Communication Design for an Educational AI Chatbot: Analyzing Cipherbot's Communication Style and Challenges

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

Essi Häyhänen
University of Vaasa, Vaasa, Finland
essi.hayhanen@uwasa.fi

Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
jjansen@acm.org

ABSTRACT

We analyzed the communication patterns of Cipherbot, an educational AI chatbot that addresses validation and transparency problems in AI-student interaction, with 44 undergraduate business students using the system. Findings show that Cipherbot delivers information grounded in learning materials, primarily focusing on statements of fact, with a 99% occurrence rate. Also, about a third (31%) of Cipherbot's messages to students contain an example, which adheres to pedagogical best practices. However, more than a third (36%) of Cipherbot's communication with students *also* contained traces of opinion, particularly normative statements about how companies or individuals should behave. Cipherbot also demonstrates some level of social etiquette, using thank yous and apologies (7%), but it rarely engages in requests for information, or clarifications and comprehension checks–as these categories might be useful for student engagement, future exploration into diversifying Cipherbot's communication style to support learning is required. Student feedback suggests that usability issues in educational AI chatbots comprise both communicational and technical issues, e.g., incompleteness and redundancy–iterative testing of prompts with student feedback could address many of these challenges. Based on our findings, we make six propositions about students' interaction with educational AI chatbots.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Mindtrek '24, October 08–11, 2024, Tampere, Finland

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1823-6/24/10
https://doi.org/10.1145/3681716.3681727

CCS CONCEPTS

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

KEYWORDS

AI chatbots, education, communication style, challenges

ACM Reference Format:
Joni Salminen, Soon-Gyo Jung, Johanne Medina, Kholoud Aldous, Jinan Azem, Waleed Akhtar, Essi Häyhänen, and Bernard J. Jansen. 2024. Communication Design for an Educational AI Chatbot: Analyzing Cipherbot's Communication Style and Challenges. In *Proceedings of the 27th International Academic Mindtrek Conference (Mindtrek '24), October 08–11, 2024, Tampere, Finland*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3681716.3681727

1 INTRODUCTION

Cipherbot is an AI chatbot facilitating students' interaction with learning materials by integrating artificial intelligence (AI) and cloud-based technologies. Cipherbot allows educators to upload learning materials. Students can then inquire about any course-related topic and receive contextualized responses based on the provided learning materials (see Figure 1). Cipherbot exemplifies the rise of AI in education [11] and the incorporation of Generative AI (GenAI) into learning and teaching practices [11, 13, 23, 23]. GenAI and its subtechnologies, such as large language models (LLMs), are rapidly transforming how students learn, shifting the focus from *documents* to *dialogue*. This means that instead of reading a textbook, students now engage with AI chatbots to learn new concepts, engage in knowledge integration, and develop their understanding of a given topic [25]. Historically, education has predominantly relied on static documents such as textbooks, assignment sheets, presentations, and videos. The transition to a dialogue-based approach, spearheaded by educational AI chatbots

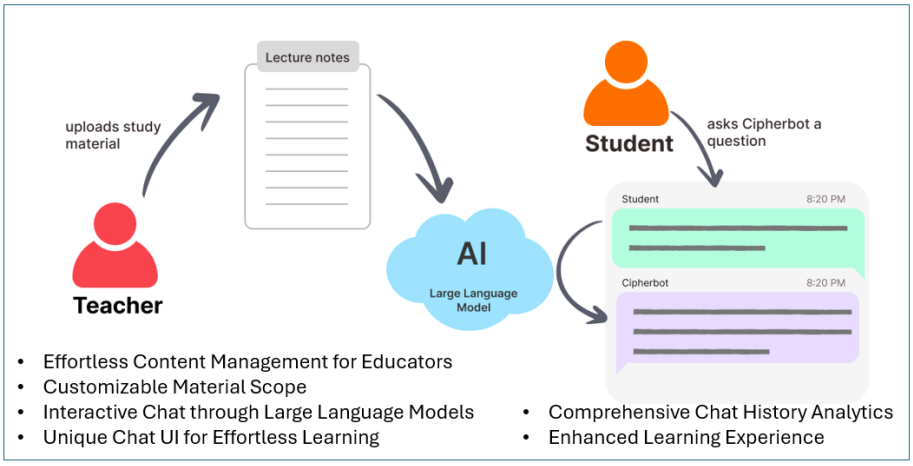


Figure 1: Cipherbot's usage workflow. Teacher uploads learning materials (e.g., lecture notes, articles, textbooks), students ask questions to learn about a topic, and Cipherbot formulates its answer based on information found from the learning materials.

enabled by GenAI and LLMs, remains understudied due to its novelty. As the educational sector continues adopting AI-based tools, techno-educational research must provide empirical evidence on patterns observed when students use these tools in real classroom settings.

In the big picture, the reader might ask, "Why are AI chatbots needed in education? What problems do they exactly solve?". While the answers to these questions can be multifaceted, the main reasoning is that AI chatbots offer personalized support to learners [28]. This personalized support is seen valuable because learners vary by their attributes such as prior knowledge, intellect, comprehension, and so, meaning that some students require more support to understand concepts than others [4, 20, 36]. Also, students also have varying interests, so their questions and specific learning objectives might vary, for which personalized support is required. Human educators are often unable to provide individual-level personalized support due to resource limitations, which has driven interest toward technology, specifically educational chatbots, as personalized support agents [26].

However, the educational community lacks guiding principles on effectively and responsibly leveraging these AI technologies. First, to provide effective learning, educational AI chatbots need to be *usable* [6]. By usability, we refer to the broad concept of the chatbot being pleasurable to use and effective for the purpose of use. Because educational AI chatbots are under development and can involve complex technical architectures with real-time connections to application programming interfaces (APIs), local and cloud server resources, and so on; technical challenges can affect students' perceived usability of these systems. Moreover, and perhaps more importantly, how the chatbot *communicates* with the student is essential: the communication style needs to be such that students *want* to engage with the chatbot. To this end, there can be multiple challenges, like the chatbot being too formulaic or robotic [5], unwelcoming [35], or simply lacking knowledge about a subject to be able to answer the student's query. While many of these usability factors are recognized in the context of the classical rule-based AI chatbots [30, 33], they are not well investigated in educational AI chatbots. At the same time, it is not conceptually clear what dimensions AI chatbots' usability consists of and if we can or should use traditional usability measures and concepts to understand usability in this context [15].

It should be noted that AI chatbots are indeed a different breed compared to traditional chatbots–they have achieved results in natural language processing (NLP) tasks that simply were not possible before (see review in [38]), including natural language generation. So, this increased ability to communicate fluently and accurately given the source material is likely to transcend to the educational context–or so it is believed. However, the body of knowledge on the capabilities of educational AI chatbots is still developing, and specifically, systematic research on their communication patterns is scarce. Understanding these communication patterns is important for multiple reasons; first, such information can help design better system prompts that make the communication more effective. Second, information on how students perceive chatbot communication, as well as how those perceptions differ by student groups, can help tailor multiple alternative communication capabilities within the same system. Third, knowledge of AI chatbot's communication patterns can help the research community to map and uncover the boundaries of this technology.

