

CHATED: A CHATBOT LEVERAGING CHATGPT FOR AN ENHANCED LEARNING EXPERIENCE IN HIGHER EDUCATION

K. Wang, J. Ramos, R. Lawrence

University of British Columbia (CANADA)

Abstract

With the rapid evolution of Natural Language Processing (NLP), Large Language Models (LLMs) like ChatGPT have emerged as powerful tools capable of transforming various sectors. Their vast knowledge base and dynamic interaction capabilities represent significant potential in improving education by operating as a personalized assistant. However, the possibility of generating incorrect, biased, or unhelpful answers are a key challenge to resolve when deploying LLMs in an education context. This work introduces an innovative architecture that combines the strengths of ChatGPT with a traditional information retrieval chatbot framework to offer enhanced student support in higher education. Our empirical evaluations underscore the high promise of this approach.

Keywords: large language model, chatbot, ChatGPT, information retrieval, educational personal assistant.

1 INTRODUCTION

Large Language Models (LLMs) like ChatGPT present a new frontier in education with their abilities to transform education by providing personalized learning and generating and answering questions. However, there are known challenges including concerns about misinformation, a lack of domain-specific expertise, and potential biases ([1], [2], [3]). Accuracy and specificity are paramount in education.

Virtual assistants tailored for educational contexts have emerged. A virtual assistant named Jill Watson trained the chatbot on previous and ongoing Q&A to provide customized assistance for courses ([4]). Prior to LLMs, issues with deploying chatbots included significant overhead associated with training the models for every new course, a dependency on vast and accurate data, and their often limited conversational scope. Prior virtual assistants were confined to data they started with unlike the vast amount of data on which LLMs are trained on.

Given these challenges, we designed and implemented a new chatbot architecture for education called ChatEd, which is retrieval-based and integrated with a large language model such as ChatGPT. Retrieval-based models, traditionally employed by specifically trained chatbots, select the most appropriate document (chunk of information) from a predefined set of documents (usually derived from course materials) based on the user's input, which ensures validity and relevancy. Instead of returning only the reference document and answer, the document is sent to a LLM for generating the query response. This integration with the LLM introduces the benefits of human-like conversations, contextual understanding, and depth of conversation, while preserving the key benefits of traditional information retrieval by ensuring accuracy of responses and verified references to original source documents. The combination has better usability than traditional retrieval chatbots and improved accuracy and source verification compared to using only an LLM. Providing original sources allows students to locate the information. The key system features include:

- **Enhanced Accuracy and Contextual Relevance:** By combining the power of NLP with a rich knowledge base, the system delivers contextually accurate responses.
- **Conversational Memory** The model is designed to remember and reference chat history, allowing for deeper and more meaningful interactions.
- **Ease of Integration:** A user-friendly interface allows seamless integration with existing Learning Management Systems (LMS) and leverages existing course materials to ensure accurate and educational context-specific answers.
- **Streamlined and Easy Training:** Unlike previous chatbot approaches, there is no training required on the Q&A data. The system leverages existing course materials for a more efficient chatbot building process.

The primary contribution is introducing and evaluating a new chatbot framework for educational use. We show that the system has a very high level of question answering ability through leveraging both a retrieval-based framework and ChatGPT. A distinctive feature of our model is its scalability and ease of use for instructors. Instead of relying on traditional and labor-intensive methods of chatbot training, our model facilitates the direct input of existing course materials. This approach streamlines the setup process and ensures that the chatbot's knowledge base is aligned with the course content.

A comprehensive series of evaluations is performed to assess the core framework of our proposed system, particularly its question-answering ability and context awareness. Evaluations are performed to measure if the question answers are helpful, relevant, accurate, and correct. Experiments show the ChatEd chatbot excels in all these criteria. ChatEd is compared with ChatGPT for answer quality and ability to support conversational interactions.

2 BACKGROUND

2.1 Large Language Models and Education

LLMs such as ChatGPT provide novel opportunities in education ([1], [2], [5]). These systems can provide personalized learning and function as a research assistant and 24/7 online instructor ([1], [2], [3]). However, LLMs have also been shown to generate incorrect, misleading, and biased information and rarely provide correct references to original sources. Misinformation is a significant concern in education, including false information or biased information. Another problem is that they leverage general information and lack domain specific knowledge, which poses problems especially in education as students need specific guidance in fields that is relevant to their particular course and instructional context ([2]).

