Transforming Student Support with AI: A Retrieval-Based Generation Framework for Personalized Support and Faculty Customization

by

Shukang Wang

B.A., The University of British Columbia 2023

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The College of Graduate Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA

(Okanagan)

April 2025

The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, a thesis entitled:

Transforming Student Support with AI: A Retrieval-Based Generation Framework for Personalized Support and Faculty Customization

submitted by Shukang Wang in partial fulfilment of the requirements of the degree of Master of Science.

Dr. Ramon Lawrence, Irving K. Barber Faculty of Science
Supervisor

Dr. Patricia Lasserre, Irving K. Barber Faculty of Science
Supervisory Committee Member

Dr. Bowen Hui, Irving K. Barber Faculty of Science
Supervisory Committee Member

Dr. Chen Feng, School of Engineering
University Examiner

Abstract

The rapid evolution of Natural Language Processing (NLP) has positioned Large Language Models (LLMs) like ChatGPT as transformative tools across various sectors, including education. These models offer significant potential for enhancing learning experiences through personalized assistance, yet their propensity to generate incorrect, biased, or unhelpful responses poses critical challenges for their deployment in educational contexts. The usability of AI as tutors or assistants demands better human-AI design, particularly through personalization and customization to meet diverse educational needs.

The first contribution of this thesis focuses on optimizing the AI agent for education, primarily through the design of a Retrieval Augmented Generation (RAG) pipeline as the agent's core tool. By refining its components, this research aims to ensure that AI-driven educational assistants provide more accurate and contextually relevant responses.

In addition to these technical enhancements, this research introduces a system that centralizes help-seeking channels with AI to facilitate the seamless integration of the optimized RAG framework into existing educational platforms. This resulted in the integrated HelpMe system, ensuring that the enhanced RAG system's benefits are accessible to a broader range of users without requiring extensive technical expertise.

Through both quantitative and qualitative data gathered during the deployment both the queuing system and integrated system, we find significant benefits in the integrated HelpMe system as an assistive tool for educators and students. We also identify future directions and challenges in human-AI interactions within educational contexts.

Lay Summary

This thesis explores how Artificial Intelligence, especially Large Language Models (LLMs), can be tailored for educational use. We introduce a system called ChatEd that retrieves course-specific materials and uses them to generate accurate, relevant responses to student questions. Unlike conventional chatbots, ChatEd does not require extensive training; instructors simply provide their slides, notes, and other documents. This research also presents HelpMe, an integrated platform that centralizes help requests, whether through office hours, asynchronous questions, or the new AI chatbot, ChatEd. By unifying all support channels, the system fosters more consistent, effective interactions between students and teaching staff. Quantitative evaluations demonstrate that grounding LLM answers in local course content substantially improves reliability and user study reveals high potential of the integrated system. Ultimately, our framework shows how institutions can deploy AI tools that adapt to changing course content and help instructors manage courses with ease.

Preface

The study in this thesis was conducted with the approval of the UBC Okanagan Behavioural Research Ethics Board under the certificate number H22-03323.

Published paper usage

Portions of this thesis have been previously published. Specifically:

1. **K. Wang** and R. Lawrence, "Quantitative Evaluation of using Large Language Models and Retrieval-Augmented Generation in Computer Science Education," *Proceedings of the 56th ACM Technical Symposium on Computer Science Education*, 2025. (*My contributions: Conceptualizing the study, designing the experiments, analyzing the results, paper writing.*)

2. **K. Wang** and R. Lawrence, "HelpMe: Student Help Seeking using Office Hours and Email," *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 2024, pp. 1388–1394. (*My contributions: Creating the tool and conducting data collection, performing analysis, and writing the paper.*)

3. **K. Wang**, J. Ramos, and R. Lawrence, "ChatEd: A Chatbot Leveraging ChatGPT for an Enhanced Learning Experience in Higher Education," *International Academy of Technology, Education and Development*, 2024. (*My contributions: Designing and co-implementing the chatbot framework, conducting the evaluation, and preparing the manuscript.*)

4. **K. Wang**, S. Akins, A. Mohammed, and R. Lawrence, "Student Mastery or AI Deception? Analyzing ChatGPT's Assessment Proficiency and Evaluating Detection Strategies," *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2023, pp. 1615–1621. (*My contributions: Formulating research*

questions, performing data analysis, and leading manuscript preparation.)

5. **K. Wang** and R. Lawrence, "Using Assignment Incentives to Reduce Student Procrastination and Encourage Code Review Interactions," *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2023, pp. 1628–1633. (*My contributions: Developing the experiments, collecting data, analyzing data, and writing the publication.*)

AI usage

Generative AI tools were utilized in the preparation of this thesis in the following ways:

- Assisting in brainstorming and organizing outlines, as well as verifying the logical structure of the thesis,
- Refining sentence structures and improving clarity in written sections,
- Generating code to produce certain graphs and visualizations,

These tools were used to support the research and presentation process; however, all core ideas, analyses, and conclusions remain the author's own.

Table of Contents

Abstract . . . . **iii**

Lay Summary . . . . **iv**

Preface . . . . **v**
Published paper usage v
AI usage vi

Table of Contents . . . . **vii**

List of Tables . . . . **x**

List of Figures . . . . **xi**

Acknowledgements . . . . **xii**

Dedication . . . . **xiii**

Glossary . . . . **xiv**

Chapter 1: Introduction **1**
1.1 Thesis Organization 2

Chapter 2: Background **4**
2.1 Student Help Seeking 4
2.2 Student Support Systems 5
2.3 Large Language Models 5
2.4 Retrieval-Augmented Generation 6
2.5 Applications of AI in Education 6
2.6 Evaluating LLMs 8
2.6.1 Evaluating RAG systems 8

Chapter 3: HelpMe Queueing System 9
3.1 Implementation 9
3.2 Evaluation Methodology 10
3.2.1 Data Sources and Participants 10
3.2.2 Procedure 11
3.2.3 Survey Instrument 11
3.3 Results 12

Chapter 4: ChatEd and Evaluations 17
4.1 Implementation of ChatEd 17
4.1.1 Interface Design 17
4.1.2 ChatEd's RAG Architecture 17
4.2 Evaluation Methodology 19
4.2.1 ChatEd Evaluation 19
4.2.2 Evaluating AI and RAG Quantitatively 22
4.2.3 RAG Optimizations and Human Feedback 24
4.3 Results 25
4.3.1 ChatEd Evaluation Results 25
4.3.2 Discussion of ChatEd Manual Evaluation 29
4.3.3 Quantitative RAG Evaluation Results 29

Chapter 5: HelpMe Integrated Platform 34
5.1 Implementation 34
5.1.1 Integrated System (Home Page) 34
5.1.2 Chatbot Integration 34
5.1.3 Anytime Question Hub 36
5.2 Evaluation Methodology 36
5.2.1 Survey for Integrated System 38
5.3 Results 39
5.3.1 Chatbot Preference 39
5.3.2 Anytime Question Preference 39
5.3.3 Illustrative Examples 40
5.3.4 Student Survey Findings 42
5.3.5 Preliminary Staff Feedback 42

Chapter 6: Conclusion and Future Directions 44
6.1 Conclusion 44
6.1.1 Key Takeaways 45
6.2 Future Directions 45

6.2.1 Longitudinal Studies on Student Adoption and Learning Outcomes . . . 46
6.2.2 Adaptive AI and Personalized Learning . . . 46
6.2.3 Fine-Tuning vs. Advanced Prompting . . . 46
6.2.4 Evaluating AI Transparency and Trust in Student-Facing Systems . . . 47
6.2.5 Scaling AI-Driven Assistance to Large Academic Institutions . . . 47
6.2.6 Final Thoughts . . . 48

Bibliography . . . 49

Appendices . . . 57
A Consent Forms . . . 58
B Surveys . . . 61
B.1 HelpMe Survey (2023 Spring) . . . 61
B.2 HelpMe Survey (2024 Fall) . . . 64

List of Tables

Table 3.1 Reasons Detering Students from Office Hours 12

Table 3.2 HelpMe Queue System Satisfaction (1=Strong Disagree, 5=Strong Agree) 15

Table 3.3 Preferred Help-Seeking Methods: Before vs. After . . . 15

Table 4.1 Average Scores Given by Evaluators (Q&A Test) . . . 26

Table 4.2 Context-Aware Conversation Scores 27

Table 4.3 Example of a misunderstood question about the HelpMe queue rather than database tickets. 27

Table 4.4 Example of leveraging course knowledge to provide a context-specific answer. 27

Table 4.5 Straightforward SQL question receiving a concise, correct system answer. 29

Table 4.6 Unhelpful or misleading response (no mention of official protocols). 29

Table 4.7 Comparison of Non-RAG Systems on General Questions 30

Table 4.8 Comparison of RAG Systems (Light Guardrails). . . . 31

Table 4.9 Context-Specific Questions with RAG (All Models). . 31

Table 4.10 Average Model Scores Across Different Courses 31

Table 4.11 Comparison of AI-Trimming vs. Traditional Chunking (llama3:70b). 32

Table 4.12 Using Instructor-Verified “Q&A” as Part of the RAG Database (Llama3:70b). 32

List of Figures

Figure 3.1 Student View of the HelpMe System 9
Figure 3.2 Instructor/Professor View of the HelpMe System . . . 10
Figure 3.3 Help Session Question Categorization 12
Figure 3.4 Email Question Categorization 13
Figure 3.5 Session Breakdown by Student 13
Figure 3.6 Distribution of Help Session Durations 14
Figure 3.7 Wait Time Distribution 14
Figure 3.8 Final Exam Performance vs. Help Session Usage . . . 16

Figure 4.1 Chatbot User Interface and Example Conversation [55] 18
Figure 4.2 Instructor Interface in ChatEd [55] 19
Figure 4.3 ChatEd Chatbot Architecture [55] 20
Figure 4.4 Distribution of Question Categories 23
Figure 4.5 Example of AI Edit 25
Figure 4.6 Example of Follow-up Questions in a Conversation . . 28

Figure 5.1 Integrated HelpMe System Homepage 35
Figure 5.2 Course Feature Settings 35
Figure 5.3 Anytime Question Ticket (Student View) 36
Figure 5.4 Anytime Question Instructor View 37
Figure 5.5 Instructor Editing an AI Response 38
Figure 5.6 Chatbot Usage Frequency 39
Figure 5.7 Anytime Question Hub Usage Frequency 40
Figure 5.8 Preference After Using the Tools 41

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Dr. Ramon Lawrence, whose unwavering support, insightful guidance, and boundless patience have made this journey possible. Dr. Lawrence not only introduced me to the fascinating world of research but also inspired me to truly fall in love with it. I am profoundly grateful for his mentorship and encouragement throughout this process.

I would also like to extend my heartfelt thanks to my friend, Ramos, for his invaluable assistance in laying the groundwork for the chatbot - a line of work that continues to inspire and shape my academic pursuits today.

My sincere thanks to Dr. Bowen Hui and Dr. Patricia Lasserre for serving on my supervisory committee. Their thoughtful feedback, expertise, and encouragement have been instrumental in shaping this thesis.

Lastly, I am deeply appreciative of the support from my family, friends, and colleagues, who have been my source of strength and motivation throughout this journey.

Dedication

To Mom and Dad, for your unconditional love, sacrifices, and belief in me.

Glossary

- **AI** - A general term referring to Artificial Intelligence. In this thesis, “AI” may be used interchangeably with “LLM” or “chatbot” when discussing automated question-answering and reasoning systems, although it often indicates a broader set of computational intelligence techniques.

- **Anytime Question Hub** - An asynchronous question submission feature that allows students to post inquiries outside of scheduled office hours. An AI-based module (and, if necessary, human instructors) can respond, enabling flexible support without requiring real-time participation by teaching staff.

- **ChatEd** - A retrieval-augmented generation (RAG) chatbot system developed to answer student questions with course-specific context. By integrating course materials directly into the conversation prompts, ChatEd delivers answers that are both relevant and authoritative, eliminating the need for specialized model training on Q&A pairs.

- **HelpMe Queue System** - An office hour queue system.

- **Integrated HelpMe System** - An integrated help-seeking platform designed to unify various channels of student support, such as office hours (queues), email requests, AI chatbots, and asynchronous question boards. The HelpMe system provides a single, streamlined interface where students can seek assistance, track responses, and engage with both AI and human help resources.

- **Large Language Model, LLM** - A deep neural network, often transformer-based, that is trained on massive text corpora to learn statistical patterns of language. Examples include GPT-3.5, GPT-4, and open-source models like LLaMA. LLMs can generate human-like text, answer questions, or perform reasoning tasks.

– **Retrieval-Augmented Generation, RAG** - A technique in which an AI system retrieves relevant documents or text “chunks” from a specialized datastore (e.g., vector database) and feeds them into the prompt of a large language model. This ensures that the AI’s generated responses are grounded in the retrieved material, reducing hallucinations and improving domain specificity.

– **Vector Database** - A specialized data store that indexes embeddings of text snippets or documents. In RAG pipelines, it efficiently returns text chunks most relevant to a user’s query, based on cosine similarity or other distance metrics in a high-dimensional embedding space.

– **Guardrails** - Rules or instructions provided to the Large Language Model (e.g., “If unsure, say ‘I don’t know’”) that constrain its responses. Guardrails aim to minimize hallucinations or irrelevant answers by enforcing stricter generation boundaries.

– **Prompt Engineering** - The process of crafting or structuring the textual input given to an LLM so that it produces the desired outputs. In this thesis, prompt engineering often involves specifying instructions on style, domain constraints, or referencing the retrieved chunks from course documents.

– **Office Hours / Lab Queue** - A synchronous channel within the HelpMe system through which students line up virtually (or physically) for real-time assistance from TAs or instructors. The queue system is integrated alongside AI-driven features, offering a one-stop platform for all help-seeking behaviors.

Chapter 1

Introduction

Over the past decade, higher education has evolved to include diverse learning formats, ranging from large on-campus lectures to fully online courses and blended models. In parallel, rapid strides in Natural Language Processing (NLP) have yielded sophisticated Large Language Models (LLMs) capable of generating remarkably fluent text. These technologies offer exciting opportunities to improve student support and enhance learning through automated assistance or instant feedback. However, practical deployment of LLMs in education introduces several critical challenges: the potential for misinformation and biased outputs [26, 28, 38], as well as a lack of domain-specific expertise essential to addressing nuanced course materials and policies.

Adapting LLMs for course-specific needs is non-trivial. Simply scaling a model with billions of parameters does not guarantee accurate, context-aware answers for a given syllabus. Key obstacles include ensuring that responses reference trustworthy, up-to-date materials and that instructors retain oversight of the AI's outputs. This thesis addresses these limitations by introducing two primary contributions:

1. ***ChatEd*: A Retrieval-Augmented Generation (RAG) Framework**: is an approach that couples LLMs with a retrieval-based system for locating course-relevant documents. By grounding AI-generated answers in instructor-provided materials, ChatEd delivers more accurate, contextually relevant assistance to learners, while mitigating the model's tendency to hallucinate or drift off-topic.

2. **The *HelpMe* Integrated Help-Seeking Platform**: Recognizing that technology is only effective if it is adopted, the work develops a centralized platform, HelpMe, that unifies office-hour queues, an Anytime Question Hub, and an AI chatbot. All the AI components are backed by infrastructure of ChatEd. This reduces barriers for students seeking help, fosters more efficient monitoring by faculty, and offers novel pathways for collecting and refining AI-assisted interactions.

ChatEd incorporates a retrieval-based workflow that retrieves most relevant “chunks” of information—such as PowerPoint slides, lab instructions, or policy documents—before generating a final response via an LLM. This ensures that answers remain rooted in factual course content, ensuring both accuracy and relevancy. Additionally, instructors can moderate or validate ChatEd’s outputs by updating the knowledge base.