Thus, addressing this vital knowledge gap, we focus on two research questions (RQs):

RQ1: *What communication patterns does Cipherbot follow?*

RQ2: *What communication (and other) challenges do students report when using Cipherbot?*

By 'communication patterns', we refer to different utterance types that Cipherbot outputs in its dialogue with students. We employ the communication style taxonomy by Wahlster [34] that involves various categories for labeling utterances (see details in Section 3). By 'challenges', we refer to issues reported by students on a weekly basis during the seven-week period of data collection. Based on the nature of usability challenges in educational AI chatbots, we divide these into technical and communication challenges,

with various subclasses identified using open coding [32] Generally, if students find an educational system difficult or cumbersome to use, their use of it will be negatively affected [19].

To address the RQs, we present a user study of Cipherbot among 44 undergraduate students in a Finnish university course, investigating the value (and contemporary challenges) of GenAI in the educational sector. Therefore, the current work can best be understood as a 'user study' of 'user research' in which students are the users and Cipherbot is the technology, here representing a state-of-the-art system of educational AI technology. Based on our findings, we offer the following propositions concerning students' interaction with educational AI chatbots:

- **P01**: Educational AI chatbot usability consists of technical and communicational elements.
- **P02**: Educational AI chatbots can be highly factual and opinionated at the same time.
- **P03** Nature of opinions by AI chatbots err on the side of normative or moral statements.
- **P04**: Educational AI chatbots commonly provide examples to illustrate a concept.
- **P05**: Educational AI chatbots rarely engage in verifying students' learning.
- **P06**: Most common issues students report in AI chatbots relate to information quality and retrieval.

These are tentative proposals, considering that we tested only one AI chatbot. However, because Cipherbot is using OpenAI's GPT model, our findings are applicable to most educational AI chatbots built on this family of models.

In summary, our study adds to the growing body of research exploring the potential benefits and challenges of GenAI in educational settings. By presenting empirical data on student interactions with a fully functional AI-powered chatbot designed for educational purposes, we offer valuable insights into learners' experiences and perceptions. These findings are particularly relevant for educators at the undergraduate level who are interested in incorporating AI-driven learning tools into their teaching practices. Moreover, our research provides useful information for developers and designers creating and refining similar dialogue-based educational systems. Through this work, we aim to contribute to the ongoing stream of studies on the effective integration of GenAI technologies in education.

2 RELATED WORK

2.1 Background

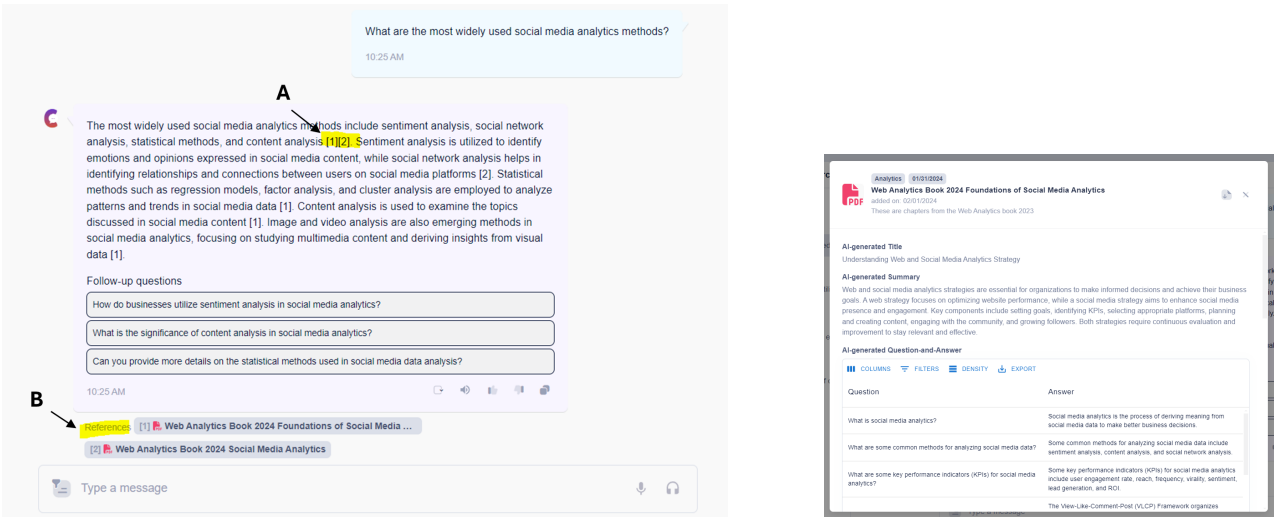
The emergence of AI technologies, such as chatbots and language models, has generated significant interest in the field of education. Education is potentially one of the most impactful applications of GenAI technologies, as these technologies can influence the learning experiences of millions of students at all levels worldwide, and this impact is already being felt. Consequently, numerous studies have explored using AI chatbots as interactive learning tools across a wide range of educational domains. Research has focused on at least the following subject areas: computer science, engineering, finance, general learning, language learning, mathematics, and practical skills training [11, 12, 21]. These studies have investigated the application of AI chatbots in various educational settings, including professional development, higher education, and K-12 education, encompassing both STEM and non-STEM subjects [29].

A systematic review by Kuhail et al. [21] showed that the majority of educational chatbots employ predefined conversational paths (cf. rules and triggers), while completely adaptive learning approaches were less common. In contrast to these "traditional" predefined systems, LLM-based educational AI chatbots like Cipherbot offer two key advantages: (1) they are context-agnostic, meaning they are compatible with virtually any learning context, and (2) they do not require pre-defined rules or conversational paths. Instead, the interaction between the educational AI chatbots and the student is entirely based on natural language, characterized by dialogue, conversation, and a general "ease" of having a discussion. Overall, educational AI chatbots like ChatGPT are rapidly being adopted by students [11, 12, 28]. However, these chatbots tend to have two main issues: (1) their responses to student queries are not generally rooted in course materials validated by educators (i.e., *the validation problem*), and (2) the responses generally do not indicate and provide references to source materials whose information the response is based on (i.e., *the transparency problem*). Cipherbot aims to address these two specific challenges: it uses retrieval augmented generation (RAG) to extract information from educator-validated materials when formulating its responses, and it provides references to the underlying materials the response is based on.

2.2 Educational AI Chatbots' Usability

It may seem paradoxical that a simple system like an AI chatbot can have usability issues, especially considering that usability is often associated with screen interface elements such as buttons, dialogues, menus, and other elements that are frequently absent from dialogue-based user interfaces [31]. However, usability issues in dialogue systems have been documented since the early 2000s [10]. The crux of the matter is that these issues rarely relate to visual aesthetics or navigation, as the interfaces to these conversational systems are minimalistic. Instead, the usability issues in dialogue systems often pertain to the quality of interaction and the users' sense of being understood [6]. Therefore, usability in these dialogue systems most frequently refers to the user's perception of the quality of the conversation.

Previous studies have emphasized a chatbot's usability in supporting the learning experience. They have identified various challenges, such as inadequate training data, lack of usability, and the need for holistic design frameworks [9, 21]. However, for modern LLMs, inadequate training data is typically not an issue, though most NLP development takes place in English [2]. The term "large" in the acronym LLM refers to the billions of parameters and gigabytes (or terabytes) of training data used to build these models [3]. Despite the advancements in LLMs, it remains unclear what constitutes good usability for an educational AI chatbot and what types of challenges might arise in this context. As a result, usability represents a 'last-mile problem' of education technology. If students and educators do not find the systems usable, there will be no adoption and active use. Previous findings, particularly those related to challenges in maintaining active usage and engagement



(a) (b)

Figure 2: Transparency in Cipherbot's answers to student queries. (a) shows an example response from Cipherbot, where A indicates references to learning materials from which the information is extracted and B shows the names of the materials that are clickable links, leading to a summary of the specific learning material (b).

