# Using Cipherbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot

Joni Salminen
University of Vaasa
Vaasa, Finland
jonisalm@uwasa.fi

Soon-gyo Jung
Qatar Computing Research Institute, Hamad Bin Khalifa University
Doha, Qatar
sjung@hbku.edu.qa

Johanne Medina
Qatar Computing Research Institute, Hamad Bin Khalifa University
Doha, Qatar
jomedina@hbku.edu.qa

Kholoud Aldous
Qatar Computing Research Institute, Hamad Bin Khalifa University
Doha, Qatar
kkaldous@hbku.edu.qa

Jinan Azem
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
jazem@hbku.edu.qa

Waleed Akhtar
University of Vaasa
Vaasa, Finland
waleed.akhtar@uwasa.fi

Bernard J. Jansen
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
bjansen@hbku.edu.qa

## ABSTRACT

Cipherbot, an educational chatbot using large language models to answer student questions concerning learning materials uploaded by the educator, was pilot tested in a classroom setting. Forty-four students used Cipherbot for seven weeks, sending 8077 messages. The average number of messages sent per student was 184 (SD = 80), with an average length of 98 characters (SD = 80). The engagement followed a non-normal distribution, with few power users, implying that most students are still hesitant to adopt tools like Cipherbot. Cipherbot was able to answer 82.5% of the student questions, demonstrating a scalable ability to address students' learning queries, with some room for improvement.

## CCS CONCEPTS

**•Human-centered computing → Human computer interaction (HCI)**

## KEYWORDS

Cipherbot, LLMs, Generative AI, Student interaction

**ACM Reference Format:**

Joni Salminen, Soon-gyo Jung, Johanne Medina, Kholoud Aldous, Jinan Azem, Waleed Akhtar, and Bernard J. Jansen. 2024. Using Cipherbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3657604.3664690


*L@S '24, July 18–20, 2024, Atlanta, GA, USA*

ACM 979-8-4007-0633-2/24/07…
https://doi.org/10.1145/3657604.3664690

## 1 INTRODUCTION

*Cipherbot* allows teachers to upload study material and students to seek answers from these materials (see Figure 1), exemplifying the integration of Generative AI and large language models (LLMs) into education [2, 15]. However, research on student engagement with tools like Cipherbot remains scarce. This study provides findings on the quality and quantity of students' interaction with Cipherbot, investigating multiple aspects.

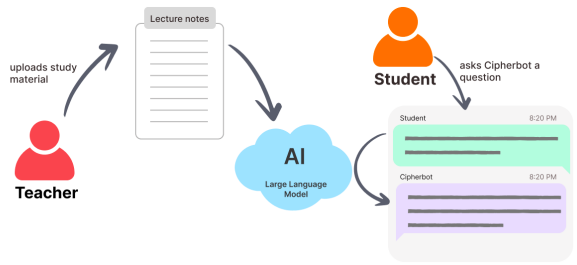


**Figure 1: The teacher uploads study materials to Cipherbot which processes the information. A student then interacts with Cipherbot by asking questions. Cipherbot analyzes the uploaded learning material and provides the student with answers, creating an interactive learning environment.**

## 2 RELATED WORK

Therep are many attempts at creating educational chatbots. A systematic review of students' interaction with educational chatbots can be found in [5] and for K-12 students [7], while a meta-analysis shows positive effects on learning outcomes [15]. Roughly speaking, we can divide the historical development of these chatbots into two main eras, which reflect the overall

development of chatbot technologies. The first era was based on *rule-based chatbots* (see a review of this era in [16] and [14]). These had limited applicability, as they could not handle complex queries or autonomously provide personalized responses to the students without predefining by the educator [11, 13].

The second and current era is the *LLM era*; essentially, the introduction of LLMs marked a divider in the abilities of educational chatbots. Virtually overnight, chatbots that struggled with producing grammatically correct, coherent, and logical sentences without predetermined responses could now engage in a natural dialogue with students [2]. Contrary to traditional rule-based chatbots, AI chatbots exhibit a much higher versatility in terms of interpreting human text inputs and, therefore, lend themselves to more meaningful interaction with students [2, 15]. The main advantage of this is the personalization of the learning experience based on specific queries that a student has [12].