Although the potential of ChatEd’s architecture is apparent, systematic benchmarking remains relatively unexplored in educational contexts. Institutions deciding whether to deploy AI assistants have to grapple with questions of cost-effectiveness, data security, and the viability of open-source versus commercial models. These considerations are particularly relevant when handling sensitive information about student performance or course materials. To fill this gap, my research conducts extensive evaluations of multiple LLMs and RAG pipelines, measuring both their performance on course-related queries and the associated infrastructural costs. This provides a holistic perspective on how to effectively adopt, maintain, and scale AI-driven educational tools.

In parallel is a presentation of the integrated HelpMe system to address pragmatic issues of user adoption. HelpMe integrates existing help channels, like office hours and one-on-one asynchronous questions often sent by email, into a single interface. This not only streamlines student inquiries but also allows faculty to observe and curate the AI’s responses without incurring additional overhead.

The architectures and tools devised in this research aim to reduce friction in the help-seeking process, improve the accuracy and relevance of AI assistance, and enhance the overall learning experience. The following chapters provide background on related work, detail the system design and methodology, present comprehensive evaluation results, and conclude with future directions for refining and extending AI-driven educational support.

1.1 Thesis Organization

This thesis is organized as follows:

1. **Chapter 1: Introduction** — Presents the context, challenges, and objectives of the research.

2. **Chapter 2: Literature Review and Background** — Explores previous literature and provides background on related fields.

3. **Chapter 3: HelpMe Queueing System** — Details the implementation, evaluation methodology, and results for the HelpMe synchronous queue-based system.

4. **Chapter 4: ChatEd and Evaluations** — Describes the implementation of the ChatEd RAG chatbot, the methodologies for its evaluation (including broader RAG experiments), and presents corresponding results.

5. **Chapter 5: HelpMe Integrated Platform** — Presents the unified platform combining HelpMe and ChatEd, its implementation, evaluation methodology, and results.

6. **Chapter 6: Conclusion and Future Work** — Summarizes the key findings, discusses their implications, and outlines directions for future research.

Chapter 2

Background

2.1 Student Help Seeking

Help sessions face significant challenges and are struggling to stay relevant, especially in the era of generative AI where answers are readily available. Research has shown that students tend to use office hours only for specific questions such as troubleshooting [43]. There is also confusion among students over communication with professors outside of academic settings [45]. An analysis of student help seeking [27] shows a strong majority of interactions happen within three days of a deadline with students primarily focused on implementation rather than planning and understanding. Getting students focused on learning rather than deadlines is a challenge.

In-person office hours are becoming outdated in the modern technological context. Smith *et al.* [45] found that the most common reason of avoiding office hours is inconvenience with students asking questions such as "Is it worth it to commute 30 minutes to ask a 5 minute question?" In person hours also poses problems for commuter and minority students, who might more trouble attending office hours [31, 36]. Thus, many instructors have found that online office hours are valuable. Early virtual office hours research has proven the effectiveness of online office hours such as the virtualization of CS50 office hours at Harvard [36]. However, the same research also noted that online office hours do not solve all problems. Students often claim long wait times, lack of physical interactions, and confusion over online queues [36]. Email and other asynchronous communication are still often the preferred ways of communication [31]. Virtual office hours only solve some of the student engagement issues with physical office hours.

There are mixed results on whether help seeking including participation in office hours and Q&A forums has a negative [46], positive [12, 20, 51], or neutral effect [14, 15] on performance. Factors include assignment difficulty, student prior knowledge, and the type of help seeking [8, 12].

2.2 Student Support Systems

Office hour queuing interfacing has been used to improve the efficiency and visibility of help sessions such as the CS50 Queue [35], My Digital Hand [44], and Queue@Illinois [24, 39]. An open source system [42] developed at Northeastern University was adapted to develop the HelpMe system used for this research. Many universities have deployed queuing technology.

The key advantages of online queuing systems are visibility for instructors and students, management of questions and student ordering, and data reporting on wait and help session times. These advantages greatly outweigh the extra time required for students to enter their questions in the system. Prior research has highlighted issues such as long wait times as discouraging student engagement. There is limited understanding on how to most effectively deploy these systems in practice, and if the systems change student behaviour on help seeking [45]. Students often still have a high volume of questions over email and reducing the email volume is often beneficial [23]. There has been limited analysis on the types of questions asked over email [21], and no studies have analyzed both office hours and email interactions.

2.3 Large Language Models

Large Language Models (LLMs) are a class of machine learning models that leverage transformers as their backbone architecture. The introduction of transformers by Vaswani et al. [50] marked a significant leap in natural language processing (NLP), enabling models to effectively utilize the attention mechanism to capture complex dependencies within text. Attention mechanisms allow LLMs to weigh the importance of various input elements, making them particularly adept at understanding context and generating coherent responses.

LLMs, such as GPT [2] and BERT [10], are pre-trained on extensive and diverse datasets, allowing them to generalize across numerous downstream tasks, including text classification, summarization, question answering, and more [5]. This broad training equips them with a foundational understanding of language, enabling fine-tuning for task-specific applications.

However, despite their remarkable capabilities, LLMs are prone to issues such as hallucination, where they generate plausible but factually incorrect information, and context misalignment, where responses deviate from user queries. These limitations present challenges, especially in applications requiring high accuracy or domain specificity.

2.4 Retrieval-Augmented Generation

To address these challenges, Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm. RAG integrates the generative capabilities of LLMs with external knowledge retrieval systems. By incorporating relevant information retrieved from external sources during the generation process, RAG mitigates hallucination and ensures contextual alignment [32]. The RAG framework operates by querying a knowledge base with a user input, retrieving relevant documents, and using this retrieved information as context for the generative model. This approach combines the strengths of retrieval-based and generative systems, ensuring more accurate and contextually appropriate outputs.

Work in the AI community to improve RAG applications uses techniques such as semantic chunking, contextual compression, and alternate embeddings [30]. RAG is highly dependent on the retrieved documents' quality, and performance is improved by retrieving better document chunks that are more related to the question. There has been limited work on specific applications of these techniques for education use cases, with most prior systems utilizing a recursive character splitter [40, 48, 55]. There has been no comparison between commercial systems using RAG for Q&A with custom RAG implementations deployed over various LLMs.

There are specific questions unique to the educational context. Since the course materials used for RAG are much smaller than required for general question answering, it is interesting to determine how best to use these materials and techniques to improve the generation pipeline. An instructor may upload the syllabus, all notes, assignments, etc., which may be fewer than 100 documents for a typical course. Uploaded course materials may suffer from a lack of details and scope to help with student questions. Guardrails restricting the LLM to using only provided materials may prevent the LLM from answering questions it could handle without RAG.

2.5 Applications of AI in Education

The integration of LLMs into education is transforming learning experiences by enabling personalized instruction and improving student engagement. Notably, Jill Watson, developed by Goel *et al.* [18], demonstrated that AI tutors can reduce instructor workload and enhance engagement. Subsequent iterations streamlined chatbot training, reducing setup time to under 24 hours [17]. While studies have demonstrated AI's proficiency in computer science question-answering [16, 37, 41, 52], comparative analyses across different instructional applications remain limited. Massive Open Online Course (MOOC) environments provide abundant Q&A data to justify chatbot training costs, but individual instructors often lack the necessary data and resources to build course-specific models. Cunningham *et al.* [9] compared chatbot frameworks such as Dialogflow, RASA, and Wit.ai, emphasizing the critical need for high-quality training data.

With advancements in transformer-based LLMs, several institutions have developed specialized AI-powered educational assistants, including Harvard's CS50 AI [33], Georgia Tech's new Jill Watson [48], and Mississippi State University's BARKPLUG [40]. These implementations commonly leverage Retrieval-Augmented Generation (RAG) to enhance question-answering with course-specific content, offering a more focused approach than general-purpose models trained on vast web-based datasets [13, 32].

Most educational RAG systems utilize in-prompt retrieval, where relevant documents stored in a vector database are queried and presented as context within prompts to transformer-based LLMs [40, 55]. Thus, their effectiveness often depends on prompt engineering to constrain response domains and ensure accuracy.

Despite these advancements, several challenges persist. Course-specific RAG implementations typically rely on limited datasets, potentially limiting retrieval quality. Strict guardrails confining LLMs to course materials may limit their ability to answer questions that general models could otherwise handle. Kuhail *et al.* [29] highlight that chatbots struggle to generalize beyond their training data, unlike human instructors who can address novel queries.

Beyond technical hurdles, adoption challenges remain a key consideration. Success depends on addressing student skepticism and instructor concerns regarding AI reliability and integration. A major barrier is ensuring trust: students are wary of incorrect answers, while faculty must validate AI outputs without increasing their workload [4, 22]. Additionally, technology fatigue can arise when faculty must frequently monitor AI-generated responses [3, 25, 57].

Overall, while LLM-based educational assistants offer promising benefits, effective deployment must address technical, pedagogical, and adoption-related challenges.

2.6 Evaluating LLMs

Evaluating Large Language Models (LLMs) has been performed with datasets like FEVER [49] and HotPotQA [59]. The Fact Extraction and Verification (FEVER) dataset evaluates a model's ability to verify claims against a set of facts from Wikipedia. HotPotQA focuses on multi-hop question answering, evaluating a model's ability to retrieve and reason over multiple documents [59]. There is no standardized data set and testing framework for education Q&A, which would be valuable as these generic testing frameworks do not capture the questions asked in education.

2.6.1 Evaluating RAG systems

Chen et al. [7] devised RGB, a RAG specific benchmark to evaluate LLMs' ability to handle context that can include noise, counterfactual content, and negative rejection. The tests are generated from prompting ChatGPT together with related news articles. They asked ChatGPT to generate test cases and checked the test cases manually. During tests, Google Search API is used to retrieve relevant information to accompany the queries. Similarly, RECALL was introduced to focus on RAG systems efficacy when dealing with counterfactual knowledge in context. Results show that LLMs are easily influenced by counterfactual information [34]. CRAG, produced by Meta, creates custom test sets. Instead of focusing on a LLM's ability to parse context, CRAG aims to test on 3 areas: web retrieval summarization, knowledge graph aided retrieval and web retrieval augmentation, and end-to-end RAG. The retrieval component uses the brave search API [58].

A recent benchmark, DomainRAG, leverages domain specific context instead of databases like Wikipedia. However, they set up test cases with preset documents, which does not evaluate the retriever component [56].

Evaluation of assistant RAG systems is focused on providing frameworks, metrics, and methods. IBM released InspectorRAGet and Meta produced Comprehensive RAG Benchmark systems. InspectorRAGet, like RAGAS [11], aims to provide a platform for which metrics of evaluation and a pipeline is provided. Langchain provide their own platform, LangSmith, that evaluates assistant RAG systems by customizing test cases[1].

[1] https://www.langchain.com/langsmith

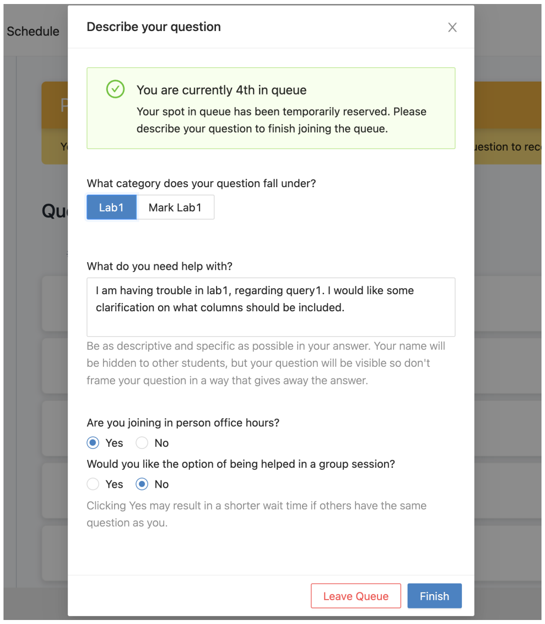
Chapter 3

HelpMe Queueing System

This chapter brings together the implementation, evaluation methodology, and results for the HelpMe queue system.

3.1 Implementation

The HelpMe Queue system [54] provides a visual queue to students and instructors (Figures 3.1 and 3.2). A course may have multiple queues for different purposes (labs, office hours, exam Q&A, etc.). Interactions can be in-person or virtual via Zoom or Microsoft Teams. A staff member (instructor or TA) "opens" a queue by checking in, then students can join the queue by submitting a question via the form shown in Figure 3.1.



Describe your question

You are currently 4th in queue

Your spot in queue has been temporarily reserved. Please describe your question to finish joining the queue.

What category does your question fall under?

Lab1 | Mark Lab1

What do you need help with?

I am having trouble in lab1, regarding query1. I would like some clarification on what columns should be included.

Be as descriptive and specific as possible in your answer. Your name will be hidden to other students, but your question will be visible so don't frame your question in a way that gives away the answer.

Are you joining in person office hours?

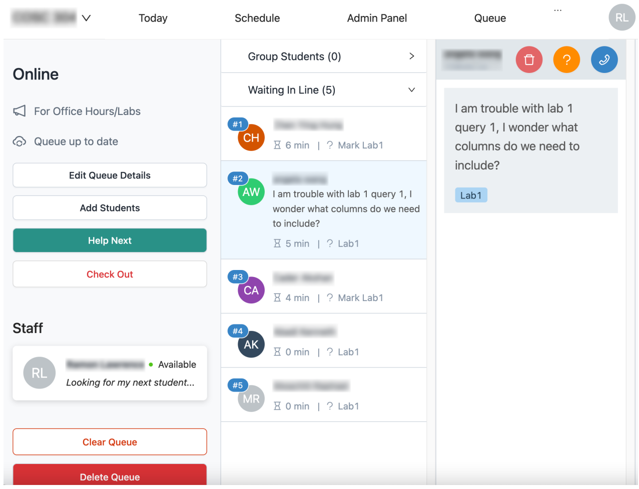
● Yes ○ No

Would you like the option of being helped in a group session?

○ Yes ● No

Clicking Yes may result in a shorter wait time if others have the same question as you.

Figure 3.1: Student View of the HelpMe System



Looking for my next student...

I am trouble with lab 1 query 1, I wonder what columns do we need to include?

I am trouble with lab 1 query 1, I wonder what columns do we need to include?

Figure 3.2: Instructor/Professor View of the HelpMe System

A key feature is that the student's place in line is guaranteed from when they start filling out the form (not after). This encourages more descriptive questions. Staff can see the queue, choose whom to help next, and provide short, focused sessions. The system logs wait times and help times. A configurable 15-minute timer helps ensure sessions do not become excessively long, based on best practices from prior work [44].

Implementation Details. The system is built with a web frontend and a backend that stores questions, timestamps, categories, and user data. It integrates with institutional single sign-on (e.g., Shibboleth). Staff have an admin view (Figure 3.2) with full queue control, while students see only their own question details plus aggregated wait times.

3.2 Evaluation Methodology

This evaluation focuses on the HelpMe queue system's effectiveness for in-person and virtual office-hour management using both survey and quantitative data, approved under a university ethics study.

3.2.1 Data Sources and Participants

System Logs The primary dataset derives from HelpMe system logs, which capture instructor-student interactions, wait times, session duration, question categories, and timestamps. A secondary dataset is exam and final grades from consenting participants.

Participants The course had 107 registered students of whom 67 (62%) consented to have their grades correlated with HelpMe usage. Overall, 83 students used the system at least once.

3.2.2 Procedure

The instructor and TAs were trained in using HelpMe. All office-hour sessions, both in-person and virtual, employed HelpMe to queue and log help requests for the entire semester.

- Students could self-queue via the system interface; for any student arriving without having queued, staff manually added them prior to providing assistance.

- Email-based questions were also manually logged as "email-based help sessions" to unify all help interactions in the same log.

Data Cleaning Interaction times under 30 seconds were removed (likely test queries or errors). Sessions over 40 minutes were capped at 40 minutes if staff forgot to close them in a timely fashion.

3.2.3 Survey Instrument

During the final two weeks of the course, students completed a survey about HelpMe's usability, impact, and their overall help-seeking preferences. Forty students responded. The survey also included the System Usability Scale (SUS) [6]. The survey questions and details are in Appendix B.1. Survey contents includes questions in the following directions:

- General help usage, preference for office hours vs. other modes of support.

- Perceived effectiveness of HelpMe in managing queues.