2.2 Virtual Assistants in Higher Education

There is prior research on developing virtual assistants (teaching agents) for courses. Jill Watson, made by Ashok *et al.* ([4]), is one of the most renowned virtual assistants for educational questions. Jill Watson was originally designed for large MOOC courses using IBM Bluemix toolsuite. Experimentation demonstrated that Jill Watson is effective in reducing teacher workload, promotes engagement, and is malleable ([4]). Jill Watson is developed based on training of Q&A and a feedback loop, and it works within a discussion forum to answer questions that it is confident in answering. Through continuous developing and exploration, Jill Watson's versions became easier to train, and required under 24 hours to develop a new chatbot for a syllabi ([6]).

Cunningham *et al.* ([7]) covered the common chatbot frameworks used, such as Dialogflow, RASA, and Wit.ai. They described applications for chatbots in education such as a FAQ Chatbot, but also cite the drawbacks such as needing accurate data for training.

Kuhail *et al.* ([8]) conducted a comprehensive review of chatbots in education including their limitations. A common problem with chatbots is their limited capabilities to expand beyond the initial data they are trained on. Instructors can discuss the problem and answer new related questions. This is hard to do unless the chatbot is provided with an extremely large amount of data. By deploying Jill Watson in a MOOC, student Q&A was abundant and the cost of training a chatbot can be justified. Even so, the conversational ability is limited. The cost and challenges of training chatbots effectively blocks individual instructors from building specialized chatbots for their courses, as there is a bottleneck for both collecting the training data and the cost of training.

2.3 Bridging the Gap

The future of educational assistance lies in harnessing the power of LLMs while addressing their inherent shortcomings. We propose a novel system that marries the vast knowledge base of models like ChatGPT with the specific context provided by traditional virtual assistants. This synergy promises a chatbot that is both knowledgeable and specialized, offering students a comprehensive educational companion without the typical constraints of current models. It also offers the ability for instructors to develop and deploy chatbots for their own courses efficiently.

3 SYSTEM DESIGN

The ChatEd chatbot architecture is designed for easy deployment in courses by instructors and features a conversational, chat interface for students. The key requirements are:

- **Correctness** - chatbot must provide correct answers and communicate its limitations when a question is out of its knowledge scope.
- **Relevancy** - responses must be specific to the course material and educational context.
- **Verifiable** - responses must be verifiable by students by providing specific source information and references used to generate the response.

Students interact with the chatbot similar to other chatbot systems by asking questions, receiving responses, and continuing a conversation with follow up questions and requests (see Figure 1). The chatbot is a service that can be integrated with Learning Management Systems (LMSs), discussion and Q&A forums, or other student interaction software.

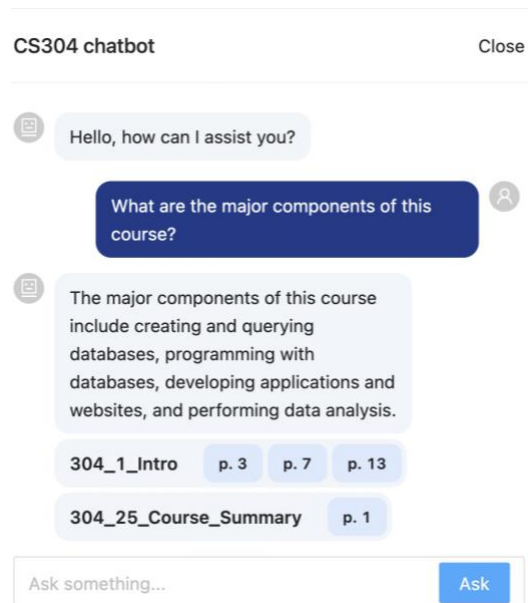


Figure 1. Chatbot User Interface and Example Conversation.

For instructors, the system provides an efficient mechanism for training on source material including notes (PowerPoint, PDF), web sites, posts and announcements, and other textual content. These source materials are processed and summarized in a database for efficient retrieval and integration with ChatGPT. Importantly, this approach eliminates any requirements for training a model by the instructor and leaves the challenging data integration and conversational task to the LLM, such as ChatGPT. Source content is provided to the system by uploading files or providing URLs of source locations. The instructor interface is in Figure 2.