[37], indicate that the usability of educational AI chatbots is far from being solved.

2.3 Pathways Toward Workable AI Chatbots

There is a clear and pressing need for cross-disciplinary collaboration between designers and educators, which is embodied in the concept of 'pedagogical design knowledge'. Drobnjak et al. [9] conducted a systematic review of 252 articles on using conversational agents (CAs) in education. Their findings revealed that there is "insufficient holistic design knowledge for pedagogical CAs" (p. 1003), which might apply to both system designers and educators. In other words, system designers might need pedagogical insights, while pedagogical practices might need to be adjusted to accommodate the use of CAs. However, in practice, many AI systems might lack in-depth understanding of pedagogical theories and instead be technology-driven [11, 29]. It is often mentioned that LLMs simply predict one word after another, without an innate understanding of fields like pedagogy. At the same time, according to Holmes and Tuomi [16], most educational AI applications are no longer speculative or at research stage, but commercially available and thus accessible to educators worldwide. Therefore, user studies or user research that informs the research community on real experiences and challenges faced by students when using these systems are called for. Our study, focusing on the communicational aspects in real conversations between undergraduate students and Cipherbot addresses the relationship between communication and usability in the educational AI chatbot context.

3 METHODOLOGY

3.1 System Overview

Cipherbot (https://cipherbot.qcri.org) is an innovative LLM-powered learning platform that integrates various technologies to enhance students' learning experiences. The platform stores and indexes educational materials provided by teachers, enabling students to ask questions and receive responses that are contextually relevant and include citations to the source materials. Cipherbot leverages natural language processing, optical character recognition, and vector embeddings to identify and retrieve the most pertinent information based on students' queries. As of this writing, the platform utilizes OpenAI's GPT-4 as its underlying LLM, with the latest version deployed in Azure. Cipherbot also incorporates a range of open-source and commercial technologies, including Redis Queue, Celery, Azure Storage, Azure Cognitive Search, and Azure OpenAI Chat Completions. While the technical details of Cipherbot's architecture are beyond the scope of this work, our primary focus is on reporting the results of student interactions.

Cipherbot's core design principle is answering student queries based on the learning materials curated by educators, setting it apart from general-purpose systems like ChatGPT (see Figure 2). In an educational context, it is crucial that the engagement with a chatbot is not only natural but also factually accurate. Providing students with accurate information is essential to avoid misleading them [11, 37]. By grounding the dialogue in learning materials, Cipherbot aims to mitigate the issue of hallucination [24], where LLMs generate factually incorrect answers. Cipherbot employs retrieval-augmented generation (RAG), an approach that locates specific information and incorporates it as context for the LLM to generate its response [22]. This approach allows Cipherbot to match students' learning queries with information extracted from specific materials, ensuring the responses are based on valid information.

Cipherbot's primary objective is to provide students with informative and contextually relevant answers that include specific references to the background materials used to generate the responses. This design principle adheres to the critical concept of transparency in AI systems [1]. By clearly citing the sources used to formulate its responses, Cipherbot ensures that students can verify the information and gain a deeper understanding of the topic at hand. Furthermore, Cipherbot's approach to matching students' queries with relevant materials and generating contextually appropriate responses with citations aligns with the technology-mediated learning framework [7]. This framework emphasizes acquiring knowledge using digital tools and platforms that offer interactive and personalized learning experiences, which, ideally, enables students to engage with educational content in a way that supports their learning.

In terms of usage, students are expected to reflect on the information provided by Cipherbot and ask follow-up questions to deepen their knowledge. This process is expected to be free of any aspect of 'asking stupid questions' or losing one's face, as the communication remains private between the student and Cipherbot (to clarify: the exchanges are analyzed for research but without considering the identity of a given student). Cipherbot is expected to provide factual information to the students while providing references to specific learning materials. Cipherbot is expected to remain neutral and simply convey information.

3.2 Participants

The participants in this study were undergraduate students enrolled in a business course focused on managing online platforms. The course had 81 enrolled students. Of these, 65 (80.2%) agreed to participate in the study. Of these, 44 (67.7%) used Cipherbot to send at least one message (Min = 15, Max = 401, M = 183.4, Median = 199.5, SD = 79.8 messages sent to Cipherbot by students). (More information on student engagement is reported in related work [27].) The majority of the students (n = 27, 61.4%) were pursuing business majors, primarily marketing, while the remaining students (n = 17, 38.6%) were studying communication. The study sample consisted primarily of Finnish students (n = 36, 21.8%), with less than a fifth (n = 8, 18.2%) non-Finnish exchange students. The average age of the participants was 22.8 years (SD = 2.9). Regarding gender distribution, the sample included 37 females (84.1%) and 7 males (15.9%).

Participation in the study was voluntary, and no extra credits or direct benefits were offered to the students. The researchers informed the participants that the purpose of the study was to evaluate an AI tool, emphasizing the significance of such evaluations in improving educational quality. All participants provided informed consent before taking part in the study. The researchers assured the students that no personally identifiable information would be used in the analysis and that access to their data would be restricted to the research team.

The course was conducted over a seven-week period and was organized into weekly themes, referred to as learning cycles (LCs). As the course progressed, the LCs gradually increased in complexity and advanced to more sophisticated topics. For each LC, students were instructed to spend approximately 30-60 minutes interacting with Cipherbot, seeking its assistance in answering the learning questions provided in a spreadsheet. This interaction was supplementary to the students' engagement with the course materials (i.e., course textbook, videos) available on the Moodle platform. The idea was, according to the logic of the from-documents-to-dialogue concept, that students can engage with Cipherbot to learn important knowledge components. This was not implemented in "instead of textbook reading" basis but "in addition to textbook" basis, so the students used Cipherbot to strengthen their learning – they were expected to nonetheless read the textbook to prepare for the exam. Moreover, the idea with the learning questions was that the educator had defined learning goals for each week and the learning questions were consistent with these goals.

To ensure an accurate assessment of Cipherbot's impact on their learning, students were instructed to refrain from using any other AI tool (e.g., ChatGPT) throughout the course. This was controlled by removing participants who indicated using an AI other than Cipherbot, based on a post-course survey. Because no student among the ones using Cipherbot indicated using another AI tool, none were excluded from this study.

3.3 Data Collection and Annotation

Messages, including those sent by students to Cipherbot and Cipherbot's responses, were automatically logged by Cipherbot's built-in analytics module. During the seven weeks, the students sent 8077 messages to Cipherbot, and Cipherbot responded to each message, yielding another 8077 messages, the total number of messages therefore being 16,154. For RQ1, we focus on a randomly selected subset (n=300) of these messages–a subset because manually coding the whole dataset would be extremely laborious and probably would not change the findings much, as a pattern can often be established with a reasonably large sample size.

To address RQ1, we mobilized categories from *VERBMOBIL*, a framework focused on translating spoken language [34]. The framework involves a set of labels for annotating dialogue acts in transcribed conversations. The purpose of these labels is to understand the structure and intention behind segments of dialogue. We employed eight categories from the VERBMOBIL framework (see Table 1). Two coders annotated a sample of the prompt-response pairs (prompts were messages from the students to Cipherbot, and responses were Cipherbot's replies). Coder 1 annotated 300 pairs using the definitions in Table 1 as guidelines. Coder 2 independently annotated 30 pairs for interrater agreement calculation. The results indicate high percentage agreement (82.54%) and moderate Fleiss' Kappa agreement (k = 0.41).