- Perceived impact of help sessions on course performance.

- Standard SUS items to gauge overall usability.

3.3 Results

Help Session Utilization

Figure 3.3 shows categories of questions asked: administrative (personal or logistical), assignments, general course content, and exams. Figure 3.4 shows the email question breakdown. In total, 539 help sessions and 146 emails were tracked.

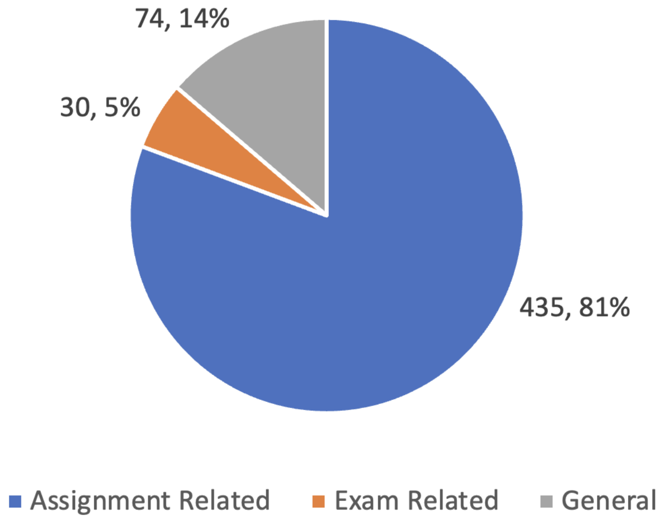


Figure 3.3: Help Session Question Categorization

Consistent with prior work on office hours [44], usage varied widely: 19% of students accounted for 64% of total sessions, and 54% used it never or only once (Figure 3.5).

A survey question on reasons for not attending office hours revealed that commute time and long waits were common deterrents (Table 3.1), matching past studies [1, 45].

Table 3.1: Reasons Detering Students from Office Hours

Reason	**Agree %**
Commute not worth it for short Q	68%
Not an effective use of time	33%
Long wait times	28%
Confusion over queue/wait	25%
Instructor not helpful	23%

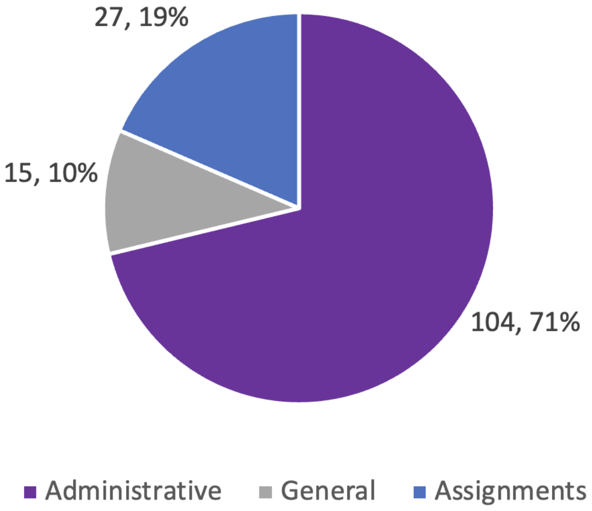


Figure 3.4: Email Question Categorization

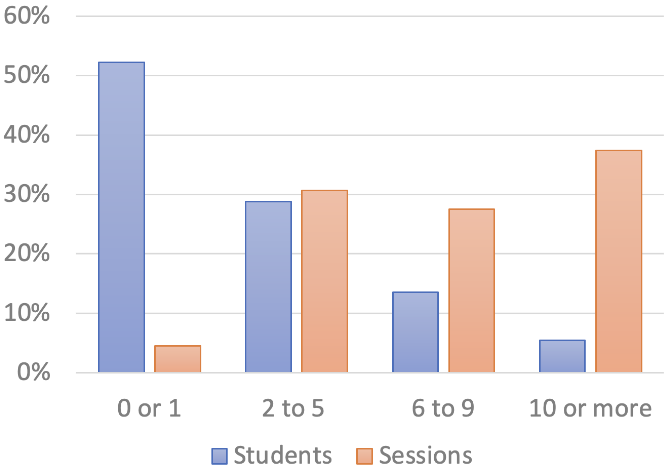


Figure 3.5: Session Breakdown by Student

Session Times

Figures 3.6 and 3.7 illustrate average help session durations (433 s) and wait times (357 s). The 15-minute timer reminder helped keep sessions short.

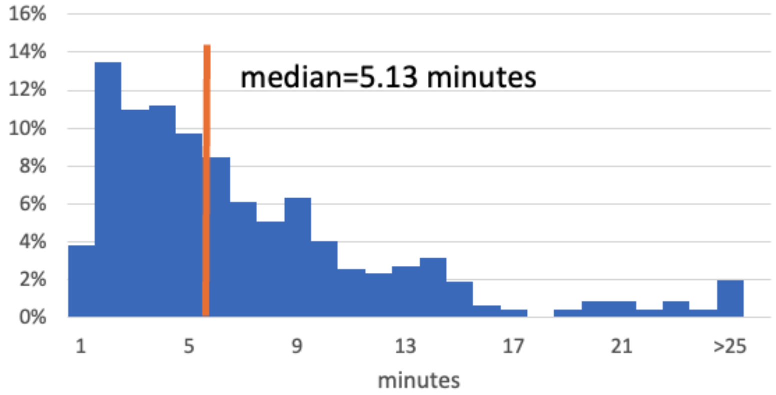


Figure 3.6: Distribution of Help Session Durations

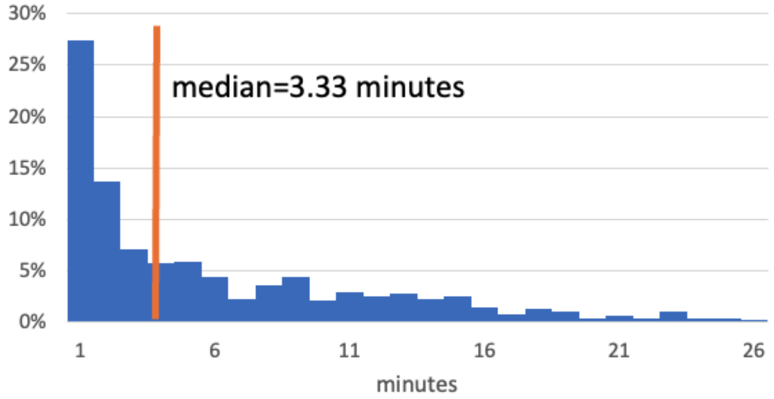


Figure 3.7: Wait Time Distribution

Student Satisfaction

Table 3.2 reports students' agreement (1–5) with statements about HelpMe. They found it provided strong visibility, efficiency, and an improved experience vs. courses lacking a queue system. The SUS score was 77.6 (out of 100). A score of 70 represents above average usability [6].

Open comments mentioned the transparency of wait times and fairness. One student noted, *"I enjoyed knowing my place in line and seeing the average wait times."* Some critiques focused on the need for better integration with Zoom.

Table 3.2: HelpMe Queue System Satisfaction (1=Strong Disagree, 5=Strong Agree)

Question	**Mean**
Provided more visibility on wait/service time	4.44
Improved office hours vs. other courses	4.26
Made office hours more efficient	4.26

Student Perception & Preferences

Table 3.3 compares students' self-reported preferred help channels before vs. after using HelpMe. While only 24% originally liked office hours (traditional), 63% now prefer hours using the HelpMe queue.

Table 3.3: Preferred Help-Seeking Methods: Before vs. After

	Before	**After**
Office Hours	24%	10%
Email	42%	16%
No help-seeking	22%	10%
After class	12%	0%
Office Hours+HelpMe	N/A	63%

This shift is likely due to reduced wait uncertainty and no need to commute (for virtual sessions). Students' email usage was reduced, although personal or sensitive matters still were common reasons for emails (Figure 3.4).

Student Performance

Figure 3.8 plots final exam grades vs. number of help sessions used. Students with two or more sessions averaged 80%, slightly above the overall 78% class average, whereas students with zero or one session averaged 73%. The difference was not statistically significant ($p = 0.068$). This suggests a mild correlation, but also that more engaged students were naturally more likely to attend office hours.

Additional factors present in the course like code review incentives during office hours [53] and student self-selection likely impacted these correlations.

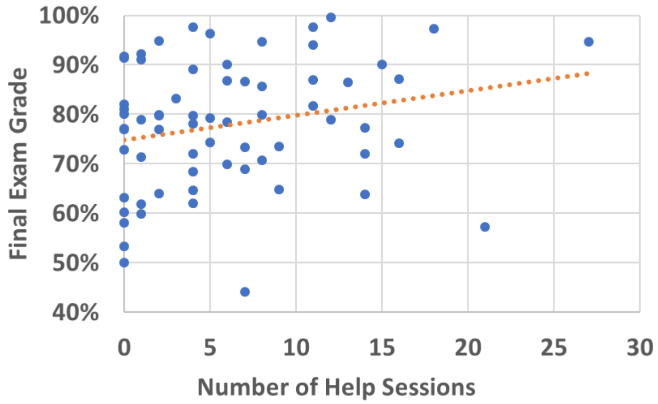


Figure 3.8: Final Exam Performance vs. Help Session Usage

Chapter 4

ChatEd and Evaluations

This chapter presents the implementation of ChatEd, the methodologies used to evaluate both ChatEd itself and broader RAG systems, and the corresponding results.

4.1 Implementation of ChatEd

ChatEd is a Retrieval-Augmented Generation (RAG) system that provides an intelligent, context-aware chatbot for educational settings [55]. It leverages a Large Language Model (LLM), such as GPT-3.5 or GPT-4, and grounds responses in course-specific materials via retrieval from a vector database with course content.

4.1.1 Interface Design

ChatEd's student interface is a familiar chat window (Figure 4.1), embedded in a Learning Management System or any website. Key features include:

– **Context Relevancy**: uses local course documents (lecture slides, PDFs, notes) so answers match the instructor's curriculum.

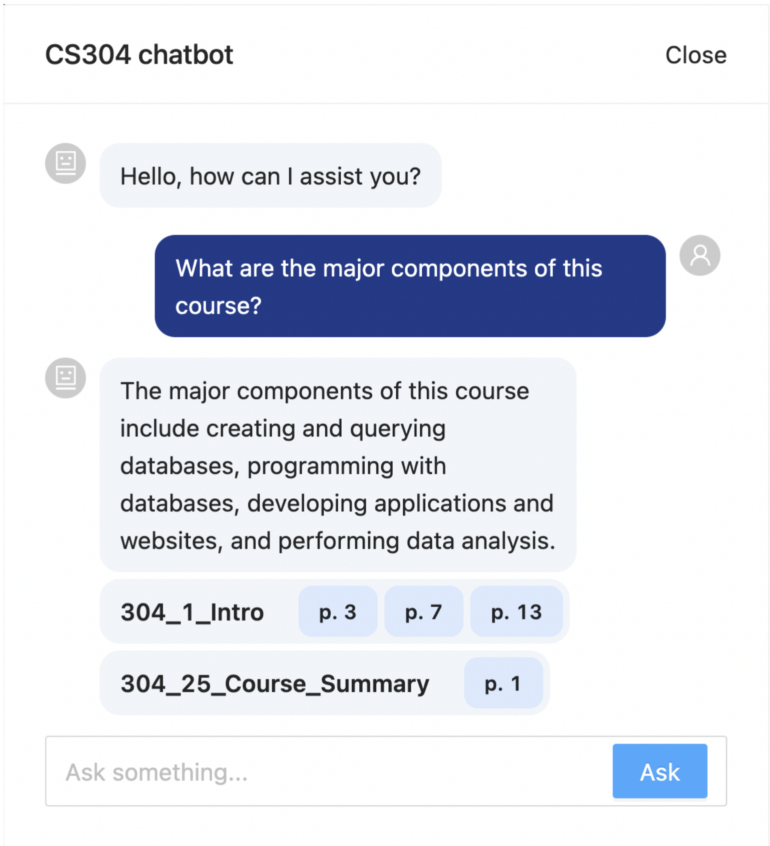
– **Customizability**: instructors upload or remove documents, set the retrieval parameters, etc.

– **Scalability**: allow different instructors to customize their own course easily.

For instructors, ChatEd offers a management interface (Figure 4.2) to upload new materials, update chunking strategies, and view usage analytics. As a result, no direct model training is needed to improve the pipeline.

4.1.2 ChatEd's RAG Architecture

Figure 4.3 summarizes ChatEd's pipeline with the main components:



Hello, how can I assist you?

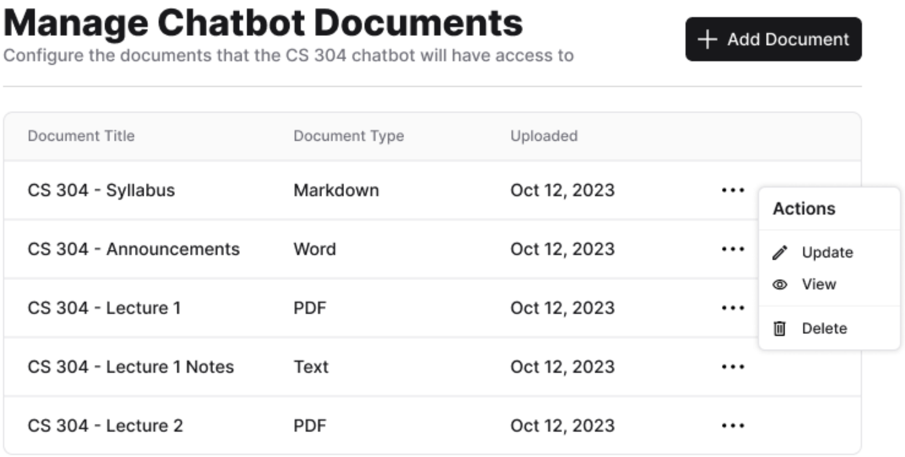
What are the major components of this course?

The major components of this course include creating and querying databases, programming with databases, developing applications and websites, and performing data analysis.

304_1_Intro p. 3 p. 7 p. 13

304_25_Course_Summary p. 1

Figure 4.1: Chatbot User Interface and Example Conversation [55]



Manage Chatbot Documents

Configure the documents that the CS 304 chatbot will have access to

Document Title	Document Type	Uploaded
CS 304 - Syllabus	Markdown	Oct 12, 2023
CS 304 - Announcements	Word	Oct 12, 2023
CS 304 - Lecture 1	PDF	Oct 12, 2023
CS 304 - Lecture 1 Notes	Text	Oct 12, 2023
CS 304 - Lecture 2	PDF	Oct 12, 2023

Figure 4.2: Instructor Interface in ChatEd [55]

1. **Document Ingestion**: Lecture slides, PDFs, web pages are chunked and embedded using Faiss or PGVector.

2. **Retrieval**: When a student asks a question, the system retrieves the top-K relevant chunks from the database.

3. **LLM Integration**: The chunks and conversation history form the prompt for the LLM, which generates a final answer referencing the context.

Since the RAG approach "grounds" the model's answers in specific course materials, ChatEd can answer many class-specific or policy questions beyond general LLM capabilities.

4.2 Evaluation Methodology

4.2.1 ChatEd Evaluation

After building the initial pipeline, ChatEd was tested with a set of 60 questions that included:

– **General questions:** e.g., "Can each table have multiple foreign keys?"