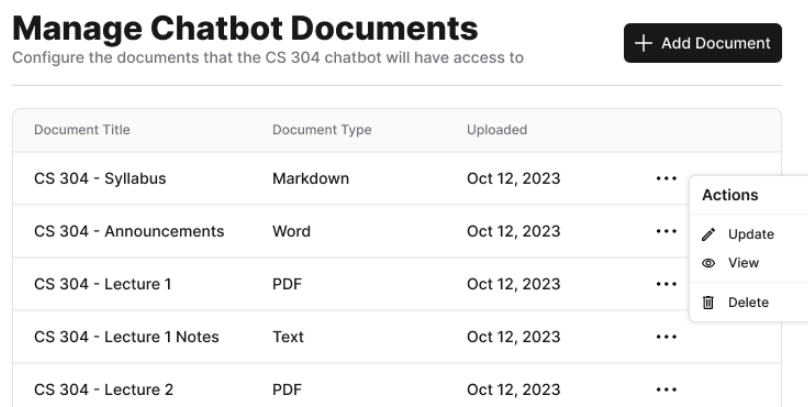


Figure 2. Instructor Interface.

3.1 ChatEd Chatbot Architecture

The unique aspect of the ChatEd architecture (Figure 3) is integrating an information retrieval system that stores and queries sources provided by the instructor with an LLM that provides the conversational support and general knowledge.

3.1.1 Context-Specific Database

The first step is for instructors to provide their sources as documents or URLs. Each document is retrieved and indexed. For example, a PDF document containing a PowerPoint presentation is divided and indexed by slide and keywords captured on each indexed element. These instructor documents provide the source context for the chatbot that is specific to the current course. The database storing these documents is PostgreSQL, and similarity search is performed using Facebook AI Similarity Search library (Faiss - <https://github.com/facebookresearch/faiss>). There is a separate database for each course or unique chatbot constructed.

3.1.2 LLM Integration

When a user poses a question, instead of sending the question directly to the LLM, which would respond using its generalized knowledge, the question is first used as a query in the database to determine similar indexed documents. Then, the question, indexed documents, and prior chat history are provided as a prompt to the LLM. This has the key advantages that the LLM has specific, contextualized knowledge to frame a response, and this response when provided back to the user can be augmented with document sources. The inclusion of chat history allows the LLM to maintain context, ensuring continuity in extended interactions. The LLM processes these inputs to generate a coherent and contextually relevant answer for the user.

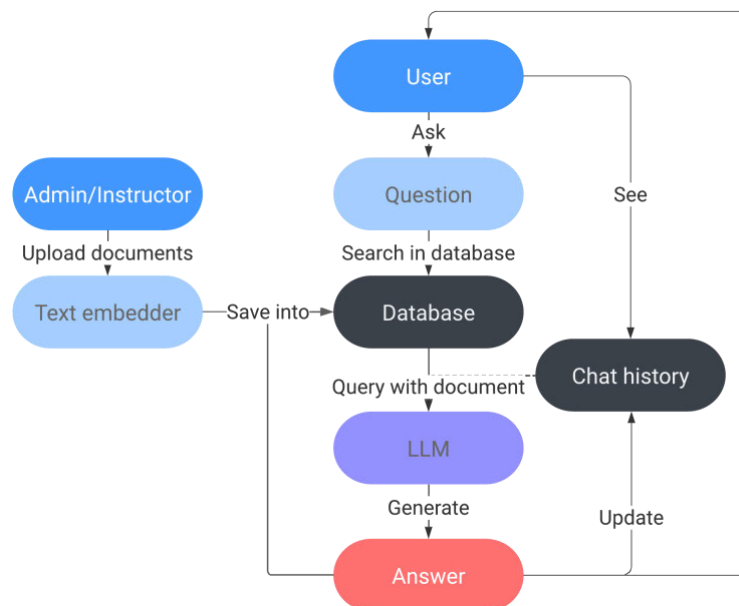


Figure 3. ChatEd Chatbot Architecture.

