs have been successful in many ways but have had some technical challenges. The adoption of LLM-powered teaching assistants and related tools has also faced some skepticism from students in higher education. For example, Kim et al. (2020) conducted a survey among undergraduate students evaluating their attitudes toward emerging technologies (such as Apple's Siri and Amazon's Alexa) and their impressions of an AI teaching assistant created by a U.S. professor. The study found that the perceived usefulness and ease of communication with AI teaching assistants play a crucial role in influencing their adoption, ultimately predicting whether students have positive attitudes toward their use.

## 3. JeepyTA

We named the AI chatbot introduced in the course discussion forum JeepyTA, a combination of "GPT" (from the OpenAI language model it is based on) and "TA" (its role as a simulated teaching assistant). JeepyTA leverages the dialogue feature of the pre-trained language model GPT and is further adapted with course-specific materials. This additional adaptation allows it to respond in ways that are relevant to the course. JeepyTA's main function in the forum is to respond to student questions and interact with their comments.

For the pilot deployment of JeepyTA we used Flarum, an open-source discussion platform. It provides an extensible architecture, suitable for the integration of additional features, such as forum bots. We developed an extension to send requests to the GPT API, generating responses to student posts and publishing them in the name of JeepyTA on the forum. Additionally, we also created functionality enabling the instructor to select categories in which JeepyTA interacts with students, define categories where JeepyTA responses require moderation (by the instructor or TA) before being published, and add specific prompts for each of the categories. Students were also given the option to choose not to have the content they generate on the platform sent to JeepyTA.

Unlike in chats, where the communication happens real-time, discussion forum users do not expect instantaneous replies. Therefore, JeepyTA's responses were not generated immediately, but with a random delay of 60-120 seconds. Asynchronous generation of responses in this scenario has benefits: we are less impacted by longer API response times, request and token limits or additional data processing pipelines.

### *3.1 Design of JeepyTA*

JeepyTA was first deployed in a graduate-level Educational Data Mining course at a large private university in the Northeastern United States in Fall 2023. A significant component of the course involves students sharing their programs, along with the methodologies and steps they used to solve the assigned problems, to exchange ideas. In JeepyTA, each action, along with the content created or modified, is recorded in the log data along with the timestamp, user ID, and the forum category in which the post was published. Images are recorded as an image preview URL in the log data.

Throughout the first semester when JeepyTA was operational, responses it generated were not immediately published. The instructor and TAs were notified via email about these responses and decided to either approve or reject each response. Instructors also had the opportunity to modify any generated response before it becomes accessible to students. This additional layer was implemented to prevent misleading, erroneous, inappropriate, biased, or non-useful responses from JeepyTA and to aid in collecting insights for ongoing improvement.

At the start of its implementation, JeepyTA was configured to respond to all student posts. This setup allowed instructors to evaluate its performance in handling different types of interactions and to identify specific areas for improvement. However, in its current version, JeepyTA allows instructors and TAs to define response parameters based on pedagogical needs. For instance, it can be set to reply only to the first post in an assignment thread, as subsequent posts are typically peer-to-peer discussions that do not require automated feedback. Meanwhile, responses can be disabled for specific categories, such as administrative announcements or casual conversation, to ensure JeepyTA’s outputs remain focused and aligned with instructional priorities.

### *3.2 Constructing JeepyTA*

Customizing the language model is essential for developing a course-specific AI teaching assistant. While GPT-based models possess a comprehensive ability to respond to questions involving general knowledge, programming, and problem-solving skills, they lack awareness of information beyond their training data. Moreover, the specific knowledge or practices taught in a course might not align with what GPT models were trained on. For instance, in Educational Data Mining, student-level cross-validation is the primary approach used to validate behavior models, as this method assesses the degree to which the model generalizes to data from unseen students. If students consult ChatGPT, however, it is likely to suggest traditional flat cross-validation methods or a flat train-test split to validate the model, which are legitimate approaches in general but less appropriate in this context. Therefore, in this specific case, our goal is to adapt the model with course-related details and knowledge such as syllabi, course schedules, lecture slides, assignment descriptions, and frequently asked questions/answers from previous years. The challenge lies not only in adapting the model to understand these contents but also in ensuring that it can provide accurate, helpful, and timely responses to both general and course-specific queries from students.

There are two primary ways for a GPT-based model to learn: updating its model weights or incorporating additional inputs into the model (Cselle & Rajgor, 2022). These correspond to fine-tuning and embedding-based search. Fine-tuning entails adjusting the model's parameters by exposing it to specialized content (e.g. see work done by Yu et al., 2021). During this process, the model's internal parameters are adjusted to better align with the new dataset. This enables the model to incorporate information from the training materials and to acquire the distinctive patterns and information relevant to them. However, fine-tuning has its limitations, particularly in tasks requiring precise factual recall, as the model may inadvertently lose some details post-training (Cselle & Rajgor, 2022).