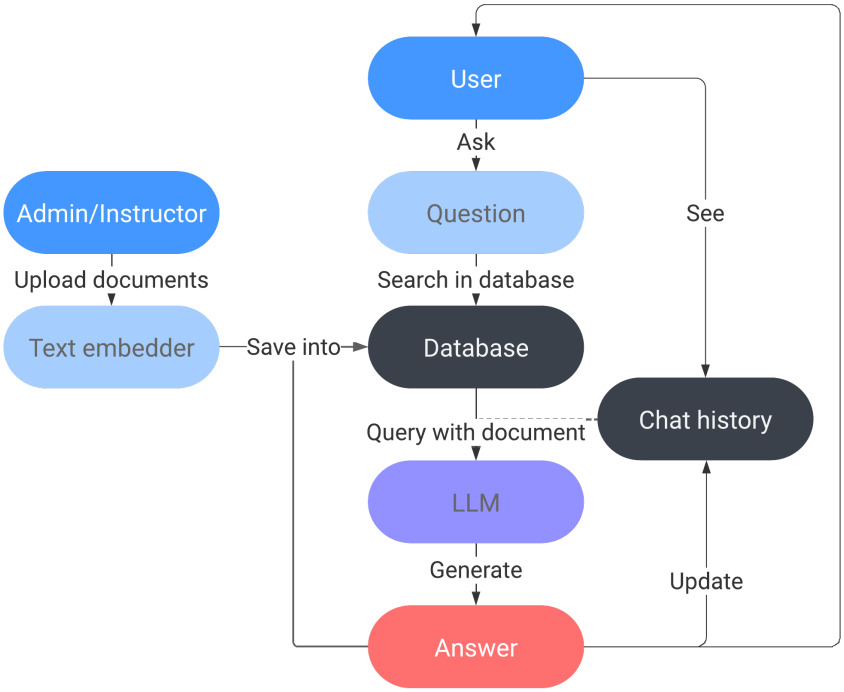


Figure 4.3: ChatEd Chatbot Architecture [55]

- **Domain-specific questions:** e.g., “Are ‘chunk’ and ‘block’ interchangeable terms?”
- **Managerial questions:** e.g., “Is the exam open-book?”

These questions were partly drawn from previous research on office-hour help interactions [54]. Two database courses (DB1 and DB2) were used. The relevant course materials (slides, labs, syllabi) were uploaded to the chatbot for retrieval.

Question-Answering Metrics Evaluators (the instructor and TAs) assessed based on three metrics:

- **Relevance:** How directly the response addressed the user’s query.
- **Accuracy:** Whether the response was factually correct given the course materials.
- **Helpfulness:** Clarity, tone, and whether the response improved user understanding.

A side-by-side comparison with ChatGPT was also done to gauge differences in accuracy and relevance for the same queries.

Context Awareness Testing

The research further examined whether ChatEd maintained context across multi-turn conversations. Five scripted interaction sets per course were designed as follows:

1. **Establish a Context (A)**: A context-setting question.
2. **Follow-up (B)**: Implicitly relies on context A.
3. **Deepen the Context (C)**: More specific questions about A.
4. **Break the Context (D)**: An unrelated query to see if ChatEd can switch.
5. **Revisit the Context (E)**: Return to context A after the interruption.

Responses were documented and analyzed for consistency and reliability in context retention.

4.2.2 Evaluating AI and RAG Quantitatively

Building on ChatEd's relatively small-scale manual testing, this phase of evaluation involved a more robust, automated approach with larger datasets and multiple LLMs. The experiments also addressed practical considerations in educational AI deployment (e.g., model selection, data security, infrastructure costs, and viability of local models).

Models and Systems Evaluated

Testing used multiple open-source and commercial LLMs on question data from four CS courses (Computer Science 1 (CS1), Computer Science 2 (CS2), Introduction Databases (DB1), Advanced Databases (DB2)):

- `gemma2` (Published 2024-06-27)
- `llama3` (Published 2024-04-28)
- `phi3` (Published 2024-04-23)
- `gpt-3.5-turbo-0125` (Published 2024-01-25)
- `gpt-4o` (Published 2024-05-13)
- `gpt-4o-mini` (Published 2024-07-18)

Local hosting tests used a server with dual Intel Xeon Platinum CPUs, 1 TB of RAM, and an Nvidia RTX 6000 GPU. Hosting costs were estimated using pricing of $1 per GPU hour.[2]

Evaluation Dataset

A total of 241 instructor- and TA-curated test questions (83% general, 17% requiring specific context) were created. Each question was associated with:

- A category (e.g., content clarification, policy, etc.).
- A ground truth answer.

Figure 4.4 shows the category distribution.

[2] `https://www.runpod.io/gpu/6000-ada`

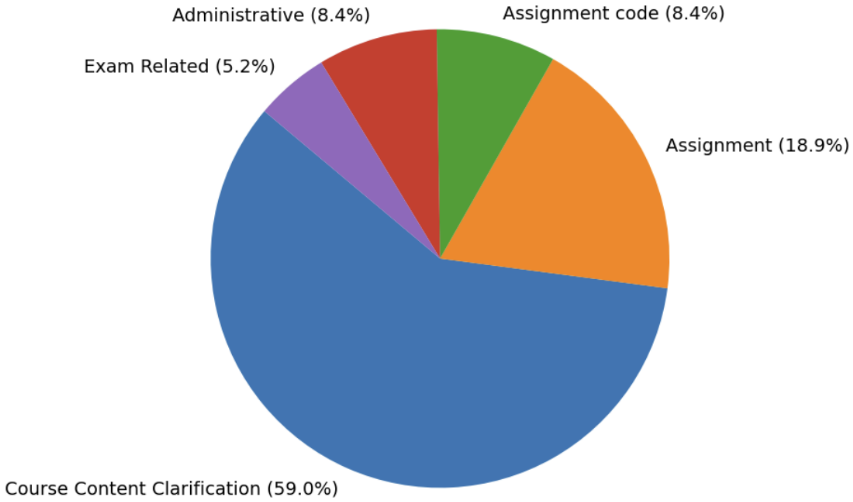


Figure 4.4: Distribution of Question Categories

Testing Framework

The benchmark was evaluated using a baseline PGVector implementation with the LangChain library and OpenAI embeddings. The system performs recursive text splitting with 1000-character chunks and a 20-character overlap, leveraging both ChatGPT and locally hosted LLMs on an Nvidia RTX 6000 GPU. The evaluation framework employs three key metrics to compare generated responses against ground truth answers:

- **TF-IDF:** Measures lexical similarity by computing cosine similarity between the ground truth and generated responses based on term frequency-inverse document frequency (TF-IDF) representations[47].

- **Similarity:** Computes cosine similarity between the embeddings of ground truth and generated responses using OpenAI's text-embedding-ada-002 model.[3] Compared to TF-IDF, this metric captures semantic relationships beyond surface-level word overlap.

- **Correctness:** Assessed using `Ragas` RAG evaluation's factual correctness metric [11] and using the GPT-4o-mini model as an LLM-based judge, following the protocol in [60]. Each factual statement in the

[3] Embedding introduced at
`https://openai.com/index/new-and-improved-embedding-model`

AI-generated response is categorized as True Positive (TP), False Positive (FP), or False Negative (FN) relative to the ground truth. The correctness score reflects overall alignment with the reference answer.

4.2.3 RAG Optimizations and Human Feedback

Further evaluation considered multiple areas of optimization of RAG systems. For commercial systems, course documents are uploaded as PDFs into the system. There is no user control of the RAG processes for these systems. For local hosted models, the base RAG system implementation encoded course documents using PGvector and OpenAI embeddings. Chunking is performed first by page, then by 1000 characters with a 20-character overlap using a recursive text splitter similar to prior work [40, 48, 55]. The prompt used in all systems is designed to incorporate relevant context effectively.

Two RAG optimizations are explored:

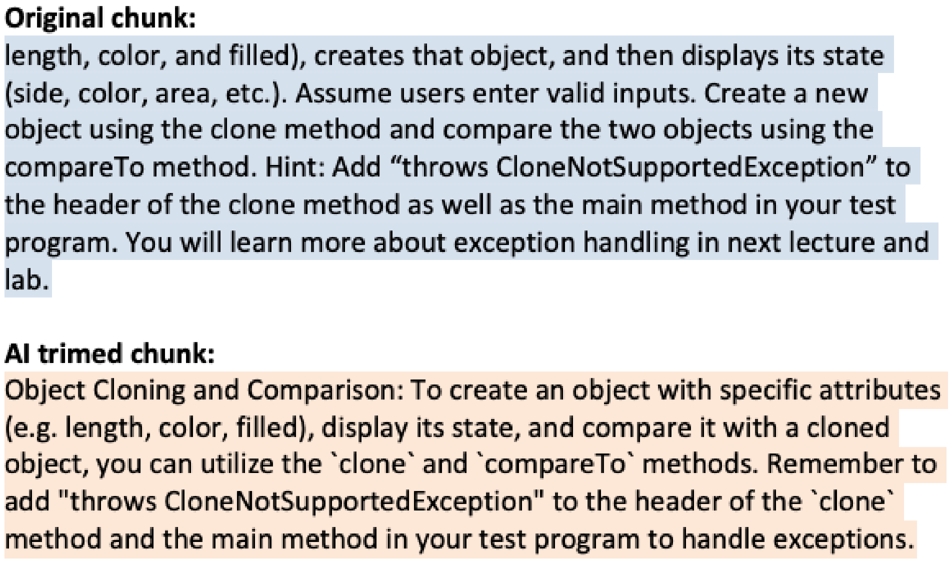
- **AI-assisted content curation:** have course contents automatically summarized by AI (see Figure 4.5 as an example)
- **Question reuse:** encodes answers to questions and stores them as content for use by RAG

Three RAG databases are evaluated:

- **Content database:** contains the course materials
- **AI-edited database:** has the course materials edited by AI
- **Content and question database:** has course materials and question-answer pairs

Two types of prompts are evaluated:

- **With heavy guardrails:** “RULES: 1) If you don’t know the answer just say that “I don’t know”, do not try to make up an answer. 2) If you are unsure of the answer, you shall PREFACE your answer with “I’m not sure, but this is what I think.” 3) Try to be as concise as possible. 4) Do not use any other resources apart from the context provided to you.”
- **With light guardrails**: “You are an educational assistant. Here are some rules for question answering: 1) Try to be as concise as possible. 2) Refer to context as you see fit.”



Original chunk:

length, color, and filled), creates that object, and then displays its state (side, color, area, etc.). Assume users enter valid inputs. Create a new object using the clone method and compare the two objects using the compareTo method. Hint: Add “throws CloneNotSupportedException” to the header of the clone method as well as the main method in your test program. You will learn more about exception handling in next lecture and lab.

AI trimed chunk:

Object Cloning and Comparison: To create an object with specific attributes (e.g. length, color, filled), display its state, and compare it with a cloned object, you can utilize the `clone` and `compareTo` methods. Remember to add "throws CloneNotSupportedException" to the header of the `clone` method and the main method in your test program to handle exceptions.

Figure 4.5: Example of AI Edit

ChatGPT’s Assistants API controls all aspects of the documents used to answer queries and prompt engineering for providing context. Thus, the Assistants API prompt is adapted to remove the last rule regarding how to use context.

Testing the content and AI-edited database uses the question set. Testing the database containing question answers is designed to determine if instructors providing feedback on correct answers can help answer similar questions. These similar questions ($N = 723$) are generated using an LLM prompted to mimic student questions and are the test set of the experiment, not the original question set.

4.3 Results

4.3.1 ChatEd Evaluation Results

This section presents the first evaluation of ChatEd using a set of curated questions. The study measures question-answering performance, conversational depth, and provides an initial case study of in-course deployment.

Question Answering Results

ChatEd was evaluated using a sample of 20 questions from the question bank. Each question was posed to the system, and its answer was evaluated on a scale of 1 to 5 (5 = best) along three criteria: *relevancy*, *accuracy*, and *helpfulness*. The same questions were also posed to ChatGPT 3.5 as a comparison. Table 4.1 shows the average scores.

Criteria	**ChatEd Score**	**ChatGPT Score**
Relevancy	5.0	4.4
Accuracy	5.0	4.4
Helpfulness	4.5	3.4

Table 4.1: Average Scores Given by Evaluators (Q&A Test)

Overall, ChatEd performed exceptionally on these 20 questions, returning highly relevant and accurate answers. In contrast, ChatGPT's responses, while still fairly strong, were sometimes overly verbose or lacked course-specific details.

Key observations include:

- **Course policies:** ChatEd excelled at "managerial" (policy-related) questions because it had direct access to the instructor's official documentation. ChatGPT, lacking these specifics in its training data, could only guess.
- **Conciseness:** ChatEd was generally more concise, referencing exactly the bullet points taught in class.
- **Reviewer notes:** TAs highlighted how ChatGPT's "exhaustive bullet points" on, e.g., "errors when retrieving database results," included extraneous cases not emphasized in class, whereas ChatEd's 5-point list matched the official course lecture.

Conversational Depth

The next experiment tested ChatEd's ability to maintain context across multiple user turns. Five multi-turn conversations were constructed, in which the user asked follow-up questions referencing earlier statements. Each conversation was scored (1–5) on *relevancy*, *accuracy*, and *helpfulness*. Table 4.2 summarizes the overall averages.

Criteria	ChatEd Avg. Score
Relevancy	4.19
Accuracy	5.00
Helpfulness	4.06

Table 4.2: Context-Aware Conversation Scores

Figure 4.6 shows an example where the user follow up with questions. ChatEd generally followed the user's reference, but occasionally it confused "this" with other topics. Performance dropped if the follow-up question was too vague (e.g., "Why do we need this again?" after multiple context shifts).

Deployment and Case Studies (Pilot)

ChatEd the chatbot was deployed in a DB1 course, where 76 student questions were asked. While many were straightforward (e.g., definitions of SQL terms), several illustrated typical "low-quality" or ambiguous user queries that led to non-ideal answers (Tables below).

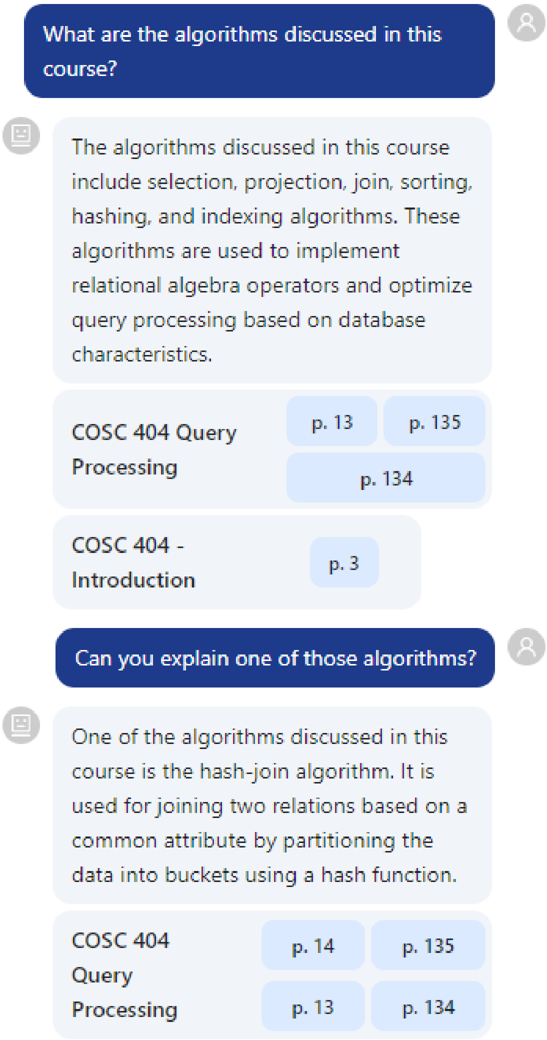
User Question	System Answer
How to MAKE A TICKET?	*[ChatEd interprets as "creating a database ticket" and gives an irrelevant DB design suggestion.]*

Table 4.3: Example of a misunderstood question about the HelpMe queue rather than database tickets.

User Question	System Answer
analysis stream vs development stream	ChatEd pulls relevant project info: "Analysis focuses on data analysis/reporting, Dev focuses on building e-commerce with SQL, etc."

Table 4.4: Example of leveraging course knowledge to provide a context-specific answer.

These examples show both ChatEd's strengths and potential pitfalls if student queries are unclear.



What are the algorithms discussed in this course?

The algorithms discussed in this course include selection, projection, join, sorting, hashing, and indexing algorithms. These algorithms are used to implement relational algebra operators and optimize query processing based on database characteristics.

COSC 404 Query Processing — p. 13 p. 135 p. 134

COSC 404 - Introduction — p. 3

Can you explain one of those algorithms?

