

Investigating the Usability of an Educational AI Chatbot by Middle School Teachers and Students for Enhanced Learning

Kholoud Khalil Aldous
Qatar Computing Research Institute
Doha, Qatar
kkaldous@hbku.edu.qa

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

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

Jinan Y. Azem
Qatar Computing Research Institute
Doha, Qatar
jazem@hbku.edu.qa

Johanne Medina
Qatar Computing Research Institute
Doha, Qatar
jomedina@hbku.edu.qa

Salar M. Khan
Qatar Foundation
Doha, Qatar
sakan5@qf.org.qa

Amani Alabed
University of Doha for Science and Technology
Doha, Qatar
amani.alabed@udst.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute
Doha, Qatar
jjansen@acm.org

Abstract—The evaluation of AI educational dialogue systems for middle-school students has been limited. This study employs a state-of-the-art AI chatbot that answers students' questions exclusively based on educator-provided learning materials. Following an initial assessment by 10 middle-school teachers using the *Chatbot Suitability Questionnaire*, we conducted a mixed-method intervention user study involving 18 middle-school students to explore usability expectations, knowledge acquisition, and learning experience. Findings reveal that interacting with the AI chatbot enhanced self-reported knowledge acquisition, improved learning outcomes measured by test scores, and maintained student interest for future use. The chatbot achieved a usability score of 71.44% (± 16.28), attributed mainly to its high answer accuracy and effective interpretation of student input. Error management emerged as the most critical usability factor.

Index Terms—Cipherbot, middle-school education, conversational agent, chatbot, AI chatbot, usability

I. INTRODUCTION

Encouraging students to ask questions is essential to effective learning, particularly in middle-school (MS) settings [1]. Question-based interaction allows students to tailor their learning experiences and improve retention [1]. However, students often hesitate to ask questions during class due to factors such as fear of negative reactions from educators or peers [2]. Educators, on the other hand, face the repetitive challenge of addressing the same questions multiple times [2]. Some educators even discourage questions that are mainly factual or procedural, as these tend to limit opportunities for deeper discussions and critical thinking [3], [4]. Furthermore, responding to numerous student inquiries can be time-intensive [5], a

difficulty compounded in online education where immediate teacher feedback may be unavailable.

AI chatbots have emerged as promising tools for addressing students' questions through real-time interactions [6]. AI chatbots may enhance cognitive abilities by supporting critical, creative, and reflective thinking in class dialogues, although concerns of their contrary effects also persist [7]. Integrating AI into classrooms poses challenges, particularly in ensuring that it enriches learning [8] and meets students' usability and user experience (UX) expectations [9].

Research on the usability of AI chatbots in MS education remains limited [10]. MS students, usually in fifth to eighth grade, are in a crucial developmental stage, making them a crucial demographic for AI-based educational tools [11]. Integrating AI into K-12 education has shown positive learning outcomes [12]. However, the adoption of AI technologies in these settings is still limited, in part due to educators' reluctance stemming from perceived challenges of implementation [13]. Therefore, while AI technologies have gained significant traction in education, there is a lack of understanding of younger (such as MS) students' and educators's expectations and experiences with AI chatbots in education [14].

Bridging this gap is crucial as MS represents a fundamental stage in which students form their technological literacy and learning habits. The MS educational context presents unique challenges, including age-appropriate content restrictions, the timeliness of knowledge bases, and response accuracy [15], making this understanding more vital. Moreover, students' expectations and understanding of AI chatbots influence their willingness to adopt such educational technologies [16]. To

this end, this research investigates what MS educators and students expect from an educational chatbot, how it impacts the learning experience, how they assess its usability, and what attributes they consider most significant. More specifically, we address the following research questions (RQs):

- RQ1: *What do middle-school educators and students expect from an AI chatbot designed for educational purposes?*
- RQ2: *How do middle-school educators and students evaluate the usability of an AI chatbot?*
- RQ3: *How does interacting with an AI chatbot influence middle-school students' knowledge acquisition, learning experience, and engagement?*

We conduct two studies to address these RQs. Study 1 centers on MS educators to gather initial expectations and validate the chatbot's design and applicability, aligning primarily with **RQ1** and **RQ2**. Study 2 evaluates MS students' knowledge acquisition and usability, aligning with **RQ3** and reinforcing insights relevant to **RQ1** and **RQ2**. We employ a mixed-method approach. Qualitative analysis of interviews with educators and students addresses **RQ1**, focusing on their expectations of AI chatbots. We deploy the *Chatbot Usability Questionnaire* (CUQ) [17] to assess the chatbot's usability, complemented by qualitative analysis of student interviews to identify key usability attributes, addressing **RQ2**. To examine **RQ3a**, we quantitatively measure students' self-reported knowledge acquisition before and after their study session. We also explore how the AI chatbot contributes to students' learning experience (**RQ3b**) and their future interest in using AI chatbots for learning (**RQ3c**).