The other way for GPT to learn, embedding, involves a process of converting words, phrases, or documents into numerical vectors suitable as input (Peng et al., 2023). The process of integrating course-related information into GPT-based models includes converting this information into embeddings and then combining them with the model's existing embeddings. This method (Retrieval Augmented Generation – RAG) does not modify the pre-trained model but instead forms a hybrid representation that fuses the model's general knowledge with specific data. As a result, there is no additional training time required. This approach is like "taking an exam with open notes", helping the model to provide consistent and intended outputs (Cselle & Rajgor, 2022). Within this project, we used the embedding-based, RAG approach. In the version of JeepyTA discussed in this paper, we selected GPT-3.5 over the GPT-4 model due to the token size limitations in GPT-4 models at the time of development, which could not accommodate our extensive course materials. However, in the current version of JeepyTA, instructors can choose between GPT-3.5, GPT-4, and GPT-4o based on their needs and preferences.

### *3.3 Embedding*

To prepare for embedding-based search, we created a collection of demonstration conversations, formatted as question-and-answer pairs, drawn from the course material (both logistics and course content). Both types of questions were based on questions asked on a discussion forum in the past three iterations of the course. This mimics potential student interactions with the model during class sessions. We decided against using the original syllabus and course schedule, as their concise and structured presentation might not be readily interpretable by the language model (although we should note that, for current version of JeepyTA, which is powered by more advanced models like GPT-4o, this preparatory step is no longer required. Such models can work directly with structured raw files such as syllabi or slides, which significantly reduced the human effort in consolidating course content into Q&A pairs while maintaining high accuracy). The final dataset consists of 279 question-and-answer pairs.

The dataset was converted into embeddings using the OpenAI text-embedding-ada-002 engine, selected for its efficiency and cost-effectiveness in diverse applications. When a student creates a forum post, its content is first encoded into embeddings using the same method as used for converting the input data. Following this, we may compute the cosine similarity in the spatial domain between the query embedding and the embeddings of the answers in the dataset. The answers are then ranked based on their similarity scores relative to the query embedding. Answers with the highest similarity scores are regarded as most pertinent to the user's query. Then, we used the gpt-3.5-turbo model to use rephrase the answer to match the style of a discussion forum. In situations where the similarity score between the query embedding and top-scoring answer embedding is below 0.70 (this threshold was chosen based on observation from 50 test cases, where answers with scores below this value tended to be less relevant or insufficiently detailed) or the top-scoring answer does not adequately address the question, we instruct JeepyTA to generate a response based on its existing knowledge base or advises students to contact the course instructor or human TA for further assistance. Figure 1 outlines the process flow JeepyTA follows to process and respond to student queries.