For RQ2, we focus on students' feedback on Cipherbot, which was asked in weekly surveys asking, among other things, about system usability and any potential challenges they observed. We combine the students' open-ended feedback into a spreadsheet for analysis, collecting 208 text responses (on average, 4.7 per student, and yielding a 67.5% response rate given 44 students and seven weekly surveys). Addressing RQ2, we used open coding. The goal of open coding is to identify, name, categorize, and describe phenomena found in the text [32], which is what we did. We asked the students to write up any issues they observed when using Cipherbot. This was done each week. The obtained responses were

Table 1: The VERBMOBIL categories employed in our study.

Category	Definition
Greetings	These are dialogue acts used at the beginning or end of conversations. They include salutations like "hello" and parting phrases like "goodbye." Greetings are important for establishing and closing communication in a polite manner.
Requests for Information	This category encompasses any dialogue act where the speaker is seeking information from the other party. It includes asking for clarifications, specific details, or any other inquiries that require a response with information.
Offers and Suggestions	Offers are proposals made by the speaker to provide something, while suggestions are recommendations for a course of action. These dialogue acts are key in negotiations and planning within conversations.
Thank Yous and Apologies	This category includes expressions of gratitude ("thank you") and apologies. These dialogue acts are essential for maintaining politeness and addressing social norms within conversations.
Confirmations and Disconfirmations	Confirmations are used to affirm or verify information or proposals, indicating agreement or correctness. Disconfirmations, on the other hand, deny, correct, or refute previously mentioned information or assumptions.
Statements of Fact	These dialogue acts convey information that the speaker believes to be true and factual. They are used to share knowledge or assert truths within a conversation.
Opinions	Unlike statements of fact, opinions express personal beliefs, preferences, or evaluations. They reflect the speaker's subjective viewpoint rather than objective facts.
Clarifications and Comprehension Checks	This category includes dialogue acts where the speaker requests further explanation or verification to ensure understanding. It also encompasses efforts to clarify one's own statements to prevent or resolve misunderstandings.

compiled into a spreadsheet and the lead author carefully read them while assigning descriptive labels (i.e., codes) based on interpreting the meaning of each feedback. Two levels of codes were used: main level which was a broader category of issues (e.g., communication) and sublevel, which described in greater detail the type of issue observed (e.g., incompleteness). When the number of labels in a given category became large, the lead author then merged them to form themes based on the similarity of the labels. The labels and themes, along with their frequencies, can be seen in the following section.

4 FINDINGS

4.1 RQ1: What communication patterns does Cipherbot follow?

Cipherbot's communication style tends to be factual (see Figure 3), with **statements of fact** appearing in almost all of its messages (n=296, 98.67%). These mostly relate to defining concepts (see underlined parts in the example): "**A business model refers to the overall approach a company takes to generate revenue and create value, including elements such as the target customer, value proposition, key resources, and distribution channels [2]. On the other hand, a revenue model specifically describes the different ways a business will generate income [1]. It is just one of many building blocks within a business model, and it is tightly connected to all the other elements [1]. (. . .)**". The cases where Cipherbot did *not* present a factual statement (n=4, 1.33%) related to empty responses (n=2, 50%) and other inabilities to answer (n=2, 50%).

Interestingly, though factual statements are the most dominant category, the second most prevalent category is **opinions** (n=107, 35.70%). This is almost exclusively because of normative statements[1] made in Cipherbot's responses: "**Additionally, power imbalances and blurred lines between intermediary and employer can lead to exploitation and a lack of worker control. Therefore, while platforms can increase efficiency, it is important to consider and address their potential negative effects on society.**"

These results indicate that, somewhat paradoxically, Cipherbot merges facts and opinions. This can be explained by Cipherbot's common argumentation style, which is to first present factual statements and then conclude with a normative (opinionated) suggestion. Consider the example (factual statements bolded, opinions underlined): "**Reputation systems matter for transaction-based platforms because they can help to ensure trust between the parties involved in a transaction [1]. Reputation systems are a form of feedback loop that allows users to rate each other based on their experiences. This feedback can be used to build trust between buyers and sellers and other platform participants. Reputation systems can also help to reduce transaction costs and create new markets [2]. However, if not architected well, reputation systems can lead to increasing worker inequality [1].** Therefore, it is important for platforms to be careful when designing reputation systems, to ensure that

[1]In the coding, we considered all normative statements ("should", "need to", "it is important") as opinions rather than facts.

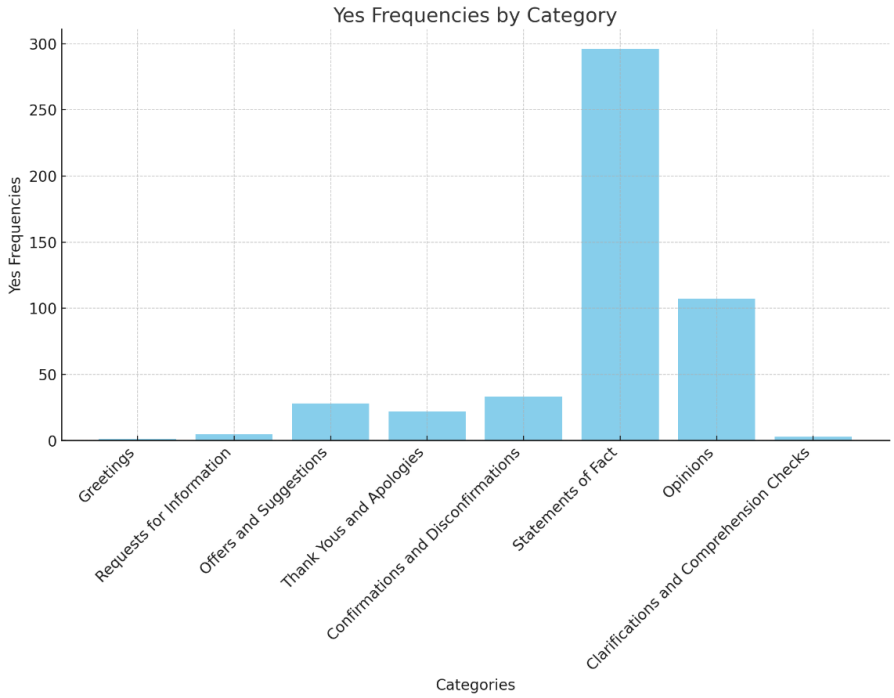


Figure 3: Cipherbot's communication markers. "Yes frequencies" in the y-axis indicate the number of occurrences.

they are fair and equitable for all users [1]." In this example, 'fair and equitable' sounds like a noble goal. However, in a business context, the organization's goal might not be the creation of fair and equitable systems but, in fact, to maximize its profit (this is the classic shareholder argument as to why firms exist [17]). We cannot say fairness is wrong, but we also cannot say it is an objective. So, in the course context, this argument is an opinion.