## 3 METHODOLOGY

### 3.1 Cipherbot

Cipherbot (https://cipherbot.qcri.org) is an LLM-powered learning platform that combines various technologies to support students' learning experiences. It stores and indexes educational materials uploaded by teachers, allowing students to ask questions and receive contextually relevant responses with citations to the source materials. Cipherbot utilizes natural language processing, optical character recognition, and vector embeddings to retrieve information that addresses students' queries. The underlying LLM is OpenAI's GPT-3.5 Turbo. Other technologies include Redis Queue, Celery, Azure Storage, Azure Cognitive Search, and Azure OpenAI Chat Completions. Cipherbot's core design principle is that student queries are answered based on the learning materials curated by the educator, making Cipherbot different from general-purpose systems like ChatGPT. It is not only important that the engagement with an educational chatbot is *natural*, but the information must be accurate in order to not mislead the student [2, 15]. Grounding the dialogue to learning materials can possibly mitigate the problem of hallucination [8] whereupon the LLM "invents" factually incorrect answers. The particular approach Cipherbot applies is *retrieval-augmented generation* (RAG) which is an approach to locating specific information and then including it as a context for the LLM to generate its response [6]. So, Cipherbot matches students' queries with information extracted from specific learning materials and presented in natural language.

### 3.2 Study Design

The study involved 60 bachelor's degree students enrolled in a business course about managing online platforms. Most of the students were business majors (n = 37, 61.7%), with the remaining being communication students (n = 23, 38.3%). Most participants were Finnish (n = 51, 85.0%), while a small number were non-Finnish exchange students (n = 9, 15.0%). The average age of the participants was 22.8 years (SD = 2.9), and the gender distribution was 14 males (23.3%), 45 females (75.0%), and one participant who did not disclose their gender.

Participation in the study was voluntary and did not offer any extra credits or direct benefits. We informed students that the purpose of the study was to evaluate an AI tool, emphasizing the importance of such evaluations in enhancing educational quality. Informed consent was secured from all participants. We assured students that no personally identifiable information would be utilized in the analysis and that only the research team would have access to their data.

The seven-week course was structured around weekly themes, or "learning cycles" (LCs). As the course progressed, the learning cycles gradually transitioned to more complex and advanced topics. Students were instructed to spend approximately 30-60 minutes interacting with Cipherbot for each learning cycle, asking it to help them answer the learning questions provided in a spreadsheet. This interaction was in addition to their engagement with the course materials in Moodle. Students were also instructed not to use any other AI tool during the course to ensure accurate measurement of Cipherbot's impact on their learning.

## 4 RESULTS

### 4.1 Quantity of Interaction

The results only contain data from the 44 students (73.3% of course participants) who consented to participate in the study and actively used Cipherbot. Figure 2a shows that, overall, the usage was quite active, with more than 8000 messages being sent to Cipherbot, the median being 200 messages per student. The interaction does not follow a normal distribution but indicates that there are relatively fewer "power users" than students who use Cipherbot less than average.

The students' engagement with Cipherbot decreased over time (Figure 3), which might stem from the typical phenomenon of student engagement dropping over the course duration [3]. On the other hand, it might also be associated with the novelty effect of AI chatbots [9, 10], essentially postulating that once the exciting newness wears off, students tend to decrease their usage of the technology. An example of interaction can be seen below:

STUDENT: "How can ReCAPTCHA be explained as a platform?"

CIPHERBOT: "ReCAPTCHA can be explained as a platform because it provides a service that connects two groups of users: website owners who need to verify that their users are human, and internet users who are asked to solve CAPTCHA puzzles to prove that they are not robots. (…)"

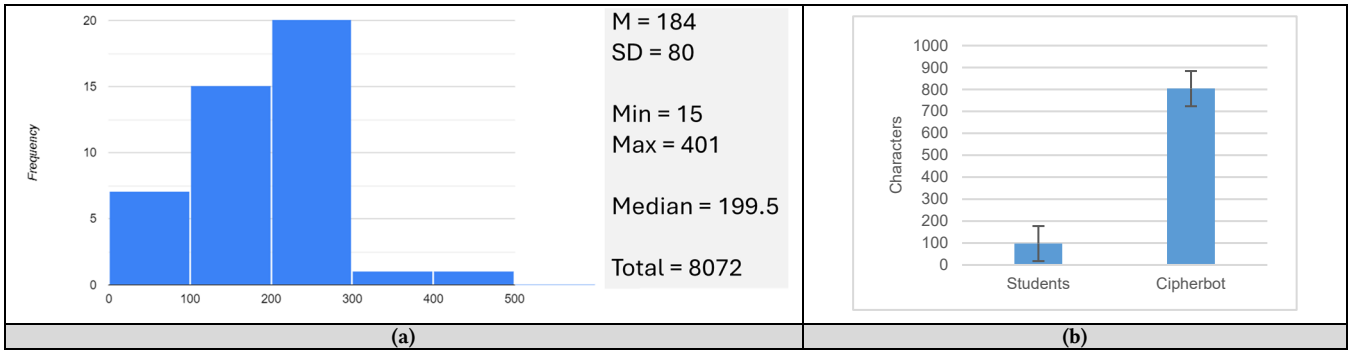


**Figure 2: (a) Distribution of messages sent by students to Cipherbot. (b) Length of messages written by students and Cipherbot.**