One of the algorithms discussed in this course is the hash-join algorithm. It is used for joining two relations based on a common attribute by partitioning the data into buckets using a hash function.

COSC 404 Query Processing — p. 14 p. 135 p. 13 p. 134

Figure 4.6: Example of Follow-up Questions in a Conversation

User Question	System Answer
What is a GRANT?	"A GRANT is a command in SQL that gives privileges on database objects to users. . . . "

Table 4.5: Straightforward SQL question receiving a concise, correct system answer.

User Question	System Answer
can i chat directly with TA?	"Yes, you can chat directly with the TA."

Table 4.6: Unhelpful or misleading response (no mention of official protocols).

4.3.2 Discussion of ChatEd Manual Evaluation

The first iteration of ChatEd demonstrated strong Q&A performance on curated questions and reasonable multi-turn context handling, outperforming raw ChatGPT on course-specific queries. However:

- Ambiguous queries led to irrelevant or incomplete answers.
- Students often typed incomplete questions (e.g., "Where is the final?").
- ChatEd occasionally conflated references when the conversation turned vague.

Despite these challenges, the results suggest RAG-based systems can be highly effective for specialized, course-targeted Q&A, especially if students formulate clear queries. These initial results led to a further investigation with a broader set of LLMs and more advanced RAG strategies.

4.3.3 Quantitative RAG Evaluation Results

Building on ChatEd, this section expands the evaluation to multiple LLMs (both commercial and open-source) across 241 curated questions plus 723 synthetic variants. The focus is on (1) raw non-RAG performance, (2) RAG performance on course materials, (3) potential instructor optimizations, (4) cost, and (5) final recommendations.

LLM Performance (Non-RAG)

Table 4.7 compares various models (GPT, Llama3, Phi3, Gemma2) on a subset of questions that do not require course context. Bold numbers indicate the best metric per column. We measure TF-IDF similarity, semantic similarity (MiniLM), correctness (using Ragas [19]), average response time, and cost per 1000 queries.

Model	TF-IDF	Similarity	Correctness	Time (s)	Cost
gpt-3.5-turbo-0125	0.466	0.844	0.531	**1.25**	$0.12
gpt-4o	0.454	0.840	**0.540**	2.25	$0.20
gpt-4o-mini	**0.467**	**0.849**	0.531	2.38	**$0.08**
llama3:70b	0.446	0.835	0.523	7.60	$2.10
llama3:8b	0.416	0.822	0.482	1.48	$0.41
phi3:14b	0.395	0.810	0.502	2.99	$0.83
phi3:3.8b	0.393	0.825	0.508	1.30	$0.36
gemma2:9b	0.383	0.807	0.514	1.66	$0.46

Table 4.7: Comparison of Non-RAG Systems on General Questions

The observations include:

- **gpt-4o-mini** and **gpt-3.5** have the best cost-performance tradeoff.
- Open-source models like llama3:70b or gemma2:9b are cheaper per query but require hosting overhead (GPU, system admin).

RAG Performance

Table 4.8 shows how these models fare when given course-specific data via retrieval-augmented generation (RAG) and light guardrails. Providing relevant context generally improves correctness and similarity, but also increases query time and cost.

We further compare “context-required” questions in Table 4.9, revealing that all models benefit from the provided domain context, but the improvement is more pronounced for DB courses where the general knowledge might be less well covered in standard model training.

Table 4.10 shows average correctness across four courses in our dataset, illustrating how basic courses (CS1/CS2) are relatively well-answered even without specialized RAG data, while DB topics see more gain from RAG. However, the result for this is not conclusive and uneven, so we will refrain from definitive conclusions.

Model	TF-IDF	Similarity	Correctness	Time (s)	Cost
gpt-3.5-turbo-0125	0.497	0.851	0.578	3.95	**$0.37**
gpt-4o	**0.525**	**0.854**	**0.588**	5.15	$5.48
gpt-4o assistantAPI	0.512	0.853	0.583	9.10	$52.07
llama3:70b	0.511	0.837	0.576	7.29	$2.10
gemma2:9b	0.481	0.830	0.580	2.24	$0.62
llama3:8b	0.474	0.821	0.562	**2.00**	$0.56
phi3:14b	0.460	0.844	0.564	2.69	$0.74
phi3:3.8b	0.428	0.814	0.543	2.01	$0.56

Table 4.8: Comparison of RAG Systems (Light Guardrails).

Model	TF-IDF	Similarity	Correctness
gpt-3.5-turbo-0125	0.416	0.714	0.519
gpt-4o	**0.448**	0.718	0.539
gpt-4o-mini	0.431	**0.749**	0.556
llama3:70b	0.428	0.724	0.522
llama3:8b	0.398	0.691	0.472
phi3:14b	0.386	0.674	0.482
phi3:3.8b	0.367	0.702	0.486
gemma2:9b	0.425	0.729	**0.557**

Table 4.9: Context-Specific Questions with RAG (All Models).

Course	Baseline Score	RAG Score
CS1	0.6000	0.6435
CS2	0.6131	0.6214
DB1	0.5645	0.6748
DB2	0.5715	0.5974

Table 4.10: Average Model Scores Across Different Courses

Instructor Optimizations

Experiments also tested whether advanced chunking, AI trimming, or previously answered Q&A storage improves performance meaningfully. Table 4.11 compares AI-trimmed content vs. typical chunking. Differences are modest; AI-trimmed content was slightly better on some "context-related" metrics but took nearly the same response time.

Chunking Method	TF-IDF	Time (s)	Similarity	Relevancy	Recall	Precision
AI-trimmed	0.48	7.3	0.84	0.49	0.68	0.93
Recursive-chunk	0.49	7.2	0.82	0.21	0.59	0.84

Table 4.11: Comparison of AI-Trimming vs. Traditional Chunking (llama3:70b).

Another major optimization is storing previously answered questions in the same vector index. Table 4.12 shows that for a test set of "similar queries," retrieving an existing instructor-verified answer can significantly boost correctness.

Database	TF-IDF	Similarity	Correctness	Prompt Type
Content only	0.311621	0.677357	0.500571	Heavy guardrail
Content+Q&A	0.439530	0.737213	0.655102	Heavy guardrail
Content only	0.296113	0.664729	0.463972	Light guardrail
Content+Q&A	0.437722	0.736551	0.640519	Light guardrail

Table 4.12: Using Instructor-Verified "Q&A" as Part of the RAG Database (Llama3:70b).

Deployment Costs

A key question is whether to self-host or use commercial APIs. GPT-3.5 is a top performer for cost-effectiveness, although GPT-4 (or "gpt-4o-mini") can yield slightly better quality. Local hosting requires a GPU server plus administration overhead. Commercial usage remains cheap on a per-query basis but can spike for large context windows or if many queries are made.

Recommendations

The findings indicate:

- For general CS questions, GPT-3.5 or gpt-4o-mini yields excellent performance at low cost.
- RAG helps mostly for advanced or course-specific queries—particularly for local LLMs or less mainstream topics.
- Storing previously answered questions can greatly improve retrieval for repeated queries.
- Unless you have resources and privacy needs that mandate local hosting, using a commercial LLM with a simple RAG pipeline is likely easiest.

Discussion

The results demonstrate ChatEd's strong performance on course-specific Q&A, the benefits of RAG for advanced topics, and the trade-offs between commercial and locally hosted LLMs. Instructor optimizations yield slight gains, and cost analysis suggests commercial models as the most cost-effective choices if data privacy is not a concern.

Chapter 5

HelpMe Integrated Platform

This chapter describes the implementation, methodology, and results of a unified platform.

5.1 Implementation

This section describes the work that integrates the HelpMe queue system (Section 3.1) with ChatEd (Section 4.1) to form a cohesive web platform for both synchronous and asynchronous help. In addition to standard office hours is a new Anytime Question Hub where students can post questions and receive an immediate AI-generated answer; if unsatisfactory, the question escalates to human review.

5.1.1 Integrated System (Home Page)

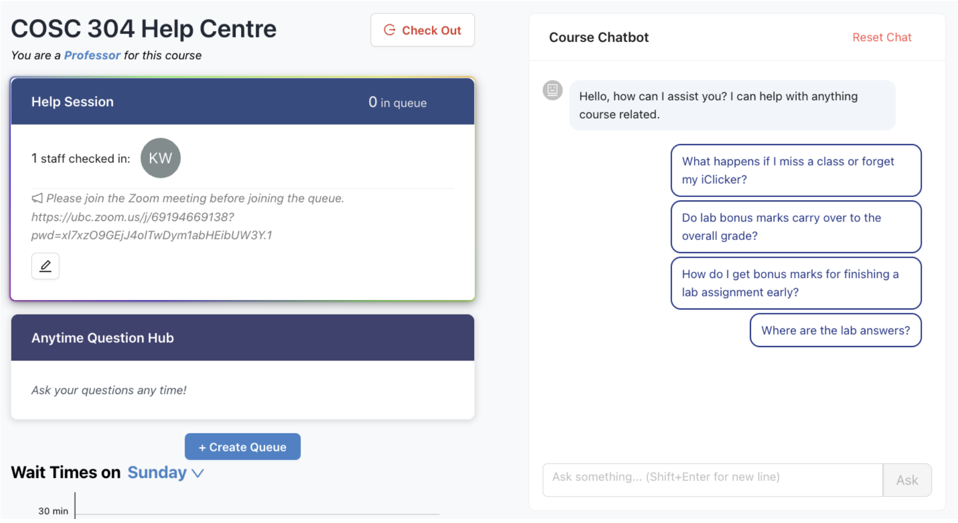
Figure 5.1 shows the system's homepage. There are two main sections:

- **Help Session**: The original queue-based approach for TAs and in-person/virtual office hours.
- **Anytime Question Hub**: Asynchronous Q&A with AI first-response, optional human follow-up.

Instructors can toggle features on/off in course settings (Figure 5.2), customizing which AI or queue features are active.

5.1.2 Chatbot Integration

Inside the integrated HelpMe interface (Figure 5.1) a persistent chatbot interface provides direct access to ChatEd's generation pipeline. Students can ask questions any time, either while waiting in line or outside scheduled help sessions. If the answer is unclear, they can still proceed to queue-based help.



COSC 304 Help Centre

You are a ***Professor*** *for this course*

Help Session — 0 in queue

1 staff checked in: KW

Please join the Zoom meeting before joining the queue. https://ubc.zoom.us/j/69194669138?pwd=xl7xzO9GEjJ4olTwDym1abHEibUW3Y.1

Anytime Question Hub

Ask your questions any time!

Wait Times on Sunday

Course Chatbot

Hello, how can I assist you? I can help with anything course related.

What happens if I miss a class or forget my iClicker?

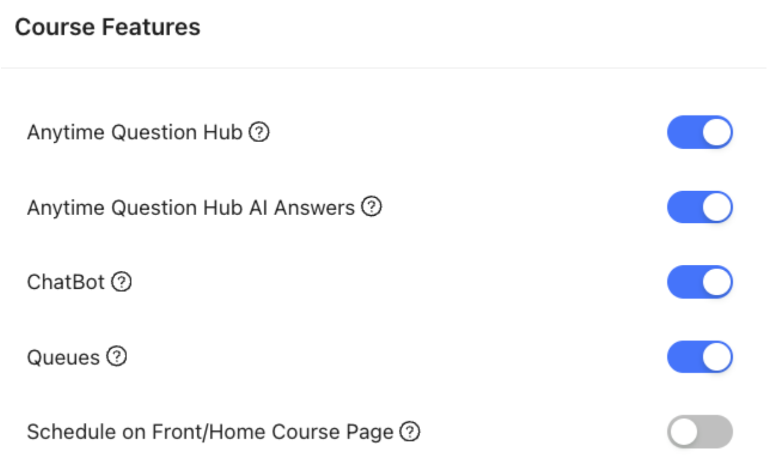
Do lab bonus marks carry over to the overall grade?

How do I get bonus marks for finishing a lab assignment early?

Where are the lab answers?

Ask something... (Shift+Enter for new line)

Figure 5.1: Integrated HelpMe System Homepage



Course Features

Anytime Question Hub ●

Anytime Question Hub AI Answers ●

ChatBot ●

Queues ●

Schedule on Front/Home Course Page ○

Figure 5.2: Course Feature Settings

5.1.3 Anytime Question Hub

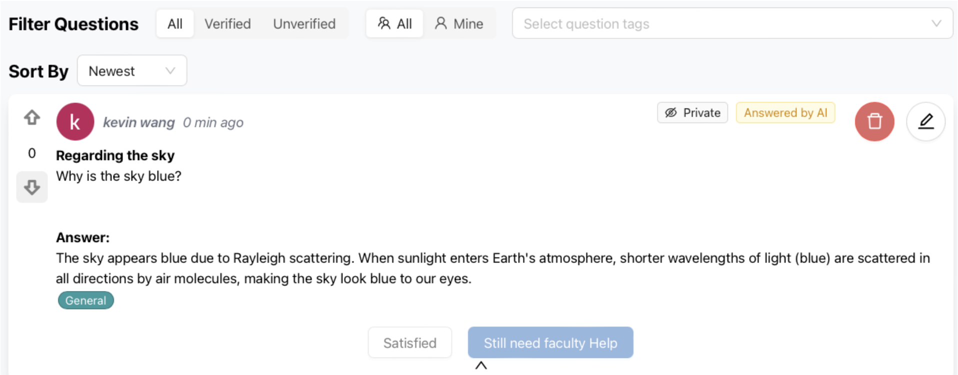
The Anytime Question Hub (Figures 5.3 and 5.4) is designed for asynchronous questions:

1. Student checks an existing public Q&A to see if the question was already answered.

2. If not found, student poses a new question; the system retrieves relevant context and produces an immediate AI response.

3. If the answer is unsatisfactory, the student clicks "Still Need Help," flagging the question for staff review.

4. Instructors and TAs see the flagged question, edit or verify the AI's draft, and optionally mark it public for future reference. All student identity remains hidden.



kevin wang *0 min ago*

Regarding the sky
Why is the sky blue?

Answer:
The sky appears blue due to Rayleigh scattering. When sunlight enters Earth's atmosphere, shorter wavelengths of light (blue) are scattered in all directions by air molecules, making the sky look blue to our eyes.

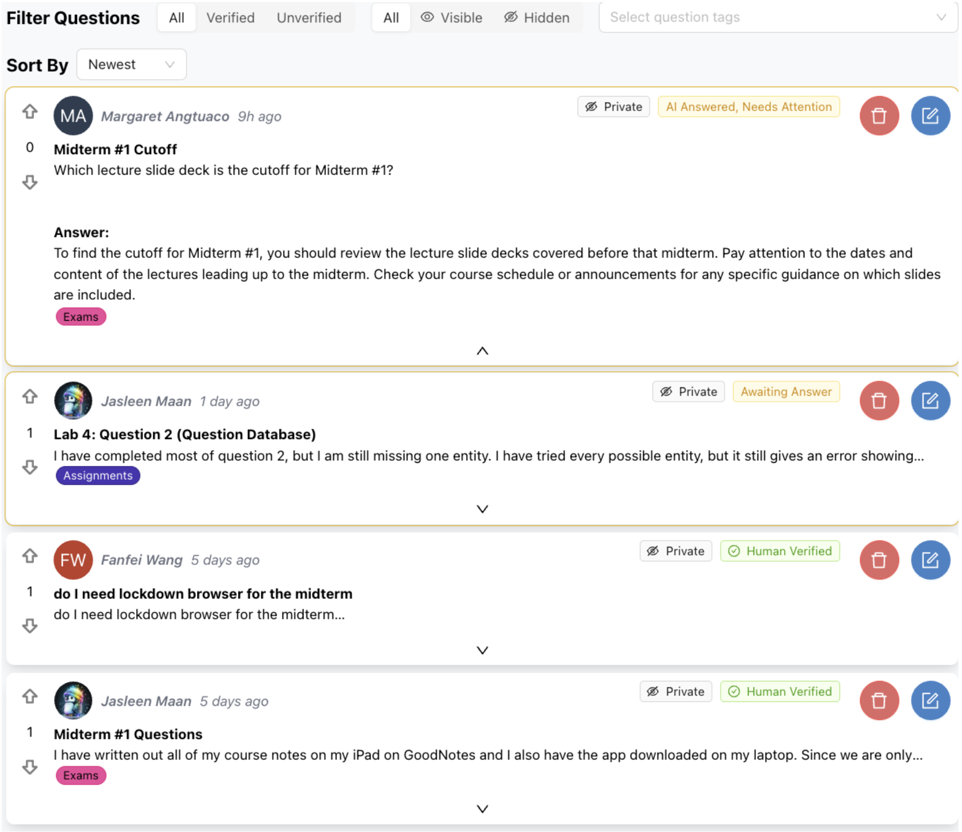
Figure 5.3: Anytime Question Ticket (Student View)

By combining AI-generated answers with a reliable human "safety net," students gain immediate feedback while still receiving correct and verified answers if the AI is inaccurate or incomplete.