II. RELATED WORK

Chatbots, also known as conversational agents (CAs) or Question and Answering (Q&A) platforms, have become prominent in Human-Computer Interaction (HCI), Information Systems (IS), and Education Technology (ET) for their ability to simulate human interactions, personalize support, and enhance engagement [18], [19]. They support diverse educational functions, from instruction to assessment [20], [21]. Multiple attempts to develop educational chatbots have been documented [6], [11], evolving from rule-based systems limited in handling complex queries [22]–[25] to more adaptive large language model (LLM)-based systems capable of natural dialogue and personalized interactions that adjust to individual learner needs [18], [26].

Recent studies show that educational chatbots can promote self-regulated learning (SRL) by guiding students' independent learning and promoting active, student-driven participation [27], [28]. However, integrating these tools into MS education presents challenges in meeting student expectations and engagement, key aspects explored in **RQ1** and **RQ2**.

Understanding MS students' expectations is essential for designing effective educational chatbots. While research identifies a range of expectations from emotional engagement to reliability [29], few studies focus on MS students' expectations. **RQ1** addresses this gap by identifying student expectations.

A systematic review of 83 chatbot studies [30] revealed a focus on simulated interactions with prototypes rather than existing chatbots, with many chatbots falling short due to poor design, limited conversational capabilities, or unrealistic user assumptions [9], [31], [32]. Gathering feedback from real-world experiences offers deeper insights on expectations and challenges, central to this study's focus.

AI chatbots have shown positive impacts on learning, particularly in higher education [18], [33]–[35], but research in MS settings remains limited. **RQ3** addresses how AI chatbots influence knowledge acquisition, learning experience, and sustained interest in educational technology. While Q&A chatbots offer personalized, on-demand support [2], their use in MS education is still underexplored [36]–[38].

Educational chatbot usability is vital for effective learning integration [39] and directly influences student motivation. While common tools like *System Usability Scale* (SUS) [40] and *User Experience Questionnaire* (UEQ) [41] often overlook conversational dynamics, the CUQ offers a more tailored evaluation [17], [42]–[46]. Though originally developed for healthcare, the CUQ aligns well with chatbot needs in education [17], [47]. Its strengths in assessing satisfaction, interaction, and error handling justify its use for evaluating usability in this MS context and addressing **RQ2**.

III. METHODOLOGY

This research included two studies to evaluate the AI chatbot's effectiveness and usability. Study 1 involved MS educators (n = 10) assessing the chatbot's classroom relevance, focusing on its conversational use for MS students. It addressed **RQ1** by exploring educators' expectations, and contributed to **RQ2** by offering early insights on usability. Study 2 involved MS students (n = 18) and examined their expectations, learning experiences, and perceptions of the chatbot's usability. A mixed-method approach using surveys, the CUQ, and interviews provides a rounded perspective to Cipherbot's educational value.

A. Cipherbot: An Educational AI Chatbot

In this study, we employ Cipherbot¹, a state-of-the-art, educational AI chatbot [48], [49]. Cipherbot differs from many other LLM-based applications primarily in its custom-made design for educational settings. Unlike general-purpose LLM tools, such as ChatGPT, Cipherbot is specifically configured to respond exclusively based on educator-vetted learning materials, ensuring that its answers are accurate, age-appropriate, and aligned with educational standards. This customization mitigates the risks of hallucinations, a challenge often associated with generic LLM-based tools. Additionally, Cipherbot integrates OpenAI's GPT-4, strengthened by a Retrieval-Augmented Generation (RAG) framework to enhance the reliability of its responses. These functionalities set Cipherbot apart from general-purpose AI chatbots, such as ChatGPT, as a unique educational AI tool to meet the specific needs of MS students and educators.

¹<https://cipherbot.qcri.org/>

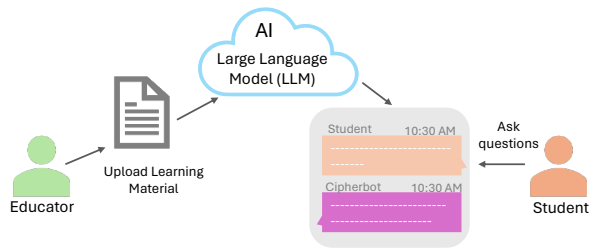


Fig. 1. CIPHERBOT’s workflow shows an educator uploading learning material and a student chatting with CIPHERBOT by asking questions.