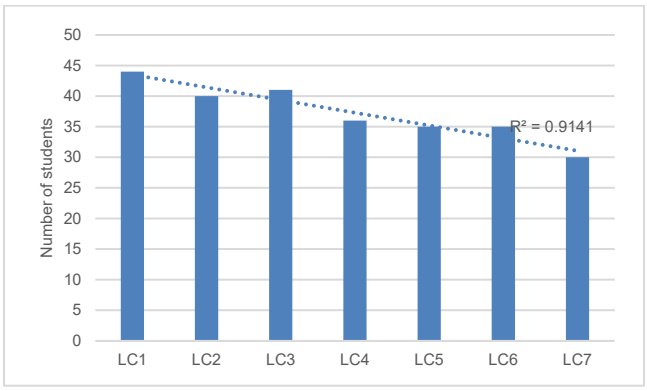


**Figure 3: Weekly active students in different learning cycles.**

A Mann-Whitney U test indicates that the messages written by students to Cipherbot are significantly shorter (M = 97.5 characters, SD = 80.1) than Cipherbot's messages back to students (M = 804.1, SD = 315.3), z = -106.8, p < .0001 (see Figure 2b). Concerning Cipherbot's ability to address student queries, the system logs, for every query, whether Cipherbot was able to generate a response to a student query ("TRUE" or "FALSE"). From the logs, we can compute that Cipherbot was able to answer student questions 82.5% of the time (n = 6666). This demonstrates an ability to accommodate a large number of questions, though we leave the further improvement of the success rate to future work. Briefly, the factors impeding a perfect score related to the system responsiveness, API/server downtime, and the nature of the question not being understood by Cipherbot.

## 4.2 Quality of Interaction

We measured the quality of interactions through a like/dislike ratio. Every time a student interacts with Cipherbot, they can indicate their satisfaction with the response by "thumbs up" (like) or "thumbs down" (dislike). Overall, 74 interactions were rated in this fashion, with 47 (63.5%) being likes and 27 (36.5%) dislikes, yielding a like-dislike ratio of 1.74 (i.e., for each one dislike, there are 1.74 likes). The dislikes were mainly associated with Cipherbot's ability to not address a student query (e.g., "I'm sorry, but the retrieved documents do not contain any information about market manipulation with cryptocurrency. Would you like me to help with anything else?"), leaving room for future analysis.

Moreover, we manually evaluated a random sample of 500 query-response pairs to assess how well Cipherbot addressed student queries from an educator's perspective. One of the authors, an educator in the course, coded these 500 pairs for five criteria: (1) **Factual correctness** ("Given the student's question, is this answer factually correct?"), (2) **On-topicality** ("Does the answer contain off-topic (=irrelevant, redundant information)?"), (3) **Specificity** ("Is the answer specific (=contains precise, detailed information)?"), (4) **Lack of repetition** ("Does the answer contain repetition (the same thing explained multiple times)?"), and (5) **Completeness** ("Does the answer fully address the student's question? ('No' means some parts of the question have not been addressed in the answer.)". The scale was binary (yes/no).

Another author, also an educator in the course, manually reviewed and corroborated the coding results. For each criterion, a success rate was calculated by dividing the "yes" instances by the total number of instances evaluated. The results indicate a satisfactory, but not perfect, performance by Cipherbot (see Figure 4).