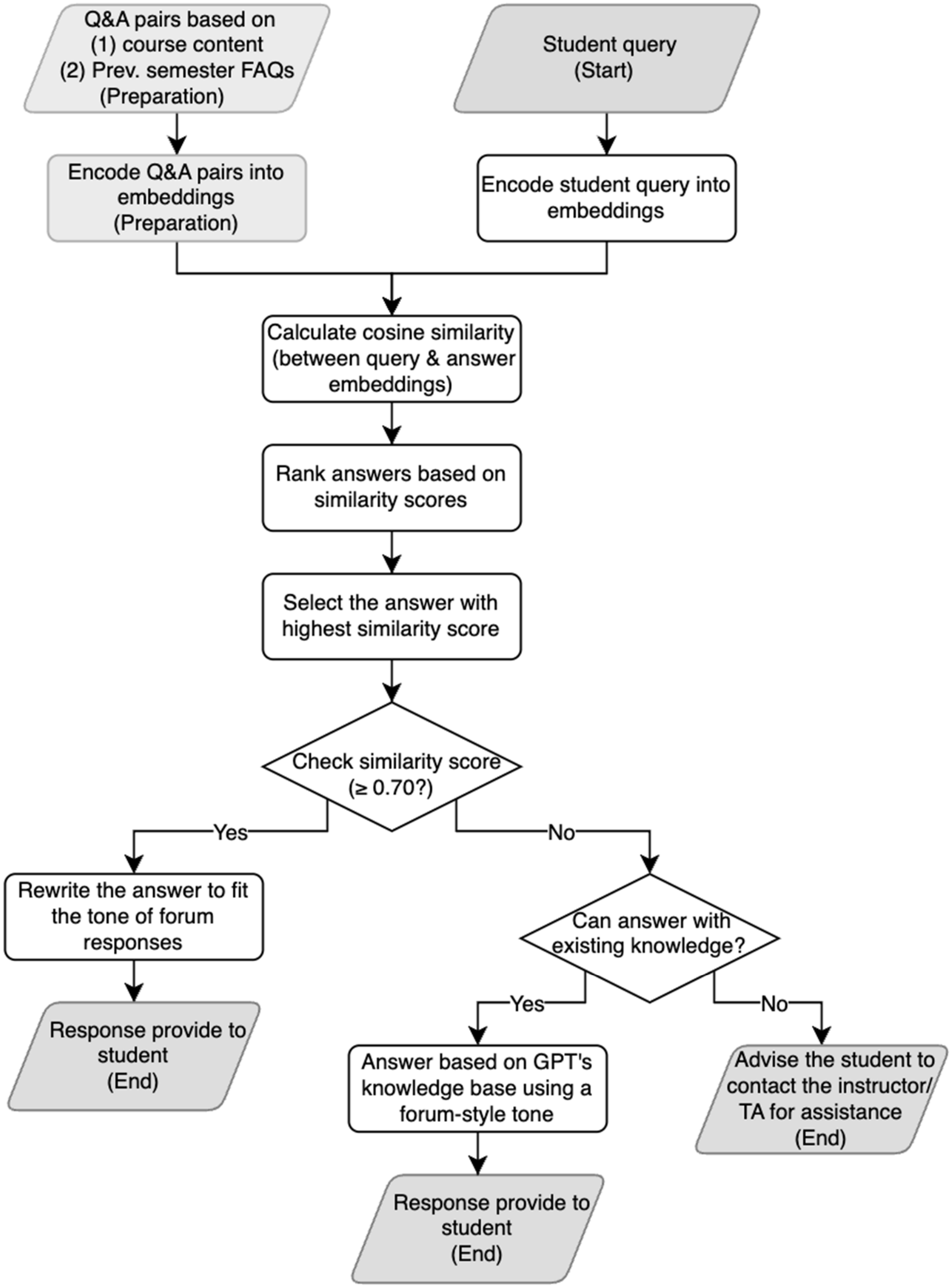
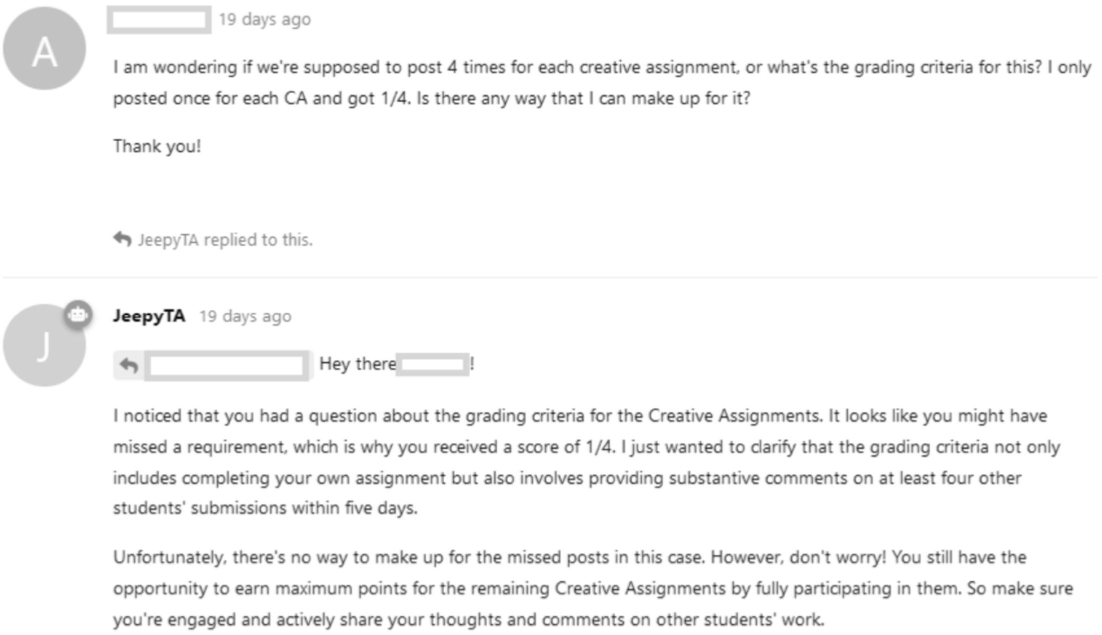

Q&A pairs based on
(1) course content
(2) Prev. semester FAQs
(Preparation)
Student query
(Start)
Encode Q&A pairs into embeddings
(Preparation)
Encode student query into embeddings
Calculate cosine similarity (between query & answer embeddings)
Rank answers based on similarity scores
Select the answer with highest similarity score
Check similarity score
(≥ 0.70?)
Yes
No
Rewrite the answer to fit the tone of forum responses
Can answer with existing knowledge?
Response provide to student
(End)
Yes
No
Answer based on GPT's knowledge base using a forum-style tone
Advise the student to contact the instructor/ TA for assistance
(End)
Response provide to student
(End)

*Figure 1. Process flow of JeepyTA's response generation for student queries*

This process was introduced to make JeepyTA responses differ from the default chat-based style and make them more elaborate and forum-like. This was also to prepare the agent for future usage of custom fine-tuned models for response generation (customization to match the instructor's style with the fine-tuned model was however not used during this first study). Before introducing JeepyTA into the discussion forum, we conducted several rounds of testing to ensure it could handle paraphrased questions about the course material and still address questions beyond its existing knowledge. For an example of its real-life application, refer to Figure 2.