TABLE I
CIPHERBOT CHAT SYSTEM PROMPT.

| Condition | Instruction |
|---------------------|---|
| Initial Prompt | You are a teaching assistant who answers based on the retrieved documents. You must scrutinize the retrieved documents based on the user’s query language. You must formulate a response in the same language as the retrieved documents’ language. You must be capable of offering answers and explanations in different languages if requested. If the question cannot be answered from the given data, be sweet and apologetic and guide them in the right direction of what questions can be asked instead. |
| School Level | If the <code>school_level</code> parameter is specified, you must use the corresponding school level language, ensuring that the vocabulary and complexity are suitable for the intended audience. |
| External Knowledge | Unless specified otherwise, you must not use external knowledge beyond the retrieved documents for your response. |
| Follow-Up Questions | If enabled, with your answer, you must provide up to three possible contextually appropriate follow-up questions based on the retrieved documents. |

Figure 1 shows CIPHERBOT’s workflow. Educators can create classes (e.g., History), add subtopics (e.g., AI Zubara Fort), upload materials (PDFs, videos, etc.), and share them via links or QR codes. Students access resources and engage in dialogue with CIPHERBOT, which responds based on uploaded content.

CIPHERBOT is programmed with conditional instructions that tailor its responses based on student queries, education level, and available information. The initial prompt defines the core behavior of the agent, such as language consistency, cross-language adaptability, and graceful handling of unknown answers. Acting as a “teaching assistant”, CIPHERBOT ensures that responses are appropriate in tone and context [50]. Table I displays the CIPHERBOT system prompt. The parameter *School Level* directs CIPHERBOT to modify language complexity to suit the audience. The parameter *External Knowledge* limits CIPHERBOT from retrieving documents unless specified otherwise. When activated, the parameter *Follow-Up Questions* allows CIPHERBOT to generate up to three relevant follow-up questions based on the retrieved documents.

IV. STUDY 1: ASSESSMENT OF THE AI CHATBOT BY MS EDUCATORS

Study 1 aimed to evaluate the practical use and effectiveness of the AI chatbot in MS settings. Its two main goals were to validate the construct of the Chatbot Suitability Questionnaire (CSQ) and assess whether CIPHERBOT is an appropriate tool for MS students. Even though Study 1 provides insights into teacher expectations, its primary focus is to validate that the chatbot is suitable for classroom use and that the CSQ effectively captures relevant aspects of chatbot usability for educators. This validation sets the stage for additional in-depth data collection in Study 2.

A. Participants

We recruited 10 subject matter experts (SMEs), experienced MS educators actively teaching in MS classrooms, through Upwork, inviting them based on the platform’s profile order. Only invited individuals could join, with an 80% response rate and \$40 USD compensation. Participants (M age = 34.8, range = 25–45) included eight from the US and two from the Philippines (nine females, one male). On average, the participants had 10.9 years of classroom experience (max = 22, min = 3, SD = 6.44), with class sizes averaging 31.4 students (max = 65, min = 15, SD = 14.4), covering diverse grades and subjects. Participants expressed a positive perception of the use of AI in education, rating its use by students at 4.0 (SD = 1.25) and by teachers at 4.8 (SD = 0.42) [51].

B. Procedure

We conducted a mixed-methods evaluation of the AI chatbot with ten MS educators via Microsoft Teams. Each remote session lasted 40–45 minutes and included a walkthrough of the same task planned for MS students (see Section V). Participants engaged in a chat dialogue with the AI chatbot, asking at least eight questions; four returned to test additional queries or the system’s error-checking capabilities.

After the task, participants completed the CUQ [17] to assess usability and the ten-item CSQ (see the Appendix), modeled after the SUS [40], to evaluate perceived MS classroom suitability. The CSQ covered three constructs: *Interface and Design* (3 items), *Language and Communication* (3 items), and *Interest & Beneficial* (4 items). The scores were scaled to a 100-point maximum.

Participants followed a think-aloud protocol throughout the study, including during surveys [52], providing real-time insights into their experiences, needs, and challenges with the AI chatbot [53]. They also answered open-ended questions about potential classroom use and attitudes toward AI in education.