There is no local training required for the chatbot as the question answering and conversational aspects are provided by the LLM. The chatbot leverages the LangChain (<https://www.langchain.com/>) framework, which provides libraries that abstract the process of querying for relevant information and calling to Large Language Models (LLM). In our experiment, we used ChatGPT 3.5 turbo API.

4 METHODOLOGY

The ChatEd chatbot is evaluated for performance based on questions from two courses at a research-intensive university. The questions and answers were collected as part of a previous research study ([9]) evaluating the types of questions asked during office hours and over email. The first course is CS304: Introduction to Databases which covers basic database knowledge such as SQL and database design. The second course is CS 404: Database Implementation which covers database internals and

how to build a database system. The course materials for both courses are distributed on GitHub, including lecture slides, labs, and syllabi. The course materials were uploaded to the chatbot, and specific chatbots were created for each course.

The test data set consisted of a diverse variety of questions, including *general questions* such as 'Can each table have multiple foreign keys?', *domain-specific questions* such as 'Are "chunk" and "block" interchangeable terms?', and *managerial questions* such as 'Is the exam open-book?'. In total, there were 60 individual questions.

The evaluation performed 3 different types of tests that used the question text, expected answer manually added according to lecture content, actual chatbot system response, and human evaluation of the chatbot response. The evaluation was performed by the instructor and teaching assistants of the courses. The specific test types are outlined below. To benchmark the system, a side-by-side comparison was conducted with ChatGPT, assessing the responses, accuracy, and relevance of both systems to identical questions.

4.1 Question Answering Abilities

ChatEd's question-answering ability was measured using the question set with responses, and three metrics were used in the evaluation:

- **Relevance:** Evaluates the pertinence of ChatEd's responses to the question asked by the user. It assesses how well the response aligns with the specific query, ensuring that the answer directly addresses the main point of the question.
 - **Directness:** Does the response directly answer the user's question?
 - **Conciseness:** Is the response free from unnecessary wordiness or details?
 - **Contextual Awareness:** Does the response consider the previous questions or statements?
 - **Detail Tailoring:** Does the response address the nuances and specific details of the user's question?
 - **Topic Alignment:** Does the response stay within the bounds of the subject matter?
- **Accuracy:** Evaluates the extent to which ChatEd's responses provide factually correct information, particularly within the context of the specific topic or domain in question.
 - **Factual Correctness:** Does the response provide information that is correct based on the documents provided?
 - **Acknowledgment of Limitations:** Does the response transparently acknowledge any limitations in its knowledge base or ability to help?
 - **Logical Consistency:** Does the response contain conflicting statements or information?
 - **Completeness:** Is the response comprehensive and thorough, leaving no critical aspects of the user's query unaddressed?
 - **Information Source:** Does the response use any information not contained in the documents provided?
- **Helpfulness:** Evaluates how well ChatEd's responses address the user's question and how well it enhances the user's experience ensuring the overall utility of the response.
 - **Clarity:** Is the response presented in a clear and understandable manner, avoiding jargon or overly complex language?
 - **Empathy and Tone:** Is the response delivered in a friendly, respectful, and empathetic tone?
 - **Follow-Up:** Is the response more detailed when the user asks for clarifications on previously asked questions?
 - **Error Correction:** Does the response correct any inaccuracies or misconceptions in the user's question?
 - **Source Citation:** Does the response provide the sources that it referenced?

4.2 Context Awareness Test Procedure

The ability of ChatEd to understand, maintain, and switch contexts during interactions was tested using these steps:

- **Establish a Context (A):** Initiate with a context-setting question.
- **Follow-up Without Explicit Context (B):** Pose questions that implicitly rely on the established context.
- **Deepen the Context (C):** Delve deeper into the context with more specific questions.
- **Break the Context (D):** Interrupt the context sequence with an unrelated query.
- **Revisit the Context (E):** Return to the original context after the interruption.

These steps were executed in various sequences to test the flexibility and memory of the chatbot. Responses were documented and analyzed to ascertain the ChatEd's proficiency in context-aware interactions. In each course, five distinct interactions were executed to evaluate consistency and reliability.

5 RESULTS

5.1 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, with 5 being the best, based on three criteria: relevancy, accuracy, and helpfulness. The questions were also directly posed to ChatGPT v3.5 for comparison. The results are in Table 1 and show the average score over all questions. An example test interaction is in Figure 4.