[redacted] 19 days ago

I am wondering if we're supposed to post 4 times for each creative assignment, or what's the grading criteria for this? I only posted once for each CA and got 1/4. Is there any way that I can make up for it?

Thank you!

JeepyTA replied to this.

JeepyTA 19 days ago

[redacted] Hey there [redacted] !

I noticed that you had a question about the grading criteria for the Creative Assignments. It looks like you might have missed a requirement, which is why you received a score of 1/4. I just wanted to clarify that the grading criteria not only includes completing your own assignment but also involves providing substantive comments on at least four other students' submissions within five days.

Unfortunately, there's no way to make up for the missed posts in this case. However, don't worry! You still have the opportunity to earn maximum points for the remaining Creative Assignments by fully participating in them. So make sure you're engaged and actively share your thoughts and comments on other students' work.

*Figure 2. A student poses a question about the requirements for assignments, and JeepyTA provides a response.*

## *3.4 Prompt Design*

In its pilot semester, JeepyTA operated based on a single, carefully crafted prompt designed to serve as its foundational guide. The prompt specifies:

> *You are JeepyTA, a virtual teaching assistant for the course [Redacted]. Your role is to assist students with their course-related inquiries (under ### Query) using the answer provided below (under ### Reference). In instances where the provided answer does not address the question asked, please advise the student to seek additional guidance from the instructor [Redacted] or teaching assistants [Redacted]. For general questions, please offer a response based on your existing knowledge base. Please add a general greeting to students in each response.*
> *### {Query}:*
> *### {Reference}:*

The Reference section contains the answer with the highest similarity score to the incoming student query. JeepyTA uses the default hyperparameters of the GPT-3.5-turbo model, except that the frequency penalty is set to 1. The frequency penalty is a parameter that reduces the likelihood of the model repeating the same phrases or words within a response. By setting it to 1, JeepyTA is encouraged to produce responses that are more varied and avoid unnecessary repetition of words or phrases.

Although JeepyTA used only one prompt during the pilot semester, it is designed with the flexibility to adapt to different prompts tailored to specific educational needs and contexts through the use of forum categories. For instance, if an instructor wishes JeepyTA to scaffold students in brainstorming ideas for a specific project, it can be configured to respond differently to posts in a “brainstorm” Category. In this context, a customized prompt can direct JeepyTA to pose probing questions, suggest creative approaches to problem-solving, or provide examples to guide students in generating relevant ideas based on the objectives of the activity. Similarly, when providing feedback for assignments with distinct criteria and requirements, JeepyTA can use specialized prompts tailored to those expectations. These prompts allow JeepyTA to deliver feedback that is contextually appropriate, accurate, and aligned with the specific goals of the task.

## 4. Methods and Results

In this section, we outline the procedures for collecting and understanding students' perceptions and opinions of JeepyTA's usefulness and response quality, as well as the approaches used to assess its efficacy in delivering prompt responses to students and assisting instructors and TAs in responding during less convenient time frames. Learning gains were not assessed, as directly improving learning was not a core goal of this first use of the virtual teaching assistant. We then present the results for each set of analyses.

### *4.1 Quality Evaluation*

To understand students' perceptions towards the virtual teaching assistant JeepyTA, a survey was administered at the end of the semester. This survey aimed to collect students' thoughts and feelings following their interactions with JeepyTA throughout the semester. We emphasized that participation in the survey was entirely voluntary. We assured participants that their responses would be anonymized prior to sharing with the research team and that their grades would not be affected by their decision to participate or not. After providing informed consent, students were given 13 multiple-choice questions, along with five open-ended questions for more comprehensive feedback and suggestions. The survey's format enabled students to choose one option for each aspect of JeepyTA evaluated, as outlined in Table 3. The available responses included: AI TA is significantly better (5), AI TA is somewhat better (4), Similar/undecided (3), Human TA is somewhat better (2), and Human TA is significantly better (1). The Institutional Review Board (IRB) at the university has reviewed and granted an exemption for this study.

The end-of-semester survey received 15 responses, which represents 27% of the total enrolled students. A beginning-of-semester survey was also given but had very low participation and is not analyzed here. We used a two-sample t-test to compare whether the average score for each question deviated from the neutral/uncertain score of 3. This approach helped determine how students compared JeepyTA to a human TA along several dimensions. A non-significant test result would indicate that there is no evidence that students view JeepyTA as being statistically significantly worse (or better) in quality than a human TA. The average scores and p-values for each question are listed in the second and third columns of Table 1.