C. Results of MS Educators Study

To analyze the qualitative data, we used thematic analysis [54] on the think-aloud transcripts, applying a line-by-line open coding process. Two researchers independently coded 2,207 relevant phrases (excluding administrative comments), resolving discrepancies through discussion and refining distinct categories. Inter-Rater Reliability, measured by Cohen’s

Kappa, was $\kappa = 0.81$, indicating strong agreement, confirming the robustness of the thematic coding.

Three main themes emerged: *Interface, Aesthetics, & Design, Language & Communication, and Interest & Beneficial*. Comments were assigned to subcategories themes, with disagreements resolved collaboratively to ensure reliability.

1) *RQ1: What do middle-school educators expect from an AI chatbot designed for educational purposes?:* Of the 2,207 statements, 745 aligned with the key themes: 155 (20.8%) on *Interface, Aesthetics, & Design*, 225 (30.2%) on *Language & Communication*, and 365 (49.0%) on *Interest & Benefits*, making sure that the themes truly represented the insights shared by the educators during interviews.

Concerning *Interface, Aesthetics, & Design*, there were many positive comments. For example: “*Yeah, I think the speed of it and just the option would be really helpful for kids especially I think it is also be very helpful for reaching kids of all ability levels.*” (P-6), “*Right. So I’m thinking that, yeah, so as I’m going through, I think it’s quick, which I think is good for the kids like attention span and whatnot.*” (P-7), and “*MS students would like to find the AI chatbot interface easy to navigate and understand, I think.*” (P-5).

One concern raised was the need for initial guidance. “*Yeah, I definitely think some more instruction at the beginning. I definitely think I needed a little help from you on that end. [...] I think if I had a little trouble in the beginning maybe. But also kids are very savvy with computers, so.*” (P-2), which we addressed by adding context instructions in the main study.

For *Language & Communication*, the participants found the dialogue appropriate (e.g., “*Yeah, all the responses. I think the responses were really good and they were like a good length for that age group. It wasn’t too much. It was like pretty easily digestible.*” (P-1), “*Let’s see the language appropriate and understandable, yes. Which is why I asked about the grade levels and the flexibility because I think that plays a pivotal role into chatting. You know the answers and whether the students can comprehend or understand them.*” (P-3)).

Possible concerns with language abilities were noted: (“*As far as the language, I think it was maybe a little advanced for like 9 year olds.*” (P-1)), prompting us to adjust the chatbot’s language by aligning it with course grade level settings.

Regarding *Interest & Benefits*, the feedback was highly positive: “*So now this is forcing me to read more. The method I’m doing now instead of just copy pasting that the copy pasting the question, getting the answer I’m reading and thinking [...]*” (P-2), “*It would be very helpful for students to be able to use that as a study guide, you know, so then no longer you can, you can provide the student with the notes and the PowerPoints and things like that and then they can use the AI chatbot to help dig in the curiosity to.*” (P-8), and “*Yeah, there’s no questions with the educational benefits and the supports it’s giving on the objectives that the teacher wanted to.*” (P-9).

Some suggested feature enhancements, such as page references: (“*Can you tell me what specific part of page 12? And you’re working on what you’re trying to achieve.*” (P-10)).

TABLE II

AVERAGED RESPONSES ACROSS ALL PARTICIPANTS ON 5-POINT LIKERT SCALE. CRONBACH’S ALPHA MEASURES THE INTERNAL CONSISTENCY OF MULTIPLE SURVEY ITEMS.

| Constructs | Mean | SD | Cronbach’s Alpha |
|------------------------------------|------|------|------------------|
| Interface & Design (3 items) | 3.97 | 1.33 | 0.74 |
| Language & Communication (3 items) | 4.70 | 0.60 | 0.71 |
| Interest & Beneficial (4 items) | 4.60 | 0.67 | 0.88 |

2) *RQ2: How do middle-school educators evaluate the usability of an AI chatbot?:* Participants rated the chatbot’s usability with an average CUQ score of 78.8 (max = 100.0, min = 46.9, SD = 16.4), indicating a *Good* rating [17]. SMEs rated its suitability for MS students with an average CSQ score of 77.0 (max = 99.0, min = 57.5, SD = 11.5). Table II summarizes the three CSQ constructs.

Cronbach’s Alpha values ($\alpha = 0.71$ – 0.88) indicate reliable constructs (see Table II). Participants found the interface user-friendly, the language age-appropriate, and the content engaging and educationally beneficial for MS students.

