QuickAns: A Virtual Teaching Assistant

ISHIKA AGARWAL, University of Illinois, Urbana-Champaign
SHRADHA SEHGAL, University of Illinois, Urbana-Champaign
VARUN GOYAL, University of Illinois, Urbana-Champaign
PRATHAMESH SONAWANE, University of Illinois, Urbana-Champaign

QuickAns is a virtual teaching assistant designed to assist course staff who use Campuswire as their Q&A platform. It reads Campuswire posts from digest emails and sends a potential answer to the course staff. At this stage, the course staff can review the answer for any logistical issues and answer a student's question in a matter of minutes.

Additional Key Words and Phrases: Teaching Aid, LLMs, Education, Vector Database

ACM Reference Format:
Ishika Agarwal, Shradha Sehgal, Varun Goyal, and Prathamesh Sonawane. 2023. QuickAns: A Virtual Teaching Assistant. In *The Third International Conference on Artificial Intelligence and Machine Learning Systems (AIMLSystems 2023), October 25–28, 2023, Bangalore, India.* ACM, New York, NY, USA, Article 111, 3 pages. https://doi.org/10.1145/3639856.3639904

1 INTRODUCTION

Campuswire is an online Q&A platform for courses. As the course staff becomes busier with course responsibilities, it becomes difficult to promptly answer Campuswire questions. Additionally, many questions tend to be repetitive or fundamental. Our application software aims to reduce the time and effort spent on answering Campuswire questions. This tool relieves them from typing answers to basic questions, allowing them to focus their attention and expertise on more challenging tasks.

The functionality it offers includes automated and timely response suggestions for questions on Campuswire, made available through email. Specifically, our application reads emails received from Campuswire, generates a potential response, and communicates the response via email to the course instructors. They can subsequently review the response for any errors or omissions before posting it. This streamlined process saves the course staff valuable time that would have been spent crafting and refining detailed replies.

This report outlines the pipeline in section 2, provides examples of how to use QuickAns in section 3, and finally, presents an evaluation in section 4.

2 METHODOLOGY

We have built this application on top of the Campuswire platform. We create a simulated user that subscribes to email alerts to all the posts made on the platform, enabling access to each question.

The email reading and scraping functionality is implemented using the Python Microsoft Graph API. A question is inputted into our question-answering module, which generates replies and emails them to the course staff. We utilize the ChatGPT model via API to generate the responses to the questions as it is relatively fast, possesses language understanding, and has been instruction tuned. Language understanding enables it to address improperly framed questions. Instruction Tuning has been qualitatively proven to outperform non-tuned models in Q&A tasks. Overall, our approach involves emailing these generated replies to the course staff.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

AIMLSystems 2023, October 25–28, 2023, Bangalore, India

© 2023 Copyright held by the owner/author(s).

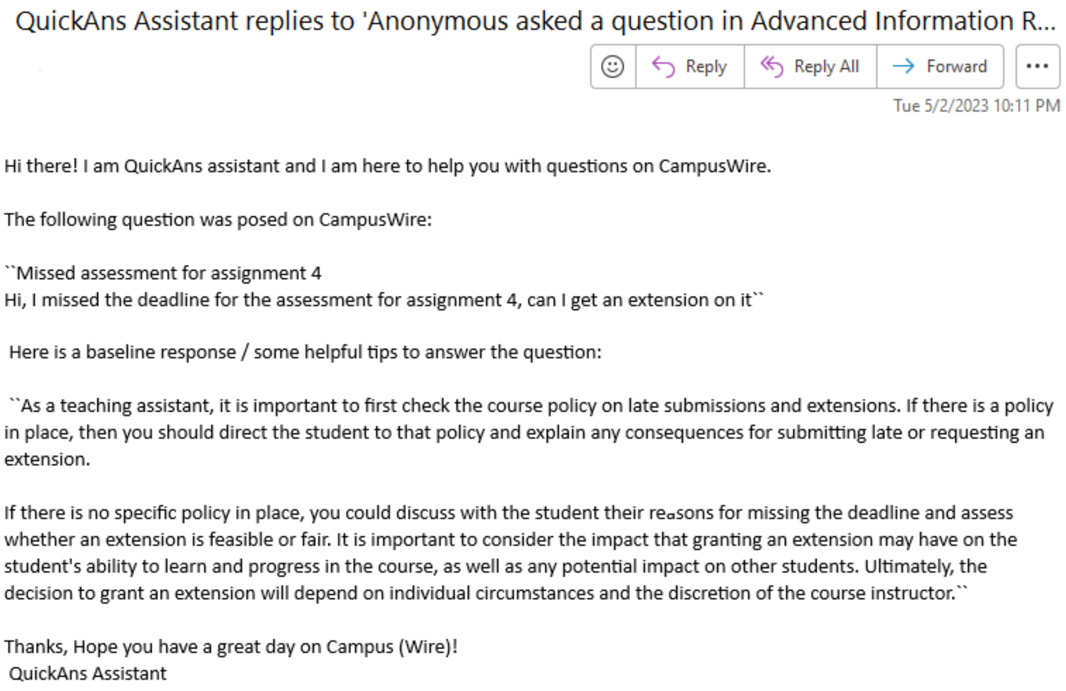
ACM ISBN 979-8-4007-1649-2/23/10
https://doi.org/10.1145/3639856.3639904

2.1 Question-Answering Module

The question-answering module utilizes ChatGPT as the core language model. To enhance its ability to answer course-related questions, we provide the corresponding textbook to the LM. As an example, we have developed this for our Information Retrieval class, so we supplied the book "Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining" [1].