5.2 Evaluation Methodology

For the 2024 Fall semester of DB1, the integrated AI features were deployed in production. Data was collected on:

– Number and category of questions asked asynchronously.



Margaret Angtuaco 9h ago · Private · AI Answered, Needs Attention · 0

Midterm #1 Cutoff

Which lecture slide deck is the cutoff for Midterm #1?

Answer:

To find the cutoff for Midterm #1, you should review the lecture slide decks covered before that midterm. Pay attention to the dates and content of the lectures leading up to the midterm. Check your course schedule or announcements for any specific guidance on which slides are included.

Exams

Jasleen Maan 1 day ago · Private · Awaiting Answer · 1

Lab 4: Question 2 (Question Database)

I have completed most of question 2, but I am still missing one entity. I have tried every possible entity, but it still gives an error showing...

Assignments

Fanfei Wang 5 days ago · Private · Human Verified · 1

do I need lockdown browser for the midterm

do I need lockdown browser for the midterm...

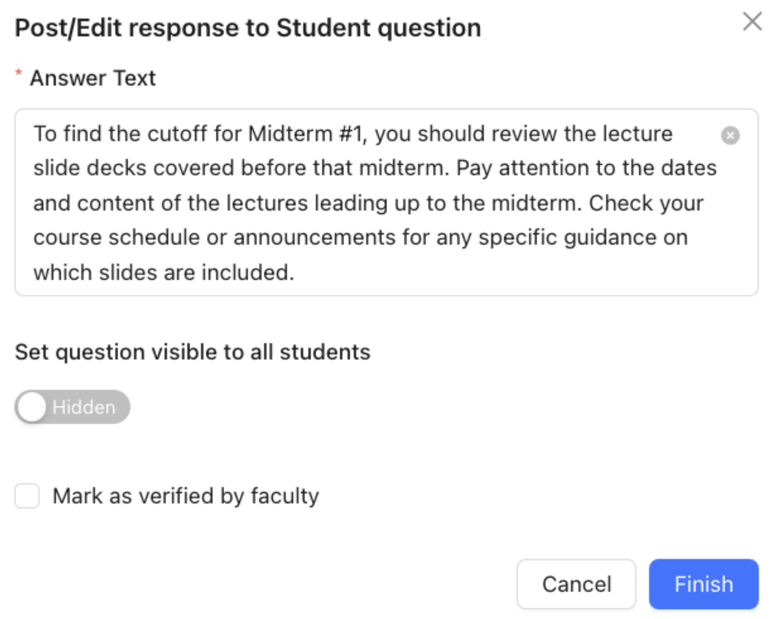
Jasleen Maan 5 days ago · Private · Human Verified · 1

Midterm #1 Questions

I have written out all of my course notes on my iPad on GoodNotes and I also have the app downloaded on my laptop. Since we are only...

Exams

Figure 5.4: Anytime Question Instructor View



Post/Edit response to Student question

* Answer Text

To find the cutoff for Midterm #1, you should review the lecture slide decks covered before that midterm. Pay attention to the dates and content of the lectures leading up to the midterm. Check your course schedule or announcements for any specific guidance on which slides are included.

Set question visible to all students

Hidden

□ Mark as verified by faculty

Figure 5.5: Instructor Editing an AI Response

- AI response satisfaction rate (thumbs up/down, or “Still need help?”).
- Percentage of questions flagged for human intervention.
- Times when the question was fully “AI resolved” vs. escalated.

The data also tracked user interactions with the chat widget and the queue to assess whether immediate AI responses reduced queue load.

5.2.1 Survey for Integrated System

At the conclusion of the course, a comprehensive survey was administered to capture qualitative feedback on:

- Overall HelpMe usage (in-person vs. virtual).
- Comparisons of AI-driven asynchronous help to traditional email or office hours.
- System Usability Scale (SUS) ratings [6].
- Perceptions of AI accuracy, reliability, tone, and trustworthiness.

- Preferred tools (in-house chatbot vs. external ChatGPT).
- Suggestions for improvement and additional features.

Survey Administration. The survey was distributed via Qualtrics in the final week of the course. Participation was voluntary, with anonymized responses.

5.3 Results

5.3.1 Chatbot Preference

Figure 5.6 shows how frequently students reported using the standalone chatbot. Many tried it just a few times, typically for quick clarifications.

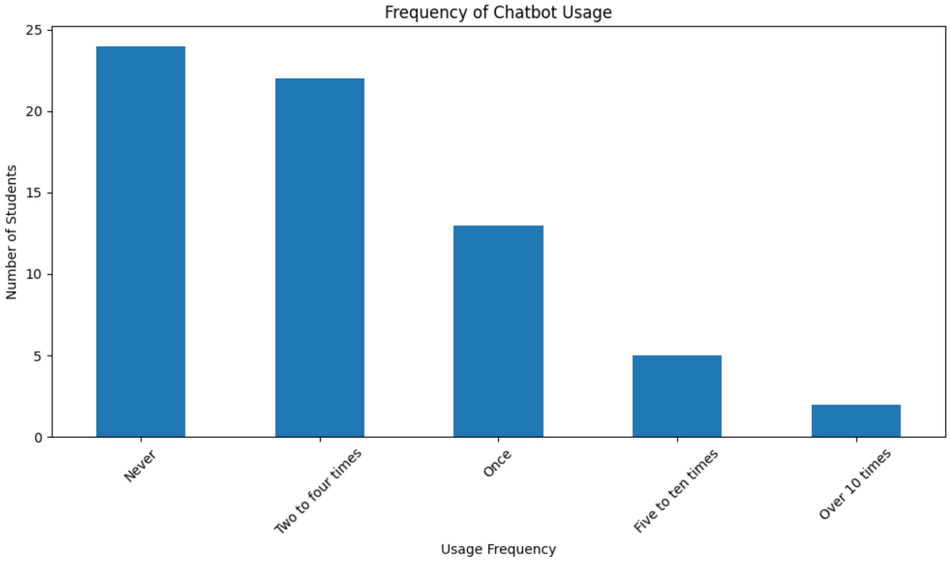


Figure 5.6: Chatbot Usage Frequency

5.3.2 Anytime Question Preference

Figure 5.8 show usage frequency and students' eventual preference after trying the Anytime Question Hub and the Chatbot. Overall, students found the Anytime Question Hub to be better than chatbot, despite both having the same generation pipeline. However, Anytime Question usage is much less than that of the chatbot (Figure 5.6) as seen in Figure 5.7.

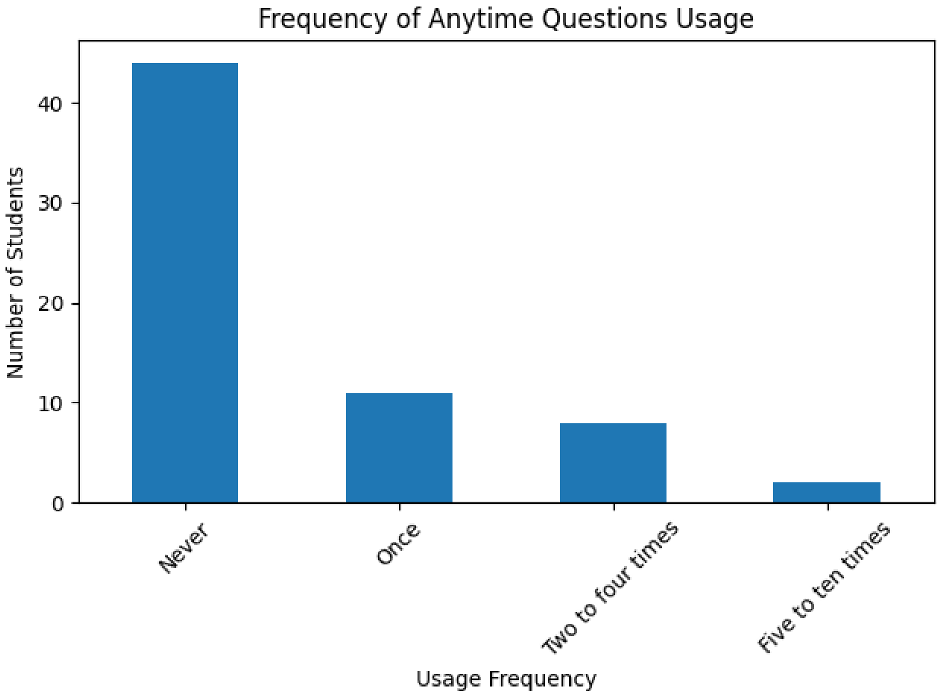


Figure 5.7: Anytime Question Hub Usage Frequency

5.3.3 Illustrative Examples

Example 1 (AI-Only Resolved)

Question Title: "Thursday's Class"

Student Query (Excerpt): *"Is Thursday's lecture only group presentations, or is there a final exam review as well?"*

AI's Answer (Unedited): *"Thursday's class includes project presentations, but it also features a final exam review. Make sure to prepare for both aspects!"*

Resolution: The student found the AI's response sufficient and did not request further human intervention. The question was marked "AIAnsweredResolved."

Example 2 (Human-Edited Resolution)

Question Title: "Why do I receive an error on Prairie Learn?"

Student Query (Excerpt): *"I have a relational algebra question in PrairieLearn LV5. I tried $\pi cname, city((customer) \bowtie balance > 2000(deposit))$, but I get an error. I believe it is correct. Why is there an error?"*

AI's Original Answer (Excerpt): *"Your answer is close, but the syntax needs adjustment. In relational algebra, you should place the selection before the join. For instance:*
$\pi_{cname,city}\big(customer \bowtie \sigma_{balance>2000}(deposit)\big).$
Make sure you check PrairieLearn's syntax requirements as well!

Human-Edited (Final) Answer: *"Your answer is on the right track, but remember to do the selection (σ) on `balance > 2000` before joining with `customer`. Try:*
$\pi_{cname,city}\big(customer \bowtie \sigma_{balance>2000}(deposit)\big).$
Also confirm that you have spelled all attributes correctly. If errors persist, see a TA during office hours to verify your syntax in PrairieLearn."

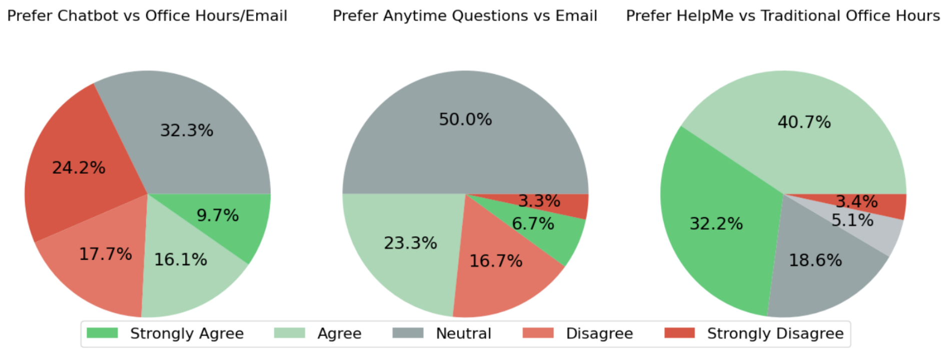


Figure 5.8: Preference After Using the Tools

5.3.4 Student Survey Findings

This subsection reports on students who reported using the system. Figure 5.8 shows a comparison of their preferences between the chatbot and the Anytime Question feature.

Overall, we observe that the Anytime Question Hub option is more favourable than chatbot questions. Although opinions vary, several students indicated that while they appreciate the escalated resolution workflow, they remain skeptical of AI's technical accuracy without instructor oversight. This feedback underscores a need for further refinements to automated help systems in educational contexts. The queue interface was confirmed to be highly favourable, consistent with findings in Section 3.3

The survey overall showed mixed sentiments regarding AI-generated help:

- **Anytime Question Hub vs. Chatbot:** Students who used both features indicated a preference for the asynchronous "Anytime Question" option over an all-AI chatbot. Many cited the reliability of eventual human oversight as a key reason for trust. Figure 5.8 illustrates overall usage and preference trends.

- **Reasons for Non-Usage:** Several students admitted they were unaware of the feature or simply defaulted to office-hour queues due to habit. Others expressed skepticism about AI accuracy but liked the escalation option.

While most respondents found the system easy to navigate, most negative opinions are regarding about the AI's correctness on technical details. Overall, the guarantee of human intervention when needed appeared to bolster trust and willingness to try the platform.

5.3.5 Preliminary Staff Feedback

Two TAs provided informal feedback:

> *"They [the AI responses] were good for the most part; the LLM seemed to handle the content accurately and I don't recall it completely fabricating details."* (TA1)

> *"Having the AI's draft already there saved me time. I'd often just correct a few details or add specifics. The interface was also straightforward."* (TA2)

However, TAs raised suggestions for improvement:

- **Markdown and Email Alerts:** Clearer instructions on formatting (e.g., markdown support) and better handling of automated notifications could improve usability.
- **Staff Verification Labeling:** Renaming "Faculty Verified" to "Staff Verified" would clarify that TAs can also finalize answers.
- **Reference & Linking:** A feature to link related questions or revert to original AI responses if needed would streamline oversight.

Overall, TAs described the system as "intuitive," noting it reduced repetitive email inquiries and facilitated consistent answers. The ability to verify or correct AI drafts led to efficient resolution of escalated questions, but they also emphasized continued refinements for better user experience.

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

The rapid advancement of Large Language Models (LLMs) and retrieval-augmented generation (RAG) has opened new possibilities for AI-powered educational tools. This thesis explores the potential of these technologies in academic settings, particularly in addressing challenges associated with student help-seeking, instructional efficiency, and course-specific knowledge dissemination. Through the development and evaluation of **ChatEd** and the **Integrated HelpMe system**, we have demonstrated that AI-driven solutions can provide meaningful assistance to students while maintaining instructor oversight and adaptability.

A major contribution of this work is the integration of RAG into educational chatbots, ensuring that student queries are answered with course-specific accuracy. Unlike general-purpose LLMs that may generate factually incorrect information, ChatEd provides responses grounded in instructor-provided materials. By leveraging structured course content, including lecture slides, policies, and assignment descriptions—ChatEd mitigates common AI pitfalls and enhances response reliability.

Benchmarking multiple LLMs, including GPT-4, LLaMA, and Phi, allowed for a comprehensive evaluation of retrieval-augmented architectures in an academic setting. Our findings indicate that while open-source LLMs can match or surpass proprietary models in accuracy when paired with an effective RAG pipeline, cost and infrastructure requirements remain key considerations. The study explored different retrieval mechanisms, question categorization strategies, and prompt engineering techniques, highlighting practical trade-offs in model performance, latency, and operational costs.

Another critical aspect of this research is the integration of AI into structured help-seeking workflows. Traditional student support mechanisms, such as office hours and email-based inquiries, are often inefficient, unstructured, and difficult to scale. The HelpMe system successfully streamlines student-instructor interactions by offering structured help queues, real-time tracking, and multi-modal assistance (live, asynchronous, and AI-supported). This thesis presents empirical evidence that such systems reduce instructor workload while improving response times and student satisfaction.