3) *MS Educators Summary of Results:* Consistent with the quantitative evaluation, the qualitative insights were largely positive. The CUQ and CSQ scores offered a measurable usability assessment, with most participants rating the AI chatbot from good to excellent. Qualitative analysis added context to the ratings, emphasizing UX, usability, language appropriateness, and response speed, which supported moving forward with testing the chatbot in Study 2 with MS students.

V. STUDY 2: EVALUATION OF THE AI CHATBOT BY MS STUDENTS

We conducted an intervention user study [55], [56] (i.e., studies that compare outcomes in participants receiving the intervention to outcomes before the intervention was introduced) with MS students at their school, following both the school’s research protocols and obtaining an Institutional Review Board (IRB) approval.

A. System Setup

We set up the AI chatbot by creating a class titled ‘Social Science’ and a sub-class called ‘AI Zubara Fort,’ which included a two-page document with relevant content. During the study, students interacted with the chatbot through Q&A based on this material. Figure 2 shows an example of a student interacting with the AI chatbot to learn about “AI Zubara Fort”, with responses based only on the uploaded material. The topic was selected for its cultural relevance and grade-level suitability, as confirmed by two school teachers.

We created individual Cipherbot credentials for each participant and provided the login details and class link during the user study session. Each student was assigned a computer at a designated workstation, where the Cipherbot chat page was pre-opened. Figure 3 shows a photo of the workstation setup.

B. Participants and Procedure

To recruit participants, we collaborated with the school’s administrators. We recruited 18 students (9 females, 9 males)

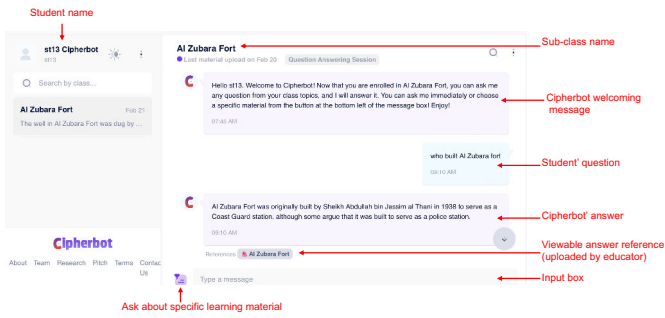


Fig. 2. A Cipherbot screenshot of the student interface showing a welcome message for the student (st13) and the student questions and Cipherbot answers.

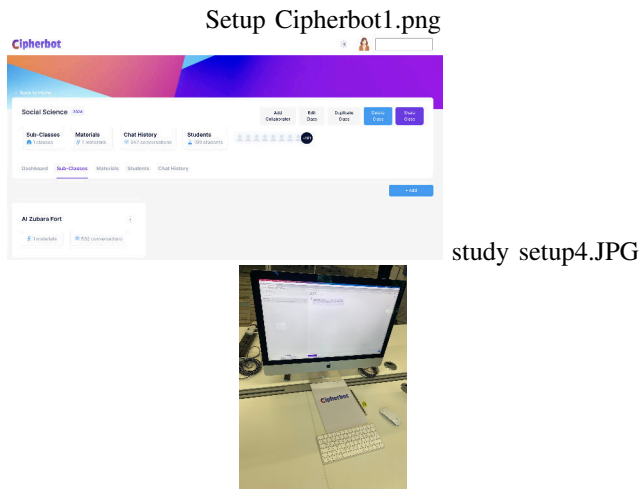


Fig. 3. On the left side, the setup of the class on Cipherbot from the instructor front-end. On the right side, the working station was set up with a computer for the user study session, displaying the Cipherbot chat page. Hard copies of the pre-session and post-session surveys and a fill-in-the-blank sheet were provided with pencil and rubber.

aged 9–13, with an average age of 11 (± 1.3) years. All participated in English, the school’s main instruction language.

The procedure included a (1) a pre-session survey assessed students’ prior knowledge and demographics; (2) students accessed the AI chatbot with no prior training; (3) they completed the fill-in-the-blank tasks on AI Zubara Fort by asking the chatbot questions; (4) a post-session survey measured knowledge gains and chatbot usability (CUQ); and (5) brief interviews gathered qualitative feedback. The session lasted about an hour. The session included a pre-survey, a chatbot task, post-survey, and interviews, lasting about an hour.

C. Measures

We used both quantitative and qualitative measures.