We employ a search-based method. We divide the textbook into chunks (5-7 sentences, with overlap) and created= embeddings using the ‘text-embedding-ada-002‘ model. Then, we create embeddings from student questions, and use a similarity function to determine the closest textbook paragraph to the question. From there, we input the question and the related paragraph into the GPT-based model to retrieve an answer.

Fine-tuning versus searching. A recent trend involves fine-tuning a language model on specific text data to enable it to understand a particular domain. However, fine-tuning may not be the best technique. Language models should be fine-tuned on tasks, not data. Embedding search-based methods are more effective for focusing on context, which is the case in QuickAns. Furthermore, we opted to search chunks of text (rather than the entire textbook, or chapter) due to the token limit. Even without the token limit, the output is expected to be largely similar.



QuickAns Assistant replies to 'Anonymous asked a question in Advanced Information R...

Tue 5/2/2023 10:11 PM

Hi there! I am QuickAns assistant and I am here to help you with questions on CampusWire.

The following question was posed on CampusWire:

``Missed assessment for assignment 4
Hi, I missed the deadline for the assessment for assignment 4, can I get an extension on it``

Here is a baseline response / some helpful tips to answer the question:

``As a teaching assistant, it is important to first check the course policy on late submissions and extensions. If there is a policy in place, then you should direct the student to that policy and explain any consequences for submitting late or requesting an extension.

If there is no specific policy in place, you could discuss with the student their reasons for missing the deadline and assess whether an extension is feasible or fair. It is important to consider the impact that granting an extension may have on the student's ability to learn and progress in the course, as well as any potential impact on other students. Ultimately, the decision to grant an extension will depend on individual circumstances and the discretion of the course instructor.``

Thanks, Hope you have a great day on Campus (Wire)!
QuickAns Assistant

Fig. 1. A sample email that a TA would receive through the QuickAns pipeline.

3 USAGE

We have made our code available on GitHub. To start the service, run python3 quickAns.py. We use the Microsoft Graph API to read and scrape (only) Campuswire emails. After executing the script, a user will go through a series of steps to grant access to their emails (they need to enable all email notifications on Campuswire). Once signed in, QuickAns will scrape emails and display the top k results in the terminal. It will select the top post, send it to the QuickAns answering module, and display the answer in the terminal. The exchange will also be emailed to the designated email account(s). A sample email that a TA would receive at the end of this pipeline is show in Figure 1. Finally, this service should be hosted on a server to continuously receive and answer queries. The server will scrape emails every 5 minutes for new questions, and we have implemented functionality to avoid answering the same question multiple times. Due to the page limit, we provide an informal video created by our team to showcase our work.

4 QUALITATIVE EVALUATION

As a quick evaluation, we posed the same question to QuickAns' base model and GPT-3.5: "In CPLSA, if each view has only one coverage, why do we have two steps (step1: choose a view, step2: choose a coverage)?" CPLSA stands for Contextual Probabilistic Latent Semantic Analysis, an acronym created by Professor Zhai (a contributor to our textbook). Figure 2 illustrates the difference in their responses.

> As an AI language model, I am not familiar with the specific context of CPLSA. However, in general, it is possible that a view may have multiple coverages, or that the choice of view may influence the available coverages. Additionally, the two steps may be necessary for organizational or procedural reasons, such as ensuring that all necessary information is collected and recorded in a systematic manner. It may be helpful to consult the specific guidelines or documentation for CPLSA to better understand the reasoning behind the two-step process.

> You are correct that each view only has one coverage in CPLSA. However, the reason we have two steps (step 1: choose a view; step 2: choose a coverage) is because each word can potentially be generated using a different view and a different topic coverage distribution depending on the contexts chosen to direct the generation process. In other words, the context that determines the choice of view of a topic can be different from the context chosen to decide the topic coverage. This allows for the discovery of different variations of the same topic in different contexts due to the dependency of a view of topics on context and different topic coverages in different contexts due to the dependency of topic coverage on context.

Fig. 2. Asking questions about CPLSA.

Analysis. GPT-3.5 responds by indicated unfamiliarity with CPLSA and provides a vague answer while requesting more information. In contrast, QuickAns accurately comprehends concept and delivers a correct response. This is something that even sophisticated prompt engineering ("In the context of Information Retrieval, ...") cannot solve.

5 CONCLUSION

In this report, we have outlined the pipeline and fine-grained details of our virtual teaching assistant, QuickAns. Through a brief qualitative analysis, we have demonstrated the benefits of our method compared to the baseline.

REFERENCES

[1] ChengXiang Zhai and Sean Massung. 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Vol. 12. Association for Computing Machinery and Morgan amp; Claypool.