To probe further into this matter, we conducted a co-occurrence analysis by computing how many times each pair of categories occurred simultaneously (i.e., in the same message). The occurrence analysis shows that two combinations took place more than 10% of the time: 'Statements of fact' and 'Opinions': 107 (35.7%), and 'Confirmations and disconfirmations' and 'Statements of fact': 33 (11.0%). So, even though Cipherbot had a high opinion rate (more than one third of its responses contained an opinion, 36%), there is an alleviating factor which is that every time Cipherbot presented an opinion, it also presented a fact (this can be seen from the fact that the opinions' co-occurrence rate with facts is the same as the occurrence rate itself). More specifically, its opinion was a normative statement based on interpreting facts, as illustrated above. Conversely, Cipherbot does not present an opinion most of the time; it represents facts–the fact presentation is a highly dominant characteristic (99%), and approximately two times out of three, a statement of fact is made *without* following up with an opinion.

Moreover, the relatively strong co-occurrence between statements of fact and confirmations and disconfirmations indicates a pattern of communication in which Cipherbot either corroborates ("**<u>Yes</u>, platforms tend to have a higher ratio of revenue per employee compared to non-platform firms [1][2]. This is because platforms are designed to facilitate interactions between multiple communities, rather than focusing on optimizing a linear value chain [1]. (. . .)**") or refutes ("**<u>No</u>, a platform does not necessarily need to own the resources it makes use of. As mentioned by Tom Goodwin in [1], 'Uber, the world's largest taxi company, owns no vehicles. Facebook, the world's most popular media owner, creates no content. Alibaba, the most valuable retailer, has no inventory. And Airbnb, the world's largest accommodation provider, owns no real estate.' Instead, these companies rely on external resources provided by the community. (. . .)**") information in a student's question and backing its argument with a factual claim rooted in the study material.

Concerning the other categories, **confirmations and disconfirmations** (n=33, 11.0%) dealt with corroborating ("Yes, network effects can be measured. According to one of the retrieved documents, network effects are critical generators of value in platform businesses [2]. (. . .)") or refuting ("No, a platform does not necessarily need to own the resources it makes use of. (. . .)") information in a student's question. We observed that in these cases, Cipherbot tends to start with a pattern of "Yes, . . ." or "No, . . .", depending on whether it is confirming or disconfirming the information. We thus searched for these patterns among the whole dataset of Cipherbot's answers and found that Cipherbot is slightly more likely to confirm (n = 392, 4.9%) than disconfirm (n = 273, 3.4%) students' questions. We cannot conclude whether this behavior is biased or not without knowing if the questions themselves contained factually correct or incorrect information, clarifying which we leave for future work.

Offers and suggestions dealt (n=28, 9.33%) with suggestions of improving the prompt or retrying the information request: "The retrieved documents do not provide information specifically related to how system development kits (SDKs) are related to platform growth or any examples. As a teaching assistant, I would suggest you refine your query or provide a different question that is more closely related to the information available in the documents."

Similarly, **thanks yous and apologies** related to errors (n = 22, 7.33%) ("I'm sorry, but I could not find any relevant documents to answer your question. Can I help you with anything else?"); interestingly, all of the occurrences were about apologizing, none was about Cipherbot thanking the user. So, Cipherbot is more apologetic than thankful.

Request for information (n = 5, 1.67%) related to Cipherbot needing additional inputs from the student: "The retrieved documents do not provide information specifically related to how system development kits (SDKs) are related to platform growth or any examples. As a teaching assistant, I would suggest you refine your query or provide a different question that is more closely related to the information available in the documents." A relatively low rate in this category can be seen as a positive sign because it implies Cipherbot can address most queries without needing to ask for more information. This is a desirable trait for the efficiency of communicating with students. To this end, Cipherbot seems to be more of an information provider than a seeker.

Clarifications and comprehension checks (n= 3, 1%) related to inability to find information, whereupon Cipherbot asked for clarification: "I'm sorry, but the retrieved documents do not provide enough information to answer your question. Could you please provide more context or clarification about what you mean by 'value unit'? Alternatively, you could ask another question that I may be able to help you with based on the retrieved documents."

The low **greeting** rate (n =1, 0.33%) might suggest that Cipherbot prioritizes functionality over social niceties. This is certainly the case, and the overall style is factual with some degree of politeness; yet, this category's frequency is somewhat misleading in the sense that greetings happen in the first message ("Hello! I'm an AI-powered teaching assistant here to help you with any questions you have. How can I assist you today?") and here we assessed a random sample of messages of which many take place in the middle of the conversation. In any case, Cipherbot is focused on efficient and effective (functional) communication.

Finally, we calculated the frequency of example terms ('example, 'for instance', 'e.g.', 'like', 'likes of') and found that example terms appear in 30.99% (n = 2503) of Cipherbot's answers. In other words, almost a third of the answers contain an example illustrating a business principle conceptually ("**(. . .) Additionally, platforms need to consider the quality of their users, as not all users are equally valuable. For example, some users may be free riders or may engage in negative behaviors that harm the platform's ecosystem (. . .).**") or through real people and organizations (**"Yes, it is possible to start marketing a platform without first building it. (. . .) For example, when Aaron Patzert wanted to build Mint.com, he hired a marketer before they even had a product. (. . .) In addition, Elon Musk sold $1 billion worth of Model 3 cars before anyone even touched a car through a prototype and building hype [1]"**). This behavior is promising, as giving examples is considered an effective pedagogical technique [39].

4.2 RQ2: What communication (and other) challenges do students report when using Cipherbot?

Overall, 208 responses were obtained on the open-ended form field, "Please write here any issues you observed with Cipherbot." These were collected over the course of the seven weeks. The responses were manually categorized by the lead author based on the presence and type of issues encountered (see Table 2 for summary statistics). The majority of interactions (58.7%, n=122) were found to have no issues (e.g., "I had no issues.", P17, Week 1). Among the interactions with issues, 26.4% (n=55) were related to communication problems, 13.9% (n=29) were due to technical difficulties, and 1.0% (n=2) involved feature-related concerns.

A closer examination of the 29 interactions that encountered technical issues (see Table 2) revealed that the most common problem was non-responsiveness (41.4%, n=12), followed by slowness (34.5%, n=10). Crashing accounted for 17.2% (n=5) of the technical issues, while user interface problems were relatively less frequent (6.9%, n=2). These results suggest that efforts to improve the responsiveness and speed of the conversational AI system, as well as reducing instances of crashing, could substantially mitigate technical issues and enhance user experience. Concerning slowness issues, this would need further investigation to determine if it was system or network related.

Concerning communication issues, 15 codes were created by the lead author in an inductive coding process. These were divided into four broader themes (see Table 3). Information Quality and Retrieval (n=30, 54.5%) issues relate to the ability to find relevant, accurate, complete, and balanced information. Communication and Comprehension (n=14, 25.5%) issues related to understanding, fluency, translation, and repetition in communication. Trust and Consistency (n=6, 10.9%) issues related to the trustworthiness and consistency of the information provided. User Interaction (n=5, 9.1%) issues related to prompting and personality in user interactions.

A closer examination of the 29 interactions that encountered technical issues revealed that the most common problem was non-responsiveness (41.4%, n=12), followed by slowness (34.5%, n=10). Crashing accounted for 17.2% (n=5) of the technical issues, while user interface problems were relatively less frequent (6.9%, n=2).

We also present illustrative quotes from the participants. Concerning technical issues, Non-responsiveness dealt with the system not being accessible ("The site showed an error during the conversation/asking process and I was not able to continue using it despite refreshing the page and logging out and in again. The bot worked normally the next day though." (P28, LC2)). The responsiveness was not stable but at times required students to retry ("It didn't work at first. The same thing has happened three times when I've been using Cipherbot with my mobile phone." (P27, LC3)), which some of them did not do and thus the downtime affected their use of the system negatively ("Every time I had problems with opening the chat and the whole website would not respond. Sometimes it stopped responding even after it had started working so sometimes

Table 2: Coding of the open-ended answers concerning issues observed by students in Cipherbot.