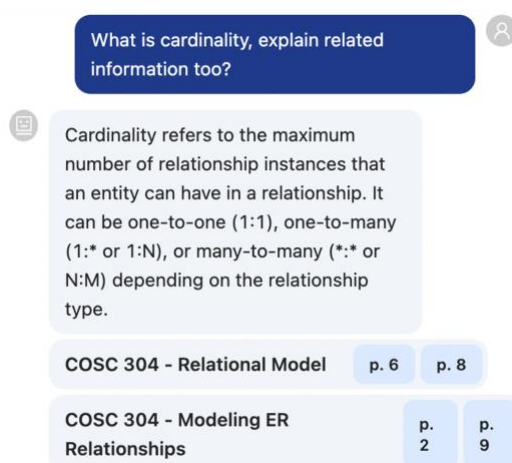


Figure 4. Example Interaction for Q&A Test.

Table 1. Average Scores Given by Evaluators for Q&A.

Criteria	ChatEd Score	ChatGPT Score
Relevancy	5	4.4
Accuracy	5	4.4
Helpfulness	4.5	3.4

ChatEd had exceptional performance in the Q&A test returning answers that were evaluated as highly relevant and accurate. The provided references to source material are helpful for students to gather more information on the subject. ChatEd answers questions regarding the course more accurately in the context of the course.

ChatGPT scored lower for reasons summarized below.

- ChatEd performs extremely well with managerial questions that LLMs cannot answer, such as policies of the course. The answers are specialized and contain information not available on the training data of ChatGPT.

- ChatEd also supplies more concise responses compared to ChatGPT, which often provides more verbose responses with more unrelated content.
- Comments by teaching assistant reviewers highlight ChatEd's ability to return more precise answers that are specific to the course content. For example, ChatGPT's answer to the question 'What are some errors when retrieving results? Explain them.' had 12 exhaustive bullet points while ChatEd provided the 5 common errors when retrieving database results covered in class.

5.2 Conversational Depth

The test for conversational depth evaluated ChatEd's ability to continue to provide relevant information as follow-up questions are asked by the user and adeptly respond when the user changes the question context during the conversation. The conversations were evaluated on a 5-point scale. The results are in Table 2. An example test interaction is in Figure 5. ChatEd displayed strong results in all criteria especially accuracy, highlighting its capabilities to relate to chat history.

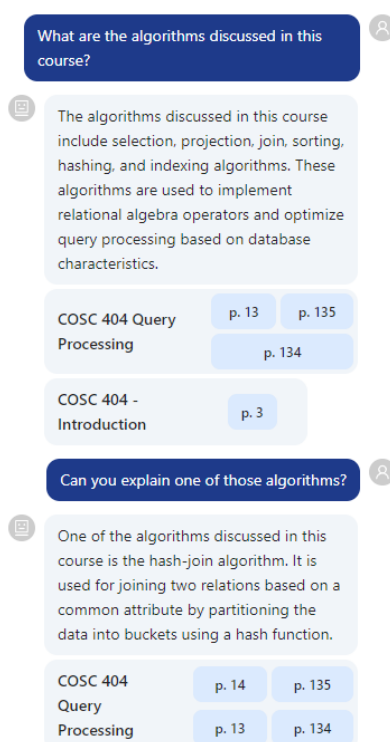


Figure 5. Example of Switching Contexts in a Conversation.

Table 2. Average Scores Given by Evaluators.

Criteria	ChatEd Score
Relevancy	4.19
Accuracy	5.00
Helpfulness	4.06

Although it had a strong ability to be aware of the context of the conversation, there is still room for improvement. For example, when asked "Why do we need this?" as a follow-up question to "What is a hash function?", ChatEd responded with the importance of a different topic. Although the answer was factually correct, ChatEd failed to understand what "this" was referring to. This resulted in an unhelpful and irrelevant response. We speculate that this is because the documents provided to ChatEd contained information about the importance of other topics. As it has access to both the history of the conversation and the relevant documents, ChatEd may have confused the documents with the original question.

Further, ChatEd is unable to consistently understand questions that refer to other questions that are too far back in the conversation. Without the context of the previous question, this results in ChatEd providing irrelevant or incomplete information. Overall, ChatEd shows promising potential as a conversational chatbot and will iteratively improve over time.