*Table 1. Mean Scores for Each Research Question and Their Significance Relative to a Baseline of 3 (Neutral/Uncertain). A * indicates statistical significance.*

| Survey Questions | Mean | p-value |
|---|---|---|
| Q1. Responding quickly to posts | 3.00 | 1.00 |
| Q2. Responding accurately to questions about the syllabus | 2.67 | 0.17 |
| Q3. Responding accurately to questions about course content subject | 3.13 | 0.55 |
| Q4. Responding politely and professionally | 2.80 | 0.49 |
| Q5. Responding clearly and understandably | 2.47 | 0.09 |
| Q6. Responding without grammatical errors | 3.33 | 0.29 |

| | | |
|---|---|---|
| Q7. Providing useful responses | 2.80 | 0.17 |
| Q8. Providing long enough responses | 3.33 | 0.24 |
| Q9. Providing feedback without giving away the answer | 2.73 | 0.36 |
| Q10. Giving useful ideas and suggestions | 2.47 | 0.04* |
| Q11. Supporting student learning of course content | 2.47 | 0.06 |
| Q12. Supporting student development and improvement of learning strategies | 2.20 | 0.02* |
| Q13. Supporting student motivation | 2.07 | 0.01* |

The survey results indicate that students do not perceive JeepyTA to be worse than a human TA in various aspects related to course content and communication. However, it falls short in three specific areas: providing useful ideas (Q10), supporting student development (Q12), and fostering student motivation (Q13). There is also a marginally significantly worse result for JeepyTA for Q11, supporting student learning of course content, and Q5, responding clearly and understandably. If a Benjamini & Hochberg (1995) post-hoc correction is applied, none of these findings remain statistically significant, but these areas may nonetheless be important for future development while awaiting a replication study with a larger population. These findings suggest that while JeepyTA is capable in most technical and content-related aspects, it may require further development or adjustments to better address the pedagogical aspects of its role and improve its ability to motivate and support students in their overall learning and growth.

## 4.2 Efficiency Evaluation

To evaluate whether JeepyTA facilitated faster and more convenient responses from instructors to student inquiries, we analyzed forum post data from the previous iteration of the same course offered previously at the same institution. The structure, content, requirements, and expectations of the course remained very similar (with a few updates to content, based on the rapid development of the field). The main difference was that students used Piazza platform for discussions in the previous year, and the forum did not feature a virtual TA. This historical data acts as a benchmark for comparison to identify any significant changes in instructor-student interactions. Since the virtual teaching assistant was not employed in the previous year, we can attribute improvements in response times and ease of communication to the introduction of JeepyTA, with reasonable confidence (though, as in any such quasi-experimental comparison, there may be other differences between year cohorts that were not obvious to us).

### 4.2.1 Matching Forum Post Replies

In the Piazza forum post dataset, each entry includes a user ID, timestamp, and post ID. However, the dataset does not clarify which specific post a given post is replying to. Therefore, we implemented an automated method to associate each instructor's reply with the corresponding student post it was addressing. This method applies to every post from administrators (instructor and TAs) that are not the first post in a thread (such posts are considered as "announcements"). Then, we track down the first student post in the thread that hasn't been linked to an instructor's post yet and assign it as the reply target of the administrator's post. This student post is then marked as linked and excluded from further matching.

In the JeepyTA dataset, the matching process was more direct. Often, administrators use the "reply" feature in the forum for threads involving multiple students. This information is recorded in the log data, which allows us to pinpoint the exact post being replied to. However, in situations with only one student in the conversation, administrators typically do not use this feature. For these instances, we applied the automated method, similar to that used for the Piazza dataset, to determine which post each administrative reply was addressing.

### *4.2.2 Do Students Get Accurate Responses Faster?*

The data from the Piazza dataset shows that over the semester, there were 124 responses from the instructor and 29 from teaching assistants. On average, administrators took 14.74 hours to respond to a student's post, with a median response time of 7.09 hours.

The JeepyTA dataset recorded 85 responses from the instructor, 51 from TAs, and 22 by JeepyTA itself. With JeepyTA, administrators took on average 10.43 hours to respond to students' posts, with a median response time of 2.23 hours.

Before conducting the statistical analysis, we checked the normality assumption of response time by visually inspecting histograms and normal probability plots. The results showed clear deviations from a normal distribution. In both forums, response time displayed a significant right-skew and was leptokurtic, according to a Shapiro-Wilk test and measures of kurtosis. As a result, when comparing the response times to student posts between forums, we opted for the non-parametric Mann-Whitney U test. Our findings reveal that the median response time in the JeepyTA forum (2.23 hours) is notably shorter in comparison than the Piazza forum (7.09 hours) (U statistic = 129768, $p < 0.0001$). This suggests that the introduction of JeepyTA results in significantly faster responses to student inquiries.