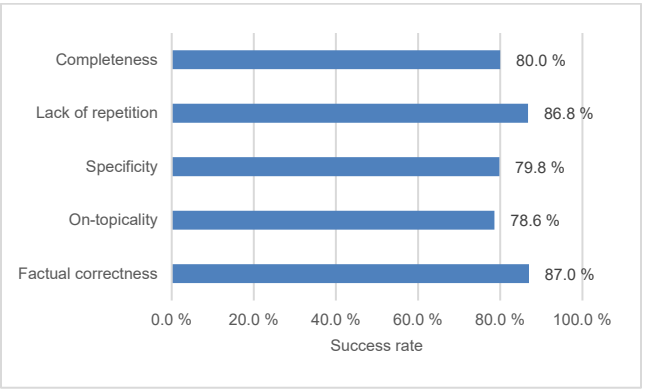


**Figure 4: Educator's evaluation of Cipherbot's response quality.**

Though most answers (87.0%) were factually correct, in some cases, the retrieved documents did not contain enough

information to answer a question. For example, in a question about whether Barbie doll is considered a platform based on the given definition, the retrieved documents do not provide enough information to determine if Barbie doll meets the specific criteria of a platform business model. Most answers were also complete (80.0%) and specific (79.8%). An example of incompleteness is when, in a question about whether venture capital funded platforms primarily seek to solve social problems, the answer did not directly state whether the answer was yes or no. Most answers (78.6%) did not contain off-topic information; the issues with staying on topic mostly related to providing general or tangentially related information when the retrieved documents did not contain directly relevant information. For example, in a question about internal and external mechanisms, the retrieved documents were empty (due to a momentary technical glitch), so the Cipherbot asked for clarification. Finally, there was mostly no repetition (13.2%) in Cipherbot's answers. For example, in a question about standardization in platform design, some phrases were repeated.

Findings indicate that Cipherbot handled thousands of questions with a performance score of above 70% in all evaluated criteria. The future challenge is going from the current 78-87% performance levels to above 90% performance. This will likely include adjustments to system prompts, the RAG implementation, and eradicating glitches that prevent the LLM from accessing the learning materials or other resources.

## 4.3 Variety of Learning Material Referencing

An important aspect of an educational chatbot is its ability to combine different source materials to provide balanced and complete answers [12, 15]. The learning materials, uploaded as PDF files, contained the course syllabus, textbook, and text transcripts for each video (each LC had seven videos). Cipherbot logs, for each response, what learning materials it draws the information from. From this, we can compute the number of times each learning material is referred to (see Figure 5).

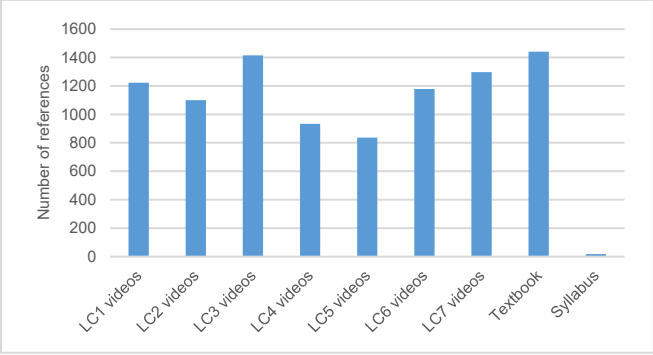


**Figure 5: The number of times learning materials were referred to by Cipherbot. As can be seen, Cipherbot referenced each learning material substantially, not only being focused on a single source. The textbook's highest reference rate makes sense because the textbook contained information that relates to each learning cycle.**

Overall, Cipherbot can draw information from one or multiple learning materials when composing its answer to the student query. On average, it used 1.17 learning materials (SD = 0.79), but more than one fourth (28.2%) of the responses included references to more than one learning material (see Table 1).

**Table 1: Number of learning materials referenced by Cipherbot in one response. Most of the time, Cipherbot refers to one learning material (54.3%).**

| # of learning materials | N | Percentage |
|---|---|---|
| Zero | 1411 | 17.5 % |
| **One** | **4386** | **54.3 %** |
| More than one | 2280 | 28.2 % |
| Total | 8077 | 100.0 % |

There were 118 unique combinations of Cipherbot referring to different learning materials in the same response (e.g., referring to LC1 *and* LC7 videos, etc.). To determine how many ways the learning materials could be referenced, we calculate the power set that has all possible combinations of the elements. Because we have 9 elements, the total number of combinations is 29 = 512. Given this "maximum variety" measure, we can calculate the actual observed variety as 118 / 512 = 23.0%, which corresponds roughly to one fourth. In other words, when answering student queries, Cipherbot combined different materials in one-fourth of the possible ways. The ability to combine learning materials is important because we want Cipherbot to reference multiple materials to avoid single-source bias [1]. On the other hand, it should not refer to unrelated materials in the same response. In a course where the material is organized in weekly themes (as it was here), the variety of materials that should be cited together is practically limited from the theoretical maximum. So, interpreted in context, we deem this result as an indication of Cipherbot's ability to combine different materials. However, it does show some tendency to rely on a single source.