Type	N	%
No issues	122	58.7%
Communication	55	26.4%
Technical	29	13.9%
• Non-responsiveness (n=12, 41.4%) • Slowness (n=10, 34.5%) • Crashing (n=5, 17.2%) • User interface (n=2, 6.9%)		
Features	2	1.0%
Total	208	100%

Table 3: Communication issues breakdown (N=55).

Information Quality and Retrieval	Communication and Comprehension	Trust and Consistency	User Interaction
- Inability to find answer: 22 (40.0%) - Misinformation: 2 (3.6%) - Informativeness: 1 (1.8%) - Precision: 1 (1.8%) - Irrelevancy: 1 (1.8%) - Incompleteness: 1 (1.8%) - One-sidedness: 2 (3.6%)	- Comprehension: 9 (16.4%) - Translation: 1 (1.8%) - Repetition: 4 (7.3%)	- Trust: 3 (5.5%) - Inconsistency: 3 (5.5%)	- Interaction fluency: 2 (3.6%) - Prompting: 2 (3.6%) - Personality: 1 (1.8%)
N=30 (54.5%)	N=4 (25.5%)	N=6 (10.9%)	N=5 (9.1%)

I ended up doing half of the questions without Cipherbot instead of fighting it since I was able to remember a lot of the info anyway." (P32, LC7)).

Slowness appeared in having to wait till Cipherbot provided the answer ("Sometimes Cipherbot was a little bit slow." (P12, LC1), especially over longer usage sessions ("When you used Cipherbot for a long time it started to slow down." (P15, LC1)). As a workaround, some students refreshed the page which seemed to address this issue ("This week Cipherbot was slow and it needed to restart many times before it answered questions." (P02, LC5)).

Crashing can be considered among the most severe system usability issues because it prevents the usage altogether. Several reports of crashing took place in different weeks ("I have sometimes issues with Cipherbot website. Sometimes it crashed without reason and it is not findable" (P02, LC1); "the site crashed on more than one occasion" (P05, LC2); "lot of time its crashed and doesn't give any answers" (P23, LC7)).

User interface (UI) issues dealt with confusion on accessing the background materials; access was provided, but this was not always clear to the students ("When the AI mention the quotes in the videos is difficult to understand at what videos it refers" (P28, LC7). Also, there was an issue with using search ("I noticed repeatedly some bugs while clicking in the search bar" (P01, LC5)).

Concerning communication issues, Comprehension issues related to factors like the lack of contextual awareness ("Cipherbot didn't understand to connect two questions to one conversation." (P04, LC1); "I've noticed there's sometimes difficulties to understand the context if I ask follow-up questions." (P20, LC3)) and inability to deal with grammatical errors in the input ("On some occasions it was difficult for the bot to understand the question if there was a minor grammar mistake." (P36, LC1)). In some cases, Cipherbot could not find an answer to the student's question ("There was 1-2 questions that the Cipherbot wasn't able to answer." (P25, LC1); "The chatbot was not able to provide answers to all questions." (P22, LC2)).

The repetition manifested as providing highly similar outputs than previously ("As I asked more specific questions about my prior ones, Cipherbot almost answered with the same paragraphs and gave no further information about the specific question asked." (P18, LC1); "On some chats it repeated same things again even though I reframed the question/asked for clarification." (P27, LC1); "Chatbot couldn't find the materials to some questions and repeated same answers even if I tried to ask clarifications." (P10, LC5)).

Concerning Misinformation, a student noticed erratic behavior when Cipherbot changed its opinion upon being questioned ("There was one occasion when I noticed it was lying to me. It was giving me examples of infrastructure platforms and when I asked why they were infrastructure platforms, it said they were not." (P33, LC1)). Some students also had trust issues with Cipherbot ("The information for some questions that I gave to Cipherbot still had to be well reworked behind to know if it was right." (P34, LC1);

"Wikipedia as a source was interesting in one question as Wikipedia is not a reliable source." (P17, LC3)).

Interaction Fluency appeared as an issue of general uneasiness perceived when interacting with Cipherbot ("it is a bit difficult to have a smooth interaction with Cipherbot" (P11, LC1); "sometimes it is hard to have a conversation with the bot." (P01, LC3)). Cipherbot's personality was seen problematic by one student ("Chatbot's personality could be even friendlier. At times it seemed too robotic." (P35, LC7)). Students also recognized their own role in successful interaction through Prompting ("No major issues, some misinterpretations caused by false langue." (P25, LC4)).

The students also expressed issues with Informativeness ("The bot was not as informative as other AI chatbots." (P35, LC1)), Translation ("It couldn't translate any words from English to Finnish." (P19, LC1)), Precision ("Doesn't always give as precise answers as you would hope." (P24, LC2)), Irrelevancy ("there was one question where it seemed to give non-relevant info." (P17, LC4)), Incompleteness ("It wasn't able to retrieve all the information." (P14, LC5)), Inconsistency ("Sometimes Cipherbot says that it doesn't have the answer but when I ask the same question later again, it has an answer." (P09, LC5)), and One-sidedness ("the answers are good as long as you understand it but when you don't understand something, it can't explain it any better." (P16, LC7)).

5 DISCUSSION

5.1 Discussion of Findings

Based on our findings, Cipherbot has high *fact orientation*: it predominantly communicates by making factual statements. However, it also often includes a normative statement about a topic that can be interpreted as an opinion rather than a fact. So, it also has a relatively high *normative orientation*. Most commonly, Cipherbot's answer first argues in a factual manner and concludes with a normative statement. This normativity has also been observed by other scholars: for example, Cooper and colleagues [8] voice a concern over LLMs imposing claims with high confidence: "as it currently stands, ChatGPT runs the risk of positioning itself as the ultimate epistemic authority, where a single truth is assumed without a proper grounding in evidence or presented with sufficient qualifications." This aspect is problematic because, as we mentioned in Section 3.1, Cipherbot is expected to remain a neutral information conveyer.

In terms of *social awareness*, Cipherbot commonly apologizes when unable to address a student's query. In these cases, it tends to offer suggestions to revise the prompt or otherwise improve the chances of getting a response. When asked a question that contains a loaded claim or asks for confirmation, Cipherbot is able to answer naturally by providing confirmation or disconfirmation. So, these abilities exhibit *communicative versatility*.

In contrast, Cipherbot does not confirm students' learning or pose questions back to the student unless it cannot address the student's query. This implies that there is a need for *pedagogy-informed prompt development* to optimize Cipherbot's communication style toward giving feedback and encouraging the student to continue learning.

In terms of communication, most issues are related to *information quality and retrieval*, with inability to find a specific answer being the primary concern. On the one hand, it is a positive sign that an educational AI chatbot like Cipherbot can admit to not finding an answer rather than inventing one (i.e., the so-called risk of hallucination [18]). On the other hand, this emphasizes the need for educators to upload a comprehensive set of materials when employing AI chabots like Cipherbot that draw their answers strictly from the learning material. This is the trade-off relative to general chat tools like ChatGPT: ChatGPT could likely answer all student questions, but it would do so using material not verified by the educator, which risks problems in accuracy and remaining on course topical scope. So, educators need to provide adequate source materials to cover many possible questions that students might have for Cipherbot.