6 DISCUSSION

Prior research on educational chatbots has shown good success and educational impact, especially in large classes. A key challenge was the cost of training chatbots on course-specific material, and their ability to adapt to new questions. The major advantage of ChatEd is that it requires no training by the instructor, who must only provide the source documents to the system for indexing. By using the source material as a prompt to a LLM rather than training the chatbot itself, the system leverages the general knowledge and conversational ability of the LLM while preserving local educational context and source references. Students receive relevant and contextually appropriate answers, enhancing their learning experience.

The improvement of ChatEd compared to ChatGPT depends on the source material provided. ChatGPT does a good job answering general questions that have accepted answers available on the Internet. It has weaker performance on questions that are on course specific content or less commonly known, and especially on questions that involve the management of the course such as deadlines, policies, and administrative factors.

7 CONCLUSION AND FUTURE WORK

This work developed and evaluated a novel architecture for leveraging a LLM augmented with localized information for particular courses. By removing the need for local training of the chatbot, instructors can easily utilize the system in their own courses simply by providing the relevant materials. The evaluation performed on actual student questions showed high performance in the criteria of relevancy, accuracy, and helpfulness with performance improving upon ChatGPT that does not have the localized context information. Future work will evaluate the system on questions in more courses, improve the question answering ability, and make the system available to instructors for use.

REFERENCES

- [1] R. Thierry-Aguilera, and F. S. Gerardou, "Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT," *Education Sciences*, vol. 13, no. 9, p. 856, 2023.
- [2] C. Kooli, "Chatbots in education and research: A critical examination of ethical implications and solutions," *Sustainability*, vol. 15, no. 7, p. 5614, 2023.
- [3] E. Kasneci, K. Seßler, S. K'uchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. G'unnemann, E. H'üllermeier et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.
- [4] A. K. Goel and L. Polepeddi, "Jill Watson," *Learning engineering for online education: Theoretical contexts and design-based examples*. Routledge, 2018.
- [5] C. K. Lo, "What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature," *Education Sciences*, vol. 13, no. 4, 2023. Available: <https://www.mdpi.com/2227-7102/13/4/410>
- [6] A. Goel, "AI-powered learning: making education accessible, affordable, and achievable," *arXiv preprint arXiv:2006.01908*, 2020.
- [7] S. Cunningham-Nelson, W. Boles, L. Trouton, and E. 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*. Engineers Australia, 2019, pp. 299–306.
- [8] M. A. Kuhail, N. Alturki, S. Alamlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Education and Information Technologies*, vol. 28, no. 1, pp. 973–1018, 2023.

- [9] K. Wang and R. 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. [Online]. Available: <https://doi.org/10.1145/3626252.3630867>

APPENDIX

Table 3 shows a set of interactions with ChatEd. Table 4 contains 5 sample questions used to test ChatEd on single Q&As.

Table 3. Sample Conversation Interaction (rated 4/5/4 for relevance/accuracy/helpfulness).

Number	Question	Actual Response
1	What are transactions?	<p>Transactions are atomic programs that execute on a database and maintain the consistency of the database. They ensure that the database remains in a consistent state before and after their execution.</p> <p>COSC 404 - Transactions p. 4, 56</p> <p>COSC 404 - Concurrency p. 62</p> <p>COSC 404 - Scaling and Distribution p. 57</p>
2	Can you explain what a consistent state is?	<p>A consistent state in a database refers to a state where the data satisfies all the constraints specified in the database schema. It means that the data in the database is in accordance with the rules or predicates defined for it.</p> <p>COSC 404 - Transactions p. 5, 6, 58</p> <p>COSC 404 - Scaling and Distribution p. 15</p>
3	How does a transaction do that?	<p>A transaction ensures consistency by preserving the consistency of the database. It executes as an atomic program, meaning it either completes all its operations or none at all, ensuring that the database remains in a consistent state throughout the transaction.</p> <p>COSC 404 - Transactions p. 3, 4, 10</p> <p>COSC 404 - Concurrency p. 62</p>
4	What's an index?	<p>An index is a data structure that allows for fast lookup of records in a file. It can also help retrieve records in sorted order. Indexing is important for file systems and databases as it improves the efficiency of queries by minimizing the number of blocks accessed.</p> <p>COSC 404 Indexing p. 2, 4, 50, 51</p>
5	Can you further explain my first question?	<p>An index is a data structure that allows for quick retrieval of records in a file. It improves query efficiency by minimizing the number of blocks accessed. Without secondary indexes, answering the query would require scanning the entire file, resulting in slower performance. With only one secondary index, the query could still be answered efficiently by retrieving the buckets for the indexed attribute and comparing the records.</p> <p>COSC 404 Indexing p. 27</p> <p>COSC 404 - Query Optimization p. 10, 14</p> <p>COSC 404 - Transactions p. 16</p>