# 5 DISCUSSION

Cipherbot shows promising results in terms of engagement and quality of answers. However, further research is needed on multiple fronts, including (1) technical development and backend functionalities, (2) frontend design and usability studies, and (3) a deeper examination of the exchanges taking place between Cipherbot and the students to understand what factors drive effects like knowledge retention [4]. These remain topics for future work. Since Ancient times, dialogue has been a central method of teaching. Leveraging LLMs, education is now facing an unparalleled situation where machines are, perhaps for the first time, teaching us humans. This is an exciting but somewhat stressful situation in which educators need to remain vigilant and examine the pros and cons of these systems. For example, Cipherbot displays the LLM's answer to a given query, but it also links the source documents from which the information was obtained. It is vital that educators stress the importance of double-checking information from other sources if a student feels uncertain or confused about the information.

## REFERENCES


[1] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. 2020. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 2020. PMLR, 528–539. Retrieved April 14, 2024 from http://proceedings.mlr.press/v119/bahng20a.html

[2] Sukhpal Singh Gill, Minxian Xu, Panos Patros, Huaming Wu, Rupinder Kaur, Kamalpreet Kaur, Stephanie Fuller, Manmeet Singh, Priyansh Arora, Ajith Kumar Parlikad, Vlado Stankovski, Ajith Abraham, Soumya K. Ghosh, Hanan Lutfiyya, Salil S. Kanhere, Rami Bahsoon, Omer Rana, Schahram Dustdar, Rizos Sakellariou, Steve Uhlig, and Rajkumar Buyya. 2024. Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems* 4, (January 2024), 19–23. https://doi.org/10.1016/j.iotcps.2023.06.002

[3] Mitchell M. Handelsman, William L. Briggs, Nora Sullivan, and Annette Towler. 2005. A Measure of College Student Course Engagement. *The Journal of Educational Research* 98, 3 (January 2005), 184–192. https://doi.org/10.3200/JOER.98.3.184-192

[4] Mohamed Ibrahim and Osama Al-Shara. 2007. Impact of Interactive Learning on Knowledge Retention. In *Human Interface and the Management of Information. Interacting in Information Environments* (*Lecture Notes in Computer Science*), 2007. Springer, Berlin, Heidelberg, 347–355. https://doi.org/10.1007/978-3-540-73354-6_38

[5] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Educ Inf Technol* 28, 1 (January 2023), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33, (2020), 9459–9474.

[7] Florence Martin, Min Zhuang, and Darlene Schaefer. 2024. Systematic review of research on artificial intelligence in K-12 education (2017–2022). *Computers and Education: Artificial Intelligence* 6, (June 2024), 100195. https://doi.org/10.1016/j.caeai.2023.100195

[8] Timothy R. McIntosh, Tong Liu, Teo Susnjak, Paul Watters, Alex Ng, and Malka N. Halgamuge. 2023. A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination. *IEEE Transactions on Artificial Intelligence* 1, 01 (November 2023), 1–13. https://doi.org/10.1109/TAI.2023.3332837

[9] Dijana Plantak Vukovac, Ana Horvat, and Antonela Čižmešija. 2021. Usability and User Experience of a Chat Application with Integrated Educational Chatbot Functionalities. In *Learning and Collaboration Technologies: Games and Virtual Environments for Learning* (*Lecture Notes in Computer Science*), 2021. Springer International Publishing, Cham, 216–229. https://doi.org/10.1007/978-3-030-77943-6_14

[10] Hua Ran, Nam Ju Kim, and Walter G. Secada. 2022. A meta-analysis on the effects of technology's functions and roles on students' mathematics achievement in K-12 classrooms. *Journal of Computer Assisted Learning* 38, 1 (2022), 258–284. https://doi.org/10.1111/jcal.12611

[11] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 02, 2019. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300587

[12] Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in Human Behavior* 117, (April 2021), 106627. https://doi.org/10.1016/j.chb.2020.106627

[13] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI '20*), April 23, 2020. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[14] Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence* 4, (2021). Retrieved March 26, 2024 from https://www.frontiersin.org/articles/10.3389/frai.2021.654924

[15] Rong Wu and Zhonggen Yu. 2024. Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology* 55, 1 (2024), 10–33. https://doi.org/10.1111/bjet.13334

[16] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* 2021, (April 2021), e8812542. https://doi.org/10.1155/2021/8812542