Moreover, actual communication issues relating to comprehension (e.g., not understanding students' prompts or not providing answers that make any sense) appear to be rare instances. Moreover, apart from minority instances, trust in Cipherbot's answers or its personality style did not raise major concerns (both in the range of 10% or below of coded instances). So, Cipherbot is generally effective at interpreting student queries.

5.2 Practical Implications for Educators and Developers

For educators, the main implication is that Cipherbot appears usable already in its current state. It communicates with students in a logical and compelling manner, rarely making mistakes. So, from our perspective, it appears "safe" to deploy in real classes depending on the content. Naturally, this requires additional thinking on the educator's part, including reflection on the following questions for Cipherbot implementation in real courses:

- **How is Cipherbot integrated?** E.g., as compulsory or optional learning support. Either approach is possible, but if educators are specifically looking for feedback on the tool, making Cipherbot a compulsory part of course taskings can provide more feedback than integrating it as "extra work" for students.
- **Do students use Cipherbot alone or in teams?** We recommend allowing individual use due to the personalized nature of Cipherbot, though it can also support teamwork where students divide the work into individual parts.
- **Are there specific activities or tasks that students should and should not use Cipherbot for?** For example, we recommend using it for tasks like concept definition and deeper topic understanding. We do not recommend using it for writing the answers on students' behalf. This helps mitigate the general concern that AI-based tools decrease students' critical thinking [14]. We encourage educators using Cipherbot to adequately inform the students that Cipherbot is meant to help, tutor, and assist in learning, not to do the student's thinking and writing. The educators ought to employ appropriate mechanisms to prevent academic dishonesty Cipherbot (and other tools like it), including physical exams where no access to AI is given.
- **Does the educator provide guiding questions for the students?** We recommend providing some guiding questions for the students to get started. This is because students

might struggle to get started, as they do not know what they do not about the study topic. "Seed questions" provided by the educator can help get started with Cipherbot and Cipherbot's provided follow-up question options can then further help student engagement with the learning material. In our experience, it is not safe to presume that "just because it's chat, it's easy to use". Rather, though students of course know *how* to use a system like Cipherbot (i.e., by typing), they do not necessarily know *what* to type.

- **Is specific training required?** Though Cipherbot has some generic help materials, including an explainer video, we recommend that each educator employing Cipherbot would record or write their own instructions that are specific to the course in question, providing specific examples in which Cipherbot can be used, as well as outlining the rules for its proper use. As these are course-specific, educators must still see the effort of properly assisting students in Cipherbot's usage (the same applies to any educational AI tool).

For developers, Cipherbot's communication and system design provide additional implications. Generally, Cipherbot communicates well. On the other hand, its communication style could be more diverse. For example, clarifications and comprehension checks could take place more frequently to enable verification of students' understanding of the topical matter. This would involve an end-to-end loop of student asking, Cipherbot answering and then verifying the student has understood. Currently, this happens rarely and focuses on lacking information rather than purposeful verification of learning. Furthermore, there appear to be technical glitches that should be addressed to ensure continuous accessibility and availability. More broadly, these highlight the fact that AI-based educational chatbot systems tend to be fairly complex–the LLM part is not the only consideration for their design, but they require multiple different engineering approaches, ranging from databases and analytics to other algorithmic support systems.

5.3 Limitations and Future Research

These findings indicate positive qualities in educational AI chatbots' ability to communicate with students. Nonetheless, the reader should be aware of certain limitations.

Firstly, future research should involve larger samples to provide more generalizable findings. There is a need to replicate these findings in other educational settings, such as different courses, subjects, and even countries. More research on the stability of educational AI chatbots' communication patterns across human factors like culture, gender, and age is direly needed.

Second, while we considered the variability and patterns in Cipherbot's communication style, we did not specifically analyze *students'* communication patterns with Cipherbot. Naturally, communication follows the sender-recipient (i.e., encode-decode) paradigm, so viewing this communication in a dualistic, two-way unit of analysis would provide additional insights into communicational factors surrounding AI-student interaction. It would also be interesting to investigate how tools like Cipherbot affect students' AI literacy: How do students "survive in the wild" (e.g., when interacting with GPT-4 or other GenAI tools) after interacting with education-adjusted AI chatbots?

Third, future research could analyze the long-term effects of educational AI chatbots on students' perceptions and behavior. For example, it is possible that students adapt their behavior, accommodating to challenges posed by this educational technology. We already saw it in some fashion in the feedback where students stated they refreshed the page multiple times to get Cipherbot working. Similarly, the perceptions of the utility of this technology can change over time–previous research has reported that students' appraisals may decrease once the novelty effect wears out [37], but it is also possible that students learn to use the technology better over time, extracting more value from it and perceiving it as more useful.

Fourth, when interpreting Cipherbot's behavior, we must bear in mind that its LLM is based on OpenAI's GPT models. Therefore, these models' general tendencies are inherited from Cipherbot's communication style to some extent, although Cipherbot's own system prompts also affect the communication patterns. Varying the system prompts systematically and testing alternative LLMs (e.g., Google's Gemini, Anthropic's Claude) pose natural follow-up study directions. These systems have different system prompts and guardrails, so their communication styles differ.

Fifth, as mentioned, aside from the underlying LLM's system prompt, Cipherbot's communication style is naturally affected by its own system prompt. Therefore, systematic testing of the effect of modifications in Cipherbot's system prompt on its communication style is needed.

Sixth, Cipherbot's answers depend on the learning materials. When information cannot be found in the learning materials, Cipherbot will not provide a substantial answer, even if the underlying LLM would have valid information on the topic (which is not grounded in the learning materials). There could be a way of relying on the LLM's general information while indicating to the students that this information has not been validated by the course educator. Overall, there is a need for studying alternative strategies to handle this *out-of-learning-materials problem*.

Seventh, GenAI has sparked ethical concerns among educators [11]. A key ethical issue with using Cipherbot is whether students might simply copy its responses for their assignments, which is unethical, or whether they would use the answers to actually advance their learning, which is considered ethical. Addressing this risk is still unresolved and remains a general challenge of educational AI tools. Overall, the use of AI in education may, at worst, lead to superficial learning habits and diminish students' social and critical thinking skills, which the educational community definitely does not desire. Therefore, mitigating these risks is a key area for future research. This concern, shared across the educational technology community [11, 37], emphasizes the fact that systems like Cipherbot are quickly moving from conceptual analysis into actual classroom settings, thereby requiring practical guidelines for implementation.

REFERENCES

[1] Behnoush Abdollahi and Olfa Nasraoui. 2018. Transparency in fair machine learning: the case of explainable recommender systems. In *Human and machine learning*. Springer, 21–35.

[2] Tanvirul Alam, Akib Khan, and Firoj Alam. 2020. Punctuation restoration using transformer models for high-and low-resource languages. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020. 132–142.