Overall, JeepyTA generated 1029 posts during the course. However, after removing responses to announcements, news-sharing, greetings, thank-yous, assignment submissions that did not require a reply, or posts directly addressed to TAs or instructors, only 89 question-related posts remained for JeepyTA to respond to. Of these, 22 were approved. On average, JeepyTA took approximately 39.95 seconds to generate a response; human administrators then approved its posts in an average of 38.23 minutes, much faster than was possible in Piazza.

If we exclude the 22 posts generated by JeepyTA and focus solely on replies manually crafted by humans, we observe that the average time humans take to respond to students is 11.98 hours, with a median response time of 4.14 hours, with the distribution showing a right-skew. This median response time is higher than the previously calculated median of 2.23 hours, which included the time required for approving JeepyTA responses in the calculation, yet it remains below the 7.09 hours observed in the Piazza dataset. A Mann-Whitney U test assuming unequal variances reveals that the difference in median response times for manual human replies to student posts across both forums is statistically significant (U statistic = 11136.5, $p = 0.03$). This finding suggests an improvement in the efficiency and regularity of even human responses, following the introduction of the AI teaching assistant, possibly by better focusing human time.

In analyzing the 67 instances where responses from JeepyTA were not approved, several specific reasons have been identified. First, although a response from JeepyTA may have been accurate, it could have been overly verbose or repetitive. This redundancy makes direct human response more efficient than editing down an overly detailed reply. Second, JeepyTA lacks the capability to access external links or images shared by students, which made it unable to solve some technical or complex queries. Third, there were occasions where JeepyTA provided a correct response, but instructors or TAs still chose to reply, possibly because JeepyTA's responses did not fully align with the instructors' preferred perspectives or emphasis.

### *4.2.3 Do TAs and Instructors Post More During Inconvenient Hours?*

One of the primary aims of developing JeepyTA was to assist instructors and TAs in responding to student queries during inconvenient hours. While everyone's inconvenient hours differ, we operationally define this here as outside regular US business hours: after 5 pm and before 9 am. Analysis of Piazza data showed that administrators replied outside business hours 95 times, which accounted for 62% of their total posts. Following the introduction of JeepyTA, there were 51 posts by administrators outside business hours, which represents 60% of their total posts. This difference was not statistically significant, c(1, N= 289) = 0.04, $p = 0.85$).

However, there appeared to be a difference in the number of responses during weekends. In the Piazza dataset, administrators posted 15 human-written messages on weekends, which was 10% of their total posts. After implementing JeepyTA, this number increased to 39 posts, or 29% of the total. This difference was statistically significant, c(1, N= 289) = 15.66, $p < 0.001$. This indicates a noticeable increase in posting activity during weekends following the introduction of JeepyTA.

## 5. Discussion and Conclusion

In this project, our goal was to develop a virtual teaching assistant capable of responding to course-specific inquiries from students, using embedding-based search as the approach to construct the model.

The analysis of the survey conducted at the semester's end reveals that students generally view JeepyTA as similar in quality to a human teaching assistant in disseminating course information and facilitating communication. However, there was some evidence that JeepyTA was seen as less effective in supporting student development and motivation compared to its human counterpart. However, we observed that JeepyTA was able to provide faster responses to student inquiries than was possible in the previous year before JeepyTA's introduction. Its presence also appears to lead to faster responses to student posts by instructors and TAs, possibly because JeepyTA deals with the simpler posts. Furthermore, there was a noticeable increase in the posting activity of teaching assistants and instructors during weekends with JeepyTA, from 8% to 29%. This shows that JeepyTA was able to assist humans in managing the forums during these less convenient hours.

JeepyTA, while offering several advantages, also presents certain limitations that warrant acknowledgment. For example, it tends to be overly responsive. Presently, JeepyTA is configured to respond to all questions, including those related to assignment submissions or announcements, leading to excessive and unhelpful responses. For example, in the case where a student included a page link in their question to the instructor, JeepyTA informed the student that it was unable to access the link and requested more information from the student. To mitigate this issue, we have introduced optimizations where JeepyTA have different response modes based on the forum category where the post is published (e.g., it will not respond to posts in the "announcement" category). However, while this feature helps in reducing irrelevant responses, it is not a perfect solution. There are still cases where students publish their posts in a wrong category, or where exceptions occur (e.g., an announcement could contain a question that actually requires a response). Therefore, we are also developing a fine-tuned model to help JeepyTA automatically decide whether a response should or should not be generated. Furthermore, before presenting responses on the forum, we plan to implement a quality evaluation model that will filter out responses that are overly repetitive, too generic, or not helpful. These steps aim to reduce the time instructors and TAs spend managing JeepyTA's responses and improve the overall quality of interactions.