6.1.1 Key Takeaways

This research presents several key insights into the practical deployment of AI-driven educational assistants:

- AI chatbots are most effective when responses are grounded in course related sources.
- AI-driven assistance reduces student dependency on traditional support channels, such as email and office hours.
- Human-AI collaboration enhances trust and usability, as students are more willing to rely on AI-generated responses when human oversight is available.
- The trade-offs between fine-tuning and advanced prompt engineering require further exploration to determine the best approach for educational applications.
- Cost-performance considerations shape AI deployment in academia, balancing accuracy, latency, and infrastructure expenses.

This thesis establishes a framework for scalable, AI-powered student support and sets the foundation for continued exploration of LLM-driven educational tools. The following section outlines several promising directions for future research.

6.2 Future Directions

While this research has demonstrated the practical advantages of RAG-based educational AI, several open questions remain regarding scalability, personalization, and human-AI interaction dynamics. Future research should address the following areas described below.

6.2.1 Longitudinal Studies on Student Adoption and Learning Outcomes

This study primarily focuses on short-term AI adoption and usability, but the long-term impact of AI-assisted learning remains largely unexplored. Key research questions include:

- How do students' reliance on AI-driven assistants evolve across multiple semesters?
- Does ChatEd improve concept retention and problem-solving skills, or does it risk encouraging passive learning?
- How does AI-based tutoring compare to human-led instruction in long-term knowledge retention?

Longitudinal studies could track students across multiple courses, analyzing whether AI-assisted help-seeking enhances academic performance or shifts study habits in unexpected ways.

6.2.2 Adaptive AI and Personalized Learning

While ChatEd currently provides course-specific assistance, future iterations could personalize responses based on individual learning needs. Key research questions include:

- Exploring whether AI models can adjust responses based on student proficiency, providing tailored explanations.
- Investigating integration with Learning Management Systems (LMS) to provide adaptive feedback based on quiz performance and engagement data.
- Evaluating reinforcement learning techniques to optimize AI responses dynamically based on prior student interactions.

By leveraging student engagement data, AI-driven assistants could offer more personalized support, enhancing individualized learning experiences.

6.2.3 Fine-Tuning vs. Advanced Prompting

This research suggests that well-structured RAG pipelines can substitute for fine-tuned models, but further studies are needed to evaluate the comparative benefits of:

- Fine-tuned, domain-specific models trained on educational datasets.
- Enhanced retrieval accuracy.
- Advanced prompt engineering techniques to optimize generation quality without additional training.

Benchmarking these approaches across different subjects and learning environments would help institutions optimize their AI strategies.

6.2.4 Evaluating AI Transparency and Trust in Student-Facing Systems

Despite ChatEd's strong accuracy, some students remained skeptical of AI-generated answers. Key research questions include:

- The impact of transparency features, such as source citations and confidence scores, on student trust.
- The effectiveness of explainable AI (XAI) techniques in making chatbot reasoning clearer to students.
- How students balance AI-generated responses with instructor guidance in decision-making.

Developing explainable AI features for educational chatbots could significantly enhance user trust and adoption.

6.2.5 Scaling AI-Driven Assistance to Large Academic Institutions

This research was conducted within a controlled university setting, but questions remain about scaling AI-based support across multiple disciplines and institutions. Key research questions include:

- Evaluating ChatEd's effectiveness in non-CS subjects, such as humanities and medical education.
- Investigating multi-lingual retrieval systems to enhance accessibility for non-native English speakers.
- Analyzing ethical considerations in deploying AI for academic support at scale, particularly regarding data privacy and student autonomy.

Addressing these challenges would enable broader deployment of AI-driven educational assistants.

6.2.6 Final Thoughts

This thesis presents a rigorous investigation into AI-driven academic support, demonstrating that retrieval-augmented chatbots can effectively enhance student help-seeking, improve response accuracy, and reduce faculty workload. By combining structured retrieval with conversational AI, ChatEd and HelpMe establish a scalable framework for integrating AI into higher education.

However, as AI adoption in education continues to grow, it is important that future research explores long-term impacts, ethical considerations, and evolving pedagogical strategies. AI-assisted learning tools must complement human educators, ensuring that students receive high-quality, trustworthy, and context-aware guidance.

By refining AI-driven educational assistants and expanding their capabilities, we can move toward a future where AI supports more inclusive, personalized, and effective learning experiences for students worldwide.

Bibliography

[1] Sabah Abdul-Wahab, Nahed Salem, Kaan Yetilmezsoy, and Sulaiman Fadlallah. Students' Reluctance to Attend Office Hours: Reasons and Suggested Solutions. *Journal of Educational and Psychological Studies [JEPS]*, 13:715, 10 2019. → pages 12

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. → pages 5

[3] Saleema Amershi, Maya Cakmak, W Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014. → pages 7

[4] D Badulescu, SA Bodog, S Dzitac, CV Toca, and A Badulescu. Students' Perception Regarding the Impact of AI in Education and Career Prospects. In *ICERI2024 Proceedings*, pages 9640–9645. IATED, 2024. → pages 7

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. → pages 5

[6] John Brooke. SUS: A quick and dirty usability scale. In *Usability Evaluation in Industry*, pages 189–194. Taylor & Francis, 1996. → pages 11, 14, 38

[7] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024. → pages 8

[8] Elizabeth B. Cloude, Ryan S. Baker, and Eric Fouh. Online Help-Seeking Occurring in Multiple Computer-Mediated Conversations Af-

fects Grades in an Introductory Programming Course. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, page 378–387, New York, NY, USA, 2023. ACM. → pages 4

[9] Sam Cunningham-Nelson, Wageeh Boles, Luke Trouton, and Emily Margerison. A review of chatbots in education: practical steps forward. In *30th Annual conference for the Australasian Association for Engineering Education (AAEE 2019): Educators becoming agents of change: innovate, integrate, motivate*, pages 299–306. Engineers Australia, 2019. → pages 7

[10] Jacob Devlin. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. → pages 5

[11] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv:2309.15217*, 2023. → pages 8, 23

[12] Carlton Fong, Cassandra Gonzales, Christie Hill-Troglin Cox, and Holly Shinn. Academic help-seeking and achievement of postsecondary students: A meta-analytic investigation. *Journal of Educational Psychology*, 115(1):1, 2021. → pages 4

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, 2024. → pages 7

[14] Zhikai Gao, Sarah Heckman, and Collin Lynch. Who Uses Office Hours? A Comparison of In-Person and Virtual Office Hours Utilization. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1*, SIGCSE 2022, page 300–306, New York, NY, USA, 2022. ACM. → pages 4

[15] Zhikai Gao, Collin Lunch, Sarah Heckman, and Tiffany Barnes. Automatically classifying student help requests: a multi-year analysis. In *The 14th International Conference on Educational Data Mining*, pages 81–92. ERIC, 2021. → pages 4

[16] Chuqin Geng, Yihan Zhang, Brigitte Pientka, and Xujie Si. Can ChatGPT Pass An Introductory Level Functional Language Programming Course?, 2023. → pages 7

[17] Ashok Goel. AI-powered learning: making education accessible, affordable, and achievable. *arXiv preprint arXiv:2006.01908*, 2020. → pages 6

[18] Ashok K Goel and Lalith Polepeddi. Jill Watson. *Learning engineering for online education: Theoretical contexts and design-based examples. Routledge*, 2018. → pages 6

[19] Exploding Gradients. RAGAS: Retrieval Augmented Generation for Answer Synthesis. `https://github.com/explodinggradients/ragas`, 2023. → pages 30

[20] Braxton Hall, Noa Heyl, Elisa Baniassad, Meghan Allen, and Reid Holmes. The Efficacy of Online Office Hours: An Experience Report. In *Proceedings of the 2021 ACM SIGPLAN International Symposium on SPLASH-E*, SPLASH-E 2021, page 59–64, New York, NY, USA, 2021. ACM. → pages 4

[21] Elkafi Hassini. Student–instructor communication: The role of email. *Computers & Education*, 47(1):29–40, 2006. → pages 5

[22] Isabel Kathleen Fornell Haugeland, Asbjørn Følstad, Cameron Taylor, and Cato Alexander Bjørkli. Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies*, 161:102788, 2022. → pages 7

[23] Lydia Eckstein Jackson and Aimee Knupsky. Weaning off of Email: Encouraging Students to Use Office Hours over Email to Contact Professors. *College Teaching*, 63(4):183–184, 2015. → pages 5

[24] Karin Jensen, Jennifer R. Amos, Lawrence Angrave, Karle Flanagan, David Mussulman, Christopher D. Schmitz, and Wade Fagen-Ulmschneider. Adoption of an Online Queue App for Higher Education: A Case Study. In *2019 ASEE Annual Conference & Exposition*, number 10.18260/1-2–32042, Tampa, Florida, June 2019. ASEE Conferences. https://peer.asee.org/32042. → pages 5

[25] Onur Karademir, Daniele Di Mitri, Jan Schneider, Ioana Jivet, Jörn Allmang, Sebastian Gombert, Marcus Kubsch, Knut Neumann, and Hendrik Drachsler. I don't have time! But keep me in the loop: Co-designing requirements for a learning analytics cockpit with teachers. *Journal of Computer Assisted Learning*, 2024. → pages 7

[26] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023. → pages 1

[27] Shao-Heng Ko and Kristin Stephens-Martinez. What Drives Students to Office Hours: Individual Differences and Similarities. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2023, page 959–965, New York, NY, USA, 2023. ACM. → pages 4

[28] Chokri Kooli. Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15(7):5614, 2023. → pages 1

[29] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018, 2023. → pages 7

[30] LangChain. LangChain Documentation - Introduction, 2023. Accessed: 2024-07-11. → pages 6

[31] Li Lei and Pitts Jennifer P. Does It Really Matter? Using Virtual Office Hours to Enhance Student-Faculty Interaction. *Journal of Information Systems Education*, 20:175–186, 2009. → pages 4

[32] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. → pages 6, 7

[33] Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. Teaching CS50 with AI: Leveraging Generative Artificial Intelligence in Computer Science Education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 750–756, New York, NY, USA, 2024. ACM. → pages 7

[34] Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. RECALL: A Benchmark for LLMs Robustness against External Counterfactual Knowledge. *arXiv preprint arXiv:2311.08147*, 2023. → pages 8

[35] Tommy MacWilliam and David J. Malan. Scaling Office Hours: Managing Live Q&A in Large Courses. *J. Comput. Sci. Coll.*, 28(3):94–101, Jan 2013. → pages 5

[36] David J. Malan. Virtualizing Office Hours in CS 50. *SIGCSE Bull.*, 41(3):303–307, Jul 2009. → pages 4

[37] Kamil Malinka, Martin Peresni, Anton Firc, Ondrej Hujnk, and Filip Janus. On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education*, page 47–53. ACM, 2023. → pages 7

[38] Rosario Michel-Villarreal, Eliseo Vilalta-Perdomo, David Ernesto Salinas-Navarro, Ricardo Thierry-Aguilera, and Flor Silvestre Gerardou. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. *Education Sciences*, 13(9):856, 2023. → pages 1

[39] David Mussulman, Karin Jensen, Jennifer R. Amos, Lawrence Angrave, Karle Flanagan, Wade Fagen-Ulmschneider, Natalia Ozymko, Rittika Adhikari, and Jacqueline Osborn. Measuring Impact: Student and Instructor Experience Using an Online Queue. In *2020 ASEE Virtual Annual Conference Content Access*, number 10.18260/1-2–34961. ASEE Conferences, June 2020. https://peer.asee.org/34961. → pages 5

[40] Subash Neupane, Elias Hossain, Jason Keith, Himanshu Tripathi, Farbod Ghiasi, Noorbakhsh Amiri Golilarz, Amin Amirlatifi, Sudip Mittal, and Shahram Rahimi. From Questions to Insightful Answers: Building an Informed Chatbot for University Resources, 2024. → pages 6, 7, 24

[41] OpenAI. GPT-4 Technical Report, 2024. → pages 7

[42] Khoury College Sandbox. Khoury Office Hours, 2019. → pages 5

[43] Jiajing G. Shi. *Office Hours: A UX Investigation on How Might We Improve the Remote Office Hours Experience in Higher Education through*

Design Thinking. PhD thesis, University of Toronto (Canada), 2021. → pages 4

[44] Aaron J. Smith, Kristy Elizabeth Boyer, Jeffrey Forbes, Sarah Heckman, and Ketan Mayer-Patel. My Digital Hand: A Tool for Scaling Up One-to-One Peer Teaching in Support of Computer Science Learning. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '17, page 549–554, New York, NY, USA, 2017. ACM. → pages 5, 10, 12

[45] Margaret Smith, Yujie Chen, Rachel Berndtson, and Kristen M. Burson. "office hours are kind of weird": Reclaiming a resource to foster student-faculty interaction. *InSight: A Journal of Scholarly Teaching*, 12:14–29, 2017. → pages 4, 5, 12

[46] David H Smith IV, Qiang Hao, Vanessa Dennen, Michail Tsikerdekis, Bradly Barnes, Lilu Martin, and Nathan Tresham. Towards Understanding Online Question & Answer Interactions and their effects on student performance in large-scale STEM classes. *International Journal of Educational Technology in Higher Education*, 17:1–15, 2020. → pages 4

[47] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. → pages 23

[48] Karan Taneja, Pratyusha Maiti, Sandeep Kakar, Pranav Guruprasad, Sanjeev Rao, and Ashok K. Goel. Jill Watson: A Virtual Teaching Assistant powered by ChatGPT, 2024. → pages 6, 7, 24

[49] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for Fact Extraction and VERification, 2018. → pages 8

[50] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. → pages 5

[51] Mickey Vellukunnel, Philip Buffum, Kristy Elizabeth Boyer, Jeffrey Forbes, Sarah Heckman, and Ketan Mayer-Patel. Deconstructing the Discussion Forum: Student Questions and Computer Science Learning. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '17, page 603–608, New York, NY, USA, 2017. ACM. → pages 4

[52] Kevin Wang, Seth Akins, Abdallah Mohammed, and Ramon Lawrence. Student Mastery or AI Deception? Analyzing ChatGPT's Assessment Proficiency and Evaluating Detection Strategies, 2023. `https://arxiv.org/abs/2311.16292`. → pages 7

[53] Kevin Wang and Ramon Lawrence. Using Assignment Incentives to Reduce Student Procrastination and Encourage Code Review Interactions. In *2023 International Conference on Computational Science and Computational Intelligence Research Track on Education*, volume 1. IEEE, 2023. → pages 15

[54] Kevin Wang and Ramon Lawrence. HelpMe: Student Help Seeking using Office Hours and Email. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education - Volume 1*. ACM, 2024. → pages 9, 21

[55] Kevin Wang, Jason Ramos, and Ramon Lawrence. ChatEd: A Chatbot Leveraging ChatGPT for an Enhanced Learning Experience in Higher Education, 2023. `https://arxiv.org/abs/2401.00052`. → pages xi, 6, 7, 17, 18, 19, 20, 24

[56] Shuting Wang, Jiongnan Liu Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation. *arXiv preprint arXiv:2406.05654*, 2024. → pages 8

[57] Wei Xu. Toward human-centered AI: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019. → pages 7

[58] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. CRAG–Comprehensive RAG Benchmark. *arXiv preprint arXiv:2406.04744*, 2024. → pages 8