Retrieved December 12, 2023 from https://aclanthology.org/2020.wnut-1.18/
[3] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will Affective Computing Emerge From Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT. *IEEE Intelligent Systems* 38, 2 (2023), 15–23.
[4] Amanda Barany, Nidhi Nasiar, Chelsea Porter, Andres Felipe Zambrano, Alexandra L. Andres, Dara Bright, Mamta Shah, Xiner Liu, Sabrina Gao, Jiayi Zhang, Shruti Mehta, Jaeyoon Choi, Camille Giordano, and Ryan S. Baker. 2024. ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development. In *Artificial Intelligence in Education*, 2024. Springer Nature Switzerland, Cham, 134–149. https://doi.org/10.1007/978-3-031-64299-9_10
[5] Jose Belda-Medina and José Ramón Calvo-Ferrer. 2022. Using chatbots as AI conversational partners in language learning. *Applied Sciences* 12, 17 (2022), 8427.
[6] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Pers Ubiquit Comput* 26, 1 (February 2022), 95–119. https://doi.org/10.1007/s00779-021-01582-9
[7] Matt Bower. 2019. Technology-mediated learning theory. *Brit J Educational Tech* 50, 3 (May 2019), 1035–1048. https://doi.org/10.1111/bjet.12771
[8] Grant Cooper. 2023. Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *J Sci Educ Technol* 32, 3 (June 2023), 444–452. https://doi.org/10.1007/s10956-023-10039-y
[9] Antun Drobnjak, Ivica Botički, Peter Seow, and Ken Kahn. 2023. Learning with Conversational AI and Personas: A Systematic Literature Review. In *The 31st International Conference on Computers in Education*, 2023. 1003–1008. Retrieved February 17, 2024 from https://www.croris.hr/crosbi/publikacija/prilog-skup/806796
[10] Laila Dybkjær and Niels Ole Bernsen. 2000. Usability issues in spoken dialogue systems. *Natural Language Engineering* 6, 3–4 (2000), 243–271.
[11] 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
[12] Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D. Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, Lennart E. Nacke, and Pan Hui. 2024. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans* 2, 1 (January 2024), 100027. https://doi.org/10.1016/j.chbah.2023.100027
[13] Ahmad Haidar. 2024. ChatGPT and Generative AI in Educational Ecosystems: Transforming Student Engagement and Ensuring Digital Safety. In *Preparing Students for the Future Educational Paradigm*. IGI Global, 70–100. https://doi.org/10.4018/979-8-3693-1536-1.ch004
[14] Hazrina Hamid, Khadjizah Zulkifli, Faiza Naimat, Nor Liana Che Yaacob, and Kwok Wen Ng. 2023. Exploratory study on student perception on the use of chat AI in process-driven problem-based learning. *Currents in Pharmacy Teaching and Learning* 15, 12 (2023), 1017–1025.
[15] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, September 10, 2019. ACM, BELFAST United Kingdom, 207–214. https://doi.org/10.1145/3335082.3335094
[16] Wayne Holmes and Ilkka Tuomi. 2022. State of the art and practice in AI in education. *European Journal of Education* 57, 4 (2022), 542–570. https://doi.org/10.1111/ejed.12533
[17] Michael C. Jensen. 2002. Value maximization, stakeholder theory, and the corporate objective function. *Business ethics quarterly* (2002), 235–256.
[18] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. 1827–1843. Retrieved May 9, 2024 from https://aclanthology.org/2023.findings-emnlp.123/
[19] Gorgi Kakasevski, Martin Mihajlov, Sime Arsenovski, and Slavcho Chungurski. 2008. Evaluating usability in learning management system moodle. In *ITI 2008 - 30th International Conference on Information Technology Interfaces*, June 2008. 613–618. https://doi.org/10.1109/ITI.2008.4588480
[20] Muhammad Fawad Akbar Khan, Max Ramsdell, Erik Falor, and Hamid Karimi. 2023. Assessing the Promise and Pitfalls of ChatGPT for Automated Code Generation. https://doi.org/10.48550/arXiv.2311.02640
[21] 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
[22] 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.
[23] Anna R. McAlister, Saleem Alhabash, and Jing Yang. 2024. Artificial intelligence and ChatGPT: Exploring Current and potential future roles in marketing education. *Journal of Marketing Communications* 30, 2 (February 2024), 166–187. https://doi.org/10.1080/13527266.2023.2289034
[24] 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
[25] Julia-Astrid Moldt, Teresa Festl-Wietek, Amir Madany Mamlouk, Kay Nieselt, Wolfgang Fuhl, and Anne Herrmann-Werner. 2023. Chatbots for future docs: exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. *Medical Education Online* 28, 1 (December 2023), 2182659. https://doi.org/10.1080/10872981.2023.2182659
[26] Ha Nguyen, Nate Stott, and Vicki Allan. 2024. Comparing Feedback from Large Language Models and Instructors: Teaching Computer Science at Scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (*L@S '24*), July 15, 2024. Association for Computing Machinery, New York, NY, USA, 335–339. https://doi.org/10.1145/3657604.3664660
[27] 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*, July 09, 2024. ACM, Atlanta GA USA, 279–283. https://doi.org/10.1145/3657604.3664690
[28] Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle &Riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. In *Artificial Intelligence in Education*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos and Ig Ibert Bittencourt (eds.). Springer Nature Switzerland, Cham, 75–90. https://doi.org/10.1007/978-3-031-64302-6_6
[29] Kyong Jin Shim, Thomas Menkhoff, Lydia Ying Qian Teo, and Clement Shi Qi Ong. 2023. Assessing the effectiveness of a chatbot workshop as experiential teaching and learning tool to engage undergraduate students. *Educ Inf Technol* 28, 12 (December 2023), 16065–16088. https://doi.org/10.1007/s10639-023-11795-5
[30] Jagdish Singh, Minnu Helen Joesph, and Khurshid Begum Abdul Jabbar. 2019. Rule-based chabot for student enquiries. In *Journal of Physics: Conference Series*, 2019. IOP Publishing, 012060. Retrieved May 5, 2024 from https://iopscience.iop.org/article/10.1088/1742-6596/1228/1/012060/meta
[31] Andreas Sonderegger and Juergen Sauer. 2010. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied ergonomics* 41, 3 (2010), 403–410.
[32] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* (1994), 273–285.
[33] Sandeep A. Thorat and Vishakha Jadhav. 2020. A review on implementation issues of rule-based chatbot systems. In *Proceedings of the international conference on innovative computing & communications (ICICC)*, 2020. . Retrieved May 5, 2024 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id$=$3567047
[34] Wolfgang Wahlster. 2013. *Verbmobil: foundations of speech-to-speech translation*. Springer Science & Business Media. Retrieved May 1, 2024 from https://books.google.com/books?hl$=$en&lr$=$&id$=$NoqrCAAAQBAJ&oi$=$fnd&pg$=$PR5&dq$=$VERBMOBIL&ots$=$qLitKoCWEd&sig$=$H9DlDGNoH7UmFkUO5LbhqMiE4_c
[35] Hua Wang, Sneha Gupta, Arvind Singhal, Poonam Muttreja, Sanghamitra Singh, Poorva Sharma, and Alice Piterova. 2022. An artificial intelligence chatbot for young people's sexual and reproductive health in India (SnehAI): instrumental case study. *Journal of Medical Internet Research* 24, 1 (2022), e29969.
[36] Micah Watanabe, Megan Imundo, Katerina Christhilf, Tracy Arner, and Danielle S. Mcnamara. 2024. Building Reading Comprehension and Knowledge with iSTART: An ITS to Provide Formative Feedback in Reading Instruction at Scale. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (*L@S '24*), July 15, 2024. Association for Computing Machinery, New York, NY, USA, 510–513. https://doi.org/10.1145/3657604.3664710
[37] 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
[38] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10, 5 (2023), 1122–1136.
[39] Rina Zazkis and Roza Leikin. 2007. Generating examples: From pedagogical tool to a research tool. *For the learning of mathematics* 27, 2 (2007), 15–21.