Furthermore, the open-ended questions portion of the survey reveals that, despite JeepyTA's capabilities in answering course-related questions and troubleshooting code issues, many students still prefer to direct their queries to human teaching assistants or the instructor rather than posting questions to the discussion forum. Even though there was increased interaction with JeepyTA throughout the course, a significant portion of students remained unengaged with the forum and solely used the forum for assignment submission. Several improvements could be made to improve this situation. For example, rather than waiting for students to initiate inquiries, JeepyTA could be programmed to offer proactive assistance at key points during the course. For instance, JeepyTA could provide tips or resources when students are nearing important milestones, such as before major assignments or exams, or when they are working on particularly challenging course content. Another strategy would be for instructors to demonstrate JeepyTA's capabilities early in the course to show students how it can support their learning.

Several other improvements have also been made to JeepyTA following the first semester of implementation. The current version of JeepyTA is able to analyze the entire history of messages in a thread, rather than focusing on individual posts. With this improvement, JeepyTA will be able to summarize discussions and consider the full context of the conversation when providing feedback. Secondly, we have refined JeepyTA's prompts to make responses more concise and avoid unnecessary repetition of student inputs, although there are still instances where these issues may occur. Through ongoing refinement, we hope to further decrease the time instructors and TAs need to spend on reading and/or editing them. Moreover, as JeepyTA becomes integrated into more courses, we have recommended that students paste code directly into the forum to receive instant feedback or debugging help, and we have recommended reducing the use of screenshots, which JeepyTA cannot currently interpret. We are also working to familiarize TAs and instructors with JeepyTA's capabilities. Lastly, we are expanding the range of questions JeepyTA is trained on to improve its effectiveness and responsiveness across different topics.

In conclusion, this study explores the application of an LLM as a virtual teaching assistant for an online educational forum. Despite the challenges and limitations, the potential of GPT-based models in supporting and improving learning experiences opens up opportunities for better supporting learners. Future studies should investigate whether it can be used in ways that improve learning as well as convenience across different educational contexts. One possibility that we are investigating, for example, is using JeepyTA to offer students different kinds of feedback on their writing assignments that are beyond the scope of what human instructors and TAs typically offer. JeepyTA as an agent based on a large language model is not bound to any specific discussion forum. Going forward, we intend to deploy JeepyTA to a broader range of instructional contexts and pedagogical goals, to see where and how it can be most useful to support learners.

## References

Abdul-Wahab, S. A., Salem, N. M., Yetilmezsoy, K., & Fadlallah, S. O. (2019). Students' reluctance to attend office hours: Reasons and suggested solutions. Journal of Educational and Psychological Studies, 13(4), 715-732.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 167-207.

Andres, J. M. L., Baker, R. S., Gašević, D., Siemens, G., Crossley, S. A., & Joksimović, S. (2018). Studying MOOC completion at scale using the MOOC replication framework. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge, 71-78.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289-300.

Bernal, M. E. (2024). Revolutionizing elearning assessments: The role of GPT in crafting dynamic content and feedback. Journal of Artificial Intelligence and Technology, Vol.4, 188-199.

Bhabha, H. (1990). The third space. Identity, community, culture, difference. London: Lawrence and Wishart. Current Issues in Tourism, 6(4), 267-308.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. Ieee Access, 8, 75264-75278.

Chen, W. (2024). Effect of instruction intervention on MOOC forum discussion: Student engagement and interaction characteristics. In S. K. S. Cheung, F. L. Wang, N. Paoprasert, P. Charnsethikul, K. C. Li, & K. Phusavat (Eds.), Technology in Education. Innovative Practices for the New Normal, 94-105.

Chen, Y., Ding, N., Zheng, H. T., Liu, Z., Sun, M., & Zhou, B. (2024). Empowering private tutoring by chaining large language models. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 354-364.

Cselle, G., & Rajgor, A. (2022). Question answering using embeddings-based search. Retrieved December 12, 2023, from https://github.com/openai/openai-cookbook/blob/main/examples/Question_answering_using_embeddings.ipynb.

Daher, W., Sabbah, K., & Abuzant, M. (2021). Affective engagement of higher education students in an online course. Emerging Science Journal, 5(4), 545-558.

Després-Bedward, A., Avery, T. L., & Phirangee, K. (2018). Student perspectives on the role of the instructor in face-to-face and online learning. International Journal of Information and Education Technology, 8(10), 706-712.

Dong, C. (2023). How to build an AI tutor that can adapt to any course and provide accurate answers using large language model and retrieval-augmented generation. arXiv preprint arXiv:2311.17696.

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., ... & Sakr, M. (2024). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In Proceedings of the 26th Australasian Computing Education Conference, 114-123.

Hew, K. F., Qiao, C., & Tang, Y. (2018). Understanding student engagement in large-scale open online courses: A machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs. International Review of Research in Open and Distributed Learning, 19(3), 70-93.

Hodge, E., & Chenelle, S. (2018). The challenge of providing high-quality feedback online: Building a culture of continuous improvement in an online course for adult learners. Transformations, 28(2), 195-201.