Table 4. Sample Questions and Responses from the Question Set with ratings on Relevance/Accuracy/Helpfulness

Question	ChatEd Response	ChatGPT Response
What are the major components of this course?	<p>The major components of this course include creating, querying, and programming with databases to develop applications, web sites, and perform data analysis. Additionally, the course covers database techniques and software including relational and NoSQL databases, SQL, JSON, and XML.</p> <p>COSC 304 - Course Introduction p. 3, 7, 13 COSC 304 - Course Summary p. 1</p> <p>Rating: 5 / 5 / 4</p>	<p>I'm sorry, but I need more specific information to provide you with an accurate answer. Could you please clarify which course you are referring to or provide more details about the subject or context of the course? Once I have more information, I'd be happy to help you identify the major components of that course.</p> <p>Rating: 1 / 1 / 1</p>
Are "chunk" and "block" interchangeable terms?	<p>No, "chunk" and "block" are not interchangeable terms. In the context of the resources provided, a "block" refers to a chunk of memory that can be read or written, while a "chunk" is not specifically defined in the given documents.</p> <p>404_2_Storage p. 33, 60, 64 404_12_Recovery p. 8</p> <p>Rating: 5 / 5 / 5</p>	<p>"Chunk" and "block" can be used interchangeably in some contexts, but they are not universally synonymous. The specific meaning of each term can vary depending on the domain or context. Here are a few examples: Memory and Cognitive Science: "Chunking" is a process where individual pieces of information are bound together into a meaningful whole. "Block" doesn't have the same... (word count: 226)</p> <p>Rating: 1 / 1 / 1</p>
When are the midterm exams?	<p>The midterm exams are on October 10th and November 2nd in class.</p> <p>Syllabus p. 2, 4, 5, 7</p> <p>Rating: 5 / 5 / 5</p>	<p>I'm sorry, but I don't have access to your specific course schedule or information about when your midterm exams are. To find out the exact date and time of your midterm exams, you should: Check your course syllabus: The syllabus provided by your instructor at the beginning of the course typically includes information about the exam dates. Visit the course... (word count: 203)</p> <p>Rating: 5 / 5 / 1</p>
How do I model a domain (explained in English) in an ER diagram?	<p>To model a domain in an ER diagram, you need to identify the entities involved in the domain and their relationships. Represent each entity as a rectangle and connect them with lines to represent the relationships between them.</p> <p>COSC 304 - Modelling ER p. 29 COSC 304 - Database Design and ER Modelling p. 1, 32 COSC 304 - ER to Relational Mapping p. 2</p> <p>Rating: 5 / 5 / 5</p>	<p>Modelling a domain in an Entity-Relationship (ER) diagram involves representing the key entities, attributes, and relationships within a specific area of interest or problem domain. Here's a step-by-step guide on how to do this in simple English: Identify the Main Entities: Start by identifying the main entities or objects within the domain. These are the things you want to represent... (word count: 424)</p> <p>Rating: 5 / 5 / 5</p>
Can each table have multiple foreign keys?	<p>Yes, each table can have multiple foreign keys.</p> <p>COSC 304 - Relational Model p. 18, 21 COSC 304 - SQL DDL p. 17, 18</p> <p>Rating: 5 / 5 / 5</p>	<p>Yes, a database table can have multiple foreign keys. Foreign keys are used to establish relationships between tables in a relational database. Each foreign key in a table references a primary key in another table, creating a link between the two tables. These relationships help maintain data integrity and allow for ... (word count: 251)</p> <p>Rating: 5 / 5 / 4</p>