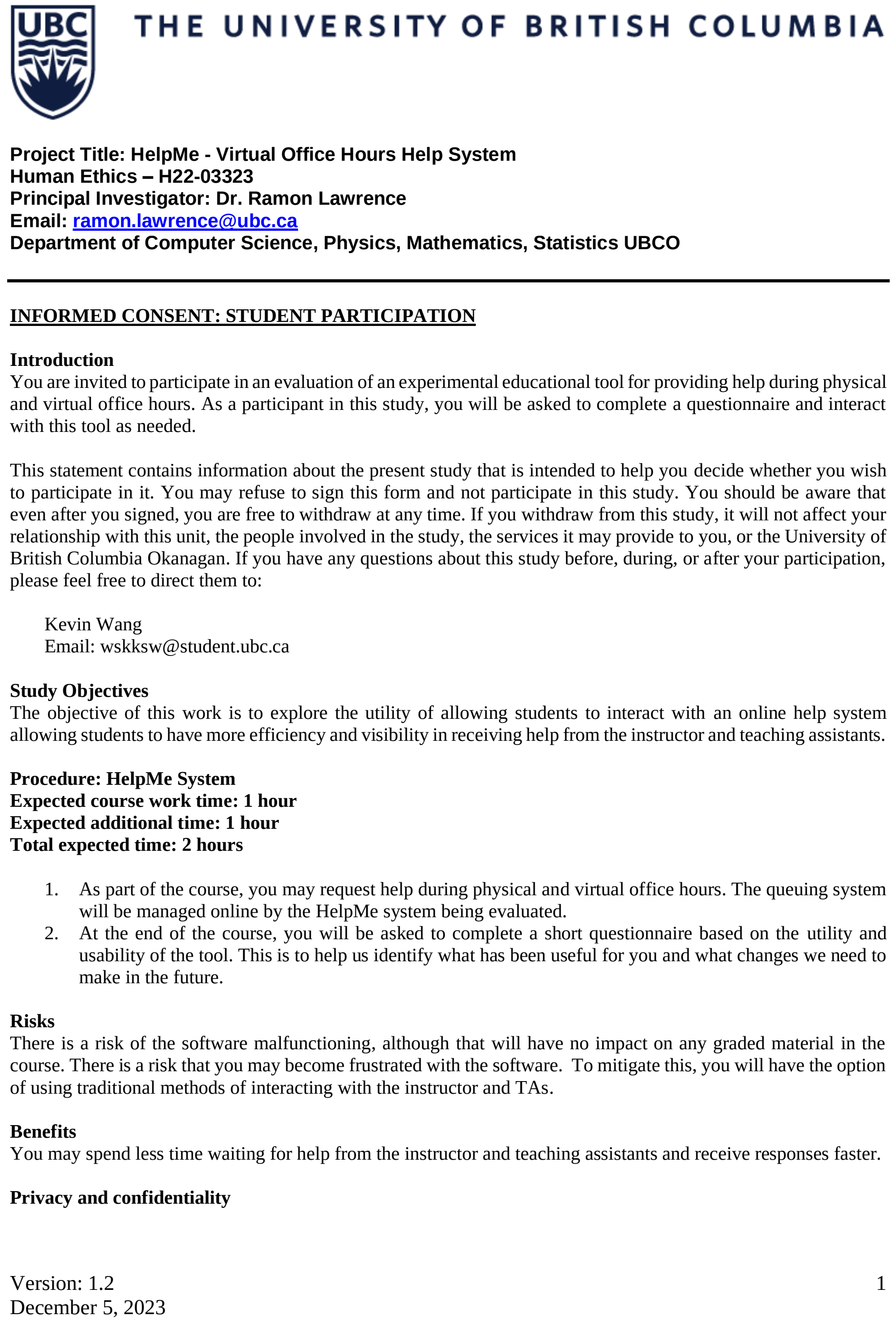
[59] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering, 2018. → pages 8

[60] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

Advances in Neural Information Processing Systems, 36, 2024. → pages 23

Appendices

A Consent Forms



THE UNIVERSITY OF BRITISH COLUMBIA

Project Title: HelpMe - Virtual Office Hours Help System
Human Ethics – H22-03323
Principal Investigator: Dr. Ramon Lawrence
Email: ramon.lawrence@ubc.ca
Department of Computer Science, Physics, Mathematics, Statistics UBCO

INFORMED CONSENT: STUDENT PARTICIPATION

Introduction

You are invited to participate in an evaluation of an experimental educational tool for providing help during physical and virtual office hours. As a participant in this study, you will be asked to complete a questionnaire and interact with this tool as needed.

This statement contains information about the present study that is intended to help you decide whether you wish to participate in it. You may refuse to sign this form and not participate in this study. You should be aware that even after you signed, you are free to withdraw at any time. If you withdraw from this study, it will not affect your relationship with this unit, the people involved in the study, the services it may provide to you, or the University of British Columbia Okanagan. If you have any questions about this study before, during, or after your participation, please feel free to direct them to:

Kevin Wang
Email: wskksw@student.ubc.ca

Study Objectives

The objective of this work is to explore the utility of allowing students to interact with an online help system allowing students to have more efficiency and visibility in receiving help from the instructor and teaching assistants.

Procedure: HelpMe System

Expected course work time: 1 hour
Expected additional time: 1 hour
Total expected time: 2 hours

1. As part of the course, you may request help during physical and virtual office hours. The queuing system will be managed online by the HelpMe system being evaluated.
2. At the end of the course, you will be asked to complete a short questionnaire based on the utility and usability of the tool. This is to help us identify what has been useful for you and what changes we need to make in the future.

Risks

There is a risk of the software malfunctioning, although that will have no impact on any graded material in the course. There is a risk that you may become frustrated with the software. To mitigate this, you will have the option of using traditional methods of interacting with the instructor and TAs.

Benefits

You may spend less time waiting for help from the instructor and teaching assistants and receive responses faster.

Privacy and confidentiality

Questionnaires are anonymous. Questionnaire data, consent data, and information collected from the system will all be stored separately and encrypted. Since the interest of this study is to identify average responses and behavior of the entire group of participants, you will not be identified in any way in any written reports of this research.

The data obtained from both components of the study will be retained for at least 5 years after publication in secured, electronic form, which only researchers associated with this research will have access to. No audio or video tapes will be used. General results of the research study may be available through publications. A summary of the results will be made available to participants upon request. Individual participant data will not be available.

Access to course work and associated grades

By checking the boxes below, you are providing consent for the researchers in this study to analyze your interactions with the HelpMe system and associated grades. The grades will only be used for research analysis purposes without including any of the participants' personal information.

Dual Role of Researcher

Since the PI is also the instructor for the course, any questions about the research study will be handled by graduate student Kevin Wang. Any significant issues will be handled by the Acting Department Head, Sylvie Desjardins, sylvie.desjardins@ubc.ca, (250) 807-8767.

Concerns about participants' rights

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics toll free at 1-877-822-8598 or the UBC Okanagan Research Services Office at 250-807-8832. It is also possible to contact the Research Complaint Line by email (RSIL@ors.ubc.ca). When doing so, include the study number H22-03323 when contacting their staff so they can better assist you.

Participant consent and signature

I have read this Consent form. I have had the opportunity to ask, and I have received answers to, any questions I had regarding the study and the use and disclosure of information about me for the study. I agree to take part in this study as a research participant. By my signature I affirm that I have received a copy of this Consent form.

□ Yes □ No	**I am willing to provide access to my information on using the HelpMe system.**
□ Yes □ No	**I am willing to provide access to my course grades.**

Whether or not you choose to participate, you will receive the same assignments and exam questions in this course. Your decision to participate or not participate in this study will in no way impact your grades or your relationship with UBC.

Participant Name: **Date:**

Participant Signature:

B Surveys

B.1 HelpMe Survey (2023 Spring)

THE UNIVERSITY OF BRITISH COLUMBIA

Project Title: HelpMe - Virtual Office Hours Help System
Human Ethics – H22-03323
Principal Investigator: Dr. Ramon Lawrence
Email: ramon.lawrence@ubc.ca
Department of Computer Science, Physics, Mathematics, Statistics UBCO

Participant ID: ________________________________ Date: ____________________________

Questionnaire - HelpMe - Virtual Office Hours Help System

Thank you for your voluntary participation in the HelpMe - Virtual Office Hours Help System research study under the direction of Dr. Ramon Lawrence.

In order to improve the software used in this study, this questionnaire has been developed to gather feedback regarding your experiences using the software. We value your honest and detailed responses. Your responses are confidential.

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1. I think I would like to use this tool frequently.	□	□	□	□	□
2. I found the tool unnecessarily complex.	□	□	□	□	□
3. I thought the tool was easy to use.	□	□	□	□	□
4. I think that I would need the support of a technical person to be able to use this system.	□	□	□	□	□
5. I found the various functions in this tool were well integrated.	□	□	□	□	□
6. I thought there was too much inconsistency in this tool.	□	□	□	□	□
7. I would imagine that most people would learn to use this tool very quickly.	□	□	□	□	□
8. I found the tool very cumbersome to use.	□	□	□	□	□
9. I felt very confident using the tool.	□	□	□	□	□
10. I needed to learn a lot of things before I could get going with this tool.	□	□	□	□	□

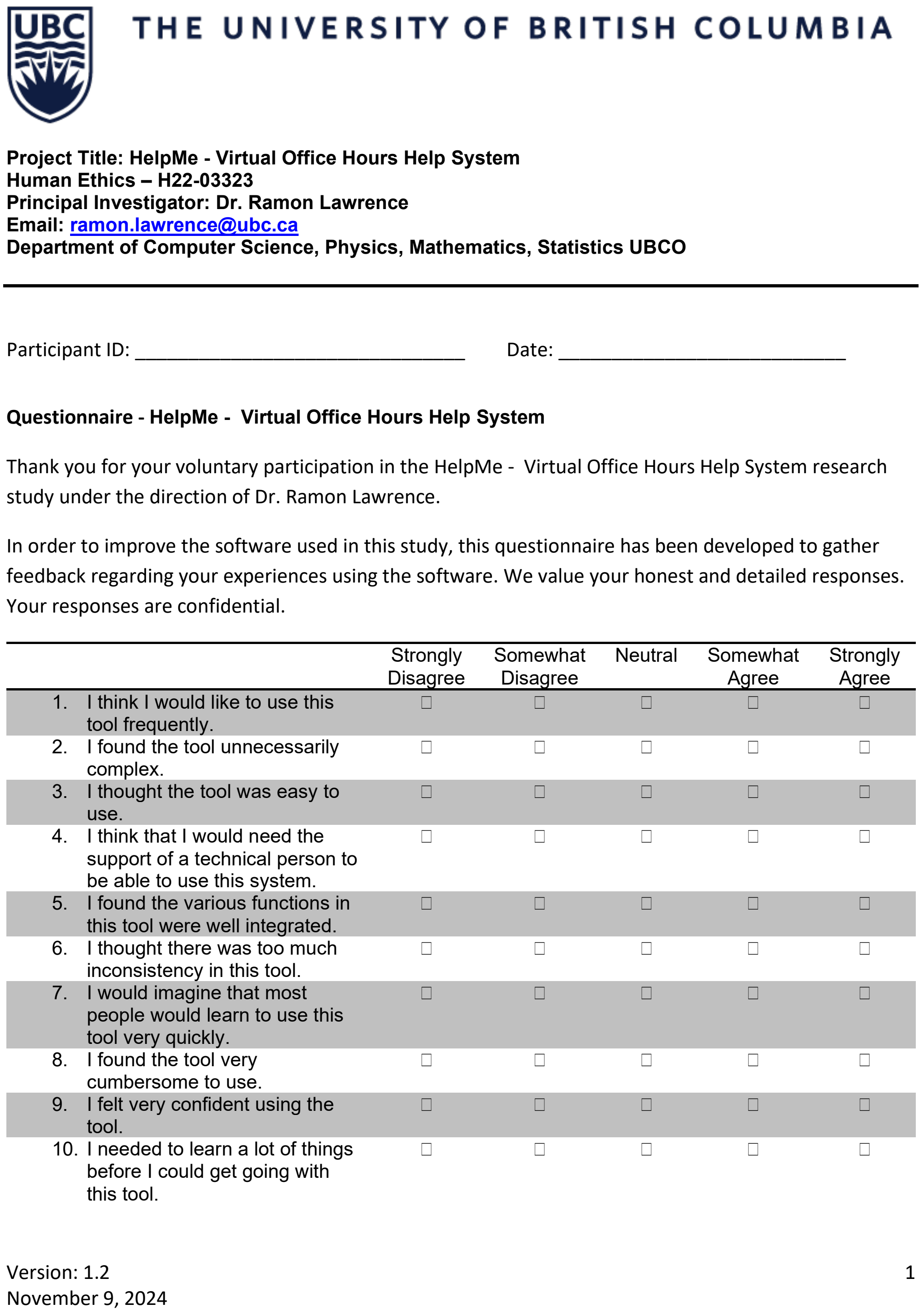
11. I attended more help sessions with HelpMe compared to traditional methods.	□	□	□	□	□
12. I prefer receiving help with HelpMe compared to traditional methods.	□	□	□	□	□
13. The response time with HelpMe was slower compared to traditional methods.	□	□	□	□	□
14. I had a better understanding of my wait time with HelpMe compared to traditional methods.	□	□	□	□	□
15. It was harder to use and get help with HelpMe compared to traditional methods.	□	□	□	□	□

16. What aspects of the software did you find helpful?

17. What additional features would you like to see with this software?

18. Please provide any general comments about the software and its use in the course.

B.2 HelpMe Survey (2024 Fall)



THE UNIVERSITY OF BRITISH COLUMBIA

Project Title: HelpMe - Virtual Office Hours Help System
Human Ethics – H22-03323
Principal Investigator: Dr. Ramon Lawrence
Email: ramon.lawrence@ubc.ca
Department of Computer Science, Physics, Mathematics, Statistics UBCO

Participant ID: ________________________________ Date: ____________________________

Questionnaire - HelpMe - Virtual Office Hours Help System

Thank you for your voluntary participation in the HelpMe - Virtual Office Hours Help System research study under the direction of Dr. Ramon Lawrence.

In order to improve the software used in this study, this questionnaire has been developed to gather feedback regarding your experiences using the software. We value your honest and detailed responses. Your responses are confidential.

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1. I think I would like to use this tool frequently.	□	□	□	□	□
2. I found the tool unnecessarily complex.	□	□	□	□	□
3. I thought the tool was easy to use.	□	□	□	□	□
4. I think that I would need the support of a technical person to be able to use this system.	□	□	□	□	□
5. I found the various functions in this tool were well integrated.	□	□	□	□	□
6. I thought there was too much inconsistency in this tool.	□	□	□	□	□
7. I would imagine that most people would learn to use this tool very quickly.	□	□	□	□	□
8. I found the tool very cumbersome to use.	□	□	□	□	□
9. I felt very confident using the tool.	□	□	□	□	□
10. I needed to learn a lot of things before I could get going with this tool.	□	□	□	□	□

Version: 1.2
November 9, 2024

1

11. I found the HelpMe system was an effective single location to go to for the help I needed.	□	□	□	□	□

12. What aspects of the software did you find helpful?

13. What additional features would you like to see with this software?

14. Please provide any general comments about the software and its use in the course.

Chatbot Questions

1. How often did you use the chatbot?

() Never () Once () 2 to 4 times () 5 to 10 times () More than 10 times

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
2. I am willing to use a chatbot for help in a course.	□	□	□	□	□
3. Chatbot answers were relevant and accurate.	□	□	□	□	□
4. Chatbot conversation was human-like.	□	□	□	□	□
5. The chatbot including source material links was helpful.	□	□	□	□	□
6. Did the chatbot assist you in meeting your learning outcomes?	□	□	□	□	□
7. Chatbot should be used in other courses.	□	□	□	□	□
8. I would prefer to use a course-provided Chatbot rather than a private option, like ChatGPT.	□	□	□	□	□
9. I would rather use the Chatbot than use office hours or email to communicate with teaching staff.	□	□	□	□	□
10. I am willing to use AI tools for course assistance.	□	□	□	□	□
11. I would recommend a struggling peer to use the Chatbot for assistance.	□	□	□	□	□

12. Please provide any general comments about the chatbot and its use in the course.

Anytime Question Hub

1. How often did you use the Anytime question feature?

() Never () Once () 2 to 4 times () 5 to 10 times () More than 10 times

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
2. I prefer email over asking anytime questions via HelpMe.	□	□	□	□	□
3. Public anytime question answers were helpful.	□	□	□	□	□
4. A single place for help (office hours, chatbot, anytime questions) is beneficial.	□	□	□	□	□
5. How would you rate the quality and relevance of the answers provided by the AI answer in Anytime question?	(Excellent, Very Good, Average, Poor)				

6. Please provide any general comments regarding how Anytime question support compares to email, the AI component or traditional, in-person office hours.