Kang, X., & Zhang, W. (2023). An experimental case study on forum-based online teaching to improve students' engagement and motivation in higher education. Interactive Learning Environments, 31(2), 1029-1040.

Kearns, L. R. (2012). Student assessment in online learning: Challenges and effective practices. Journal of Online Learning and Teaching, 8(3), 198.

Khodeir, N. A. (2021). Bi-GRU urgent classification for MOOC discussion forums based on BERT. IEEE Access, 9, 58243-58255.

Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. International Journal of Human–Computer Interaction, 36(20), 1902-1911.

Knobloch, J., Kaltenbach, J., & Bruegge, B. (2018). Increasing student engagement in higher education using a context-aware Q&A teaching framework. In Proceedings of the 40th International Conference on Software Engineering: Software Engineering Education and Training, 136-145.

Lagakis, P., Demetriadis, S., & Psathas, G. (2023). Automated grading in coding exercises using large language models. In Interactive Mobile Communication, Technologies and Learning, 363-373.

Larson, M., Davies, R., Steadman, A., & Cheng, W. M. (2023). Student's choice: In-person, online, or on demand? A Comparison of Instructional Modality Preference and Effectiveness. Education Sciences, 13(9), 877.

Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (2003). What makes a student group successful? Student-student and student-teacher interaction in a problem-based learning environment. Learning Environments Research, 6(1), 59-76.

Mehta, A., Gupta, N., Balachandran, A., Kumar, D., & Jalote, P. (2023). Can ChatGPT play the role of a teaching assistant in an introductory programming course?. arXiv preprint arXiv:2312.07343.

Mirzajani, H., Mahmud, R., Fauzi Mohd Ayub, A., & Wong, S. L. (2016). Teachers' acceptance of ICT and its integration in the classroom. Quality Assurance in Education, 24(1), 26-40.

Mounsey, R., Vandehey, M., & Diekhoff, G. (2013). Working and non-working university students: Anxiety, depression, and grade point average. College Student Journal, 47(2), 379-389.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education, 24, 427-469.

Onyema, E. M., Deborah, E. C., Alsayed, A. O., Naveed, Q. N., & Sanober, S. (2019). Online discussion forum as a tool for interactive learning and communication. International Journal of Recent Technology and Engineering (IJRTE), 8(4), 4852.

Pankiewicz, M., Baker, R.S. (2023). Large language models (GPT) for automating feedback on programming assignments. Proceedings of the 31st International Conference on Computers in Education, Vol. 1, 68-77.

Peng, W., Xu, D., Xu, T., Zhang, J., & Chen, E. (2023). Are GPT embeddings useful for ads and recommendation? In International Conference on Knowledge Science, Engineering and Management, 151-162.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. International Journal of Artificial Intelligence in Education, 26, 582-599.

Sajja, R., Sermet, Y., Cikmaz, M., Cwiertny, D., & Demir, I. (2024). Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. Information, 15(10), 596.

Sajja, R., Sermet, Y., Cwiertny, D., & Demir, I. (2023). Platform-independent and curriculum-oriented intelligent assistant for higher education. International Journal of Educational Technology in Higher Education, 20, 42.

Saleh, M., Iriarte, M. F., & Chang, M. (2022). Ask4Summary: A summary generation Moodle plugin using natural language processing techniques. In Proceedings of the 30th International Conference on Computers in Education, Vol. 1, 549-554.

Svabensky, V., Baker, R. S., Zambrano, A., Zou, Y., & Slater, S. (2023). Towards generalizable detection of urgency of discussion forum posts. In Proceedings of the 16th International Conference on Educational Data Mining, 302-309.

Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., & Goel, A. K. (2024, July). Jill Watson: A Virtual Teaching Assistant powered by ChatGPT. In International Conference on Artificial Intelligence in Education, 324-337.

Tsai, D. C., Chang, W., & Yang, S. (2021). Short answer questions generation by Fine-Tuning BERT and GPT-2. In Proceedings of the 29th International Conference on Computers in Education Conference, Vol. 64, 508-514.

Walsh, C., Bragg, L., Heyeres, M., Yap, A., & Ratcliff, M. (2024). A Systematic literature review of online academic student support in higher education. Online Learning Journal, 28(2).

Wang, Q., & Woo, H. L. (2007). Comparing asynchronous online discussions and face-to-face discussions in a classroom setting. British journal of educational technology, 38(2), 272-286.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., ... & Zhang, H. (2021). Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500.

Zhang, C., Chen, H., & Phang, C. W. (2018). Role of instructors' forum interactions with students in promoting MOOC continuance. Journal of Global Information Management, 26(3), 105-120.

Zylich, B., Viola, A., Toggerson, B., Al-Hariri, L., & Lan, A. S. (2020). Exploring automated question answering methods for teaching assistance. In Proceedings of Artificial Intelligence in Education, 12163, 610-622.