1) *Quantitative Analysis*: To assess learning gains, we used a pre-session survey to capture students’ demographics and prior knowledge of AI Zubara Fort, and repeated the same knowledge question in the post-session survey. This measured students’ self-evaluation of knowledge before and after the session (**RQ3**). The survey question we used in both the pre

and post-surveys is “How much do you know about AI Zubara Fort in Qatar?”, which has a 5-point Likert scale from “I’ve never heard of it” to “I’m like an expert!” A paired t-test assessed the chatbot’s effect on knowledge.

In the post-session survey, we used the CUQ [17] to assess the AI chatbot’s usability (**RQ2**). Completed after the task, the CUQ includes 16 equally weighted items and is comparable to the SUS, with 68/100 as the usability benchmark [17]. Higher scores indicate good usability; lower ones suggest room for improvement. Table IV lists all CUQ questions, covering chatbot personality, onboarding, navigation, input understanding, response time, error handling, and overall intelligence. Odd-numbered items are positive while even-numbered are negative. Some questions were rephrased to suit MS students.

We calculated the CUQ of each student using Equation 1.

$$CUQ = \left(\left(\sum n_{\text{odd}}^Q - 8 \right) + \left(40 - \sum n_{\text{Even}}^Q \right) \right) / 64 \times 100 \quad (1)$$

During interviews, we collected exploratory data on participants’ online habits and prior chatbot use to assess their technology and chatbot expertise. After the study, we analyzed all chatbot interactions, reporting the number of questions asked, average and SD of questions per participant, average and SD of characters per question/answer, and response time. The lead author graded the AI Zubara Fort worksheet, assigning 1 point per correctly answered blank (10 total).

2) *Qualitative Analysis*: In addition to completing the AI Zubara Fort task and pre/post-surveys, participants were interviewed about their general chatbot experience and expectations, as well as their experience and impressions of the AI chatbot’s personality, navigation, comprehension, and error handling. The interviews concluded with questions about future use.

We used interview data to address **RQ1**, **RQ2**, and **RQ3**, applying thematic analysis [54] using line-by-line open coding. Two researchers independently coded transcripts, resolved differences through discussion, and developed a shared coding guide. The final scheme was applied across all data. Cohen’s Kappa ($K = 0.823$) indicated a substantial level of agreement, confirming the robustness of the thematic coding. Three themes emerged around chatbot usability: *Ease of Use*, *Interaction Quality*, and *Error Handling*, each that detailed specific UX aspects.

D. Results of Study 2

Students reported spending an average of 3 hours ($SD = 2.4$) online daily. Most students ($n = 10$, 55.5%) indicated that they use chatbots like ChatGPT for non-academic purposes (e.g., testing, cooking, writing), while only three students (16%) used them for school-related work. Students completed the fill-in-the-blank tasks using the chatbot, with an average score of 9 out of 10 ($SD = 0.69$, $Min = 8$) and an average completion time of 13.6 minutes ($SD = 5.3$). Students asked 274 questions in total: 111 (40.5%) from females and 163 (59.5%) from males, averaging 15.2 questions per student ($SD = 7.7$). The questions averaged 29.8 characters ($SD = 22.8$),

and the responses averaged 265.6 characters (SD = 224.6). The average response time was 3.9 seconds (SD = 12.4).

1) *RQ1: What do middle-school students expect from an AI chatbot designed for educational purposes?:* Our thematic analysis of student interviews revealed key expectations for an educational chatbot: Accuracy (38.3%), speed (25.5%), ease of use and interaction quality (both 10.6%), concise answers (8.5%), and effective input understanding and error handling (6.4%). Students emphasized the need for a usable interface, accurate and fast responses, human-like interaction, and concise replies to support learning and engagement.

Table III summarizes the main themes from our qualitative interview analysis, which are detailed in the following.

Ease of Use: Students emphasized the importance of a simple, user-friendly interface, expecting the chatbot to be “easy to use and understand” (P-17). Visual clarity and intuitiveness were common expectations (P-2, P-12).

Answers’ Accuracy & Errors: Clear, correct, and trustworthy responses were a top priority. As one student put it: “Give me the correct answers” (P-10). Students were frustrated by errors, unclear explanations, or overly complex responses (e.g., P-2, P-7, P-17).

Speed of Response: Students expected near-instant replies, such as “fast answers” (P-15), and showed low tolerance for delays or lags (P-4, P-10).

Interaction Quality: A natural, human-like interaction was valued, with one student noting the chatbot should feel like “talking to a person” (P-13). They were critical of bots that couldn’t learn or crashed under pressure (e.g., P-5, P-18).

Answers Length: Students preferred concise responses, noting that the chatbot should “not talk too long” (P-11). Long answers were perceived as overwhelming (P-9, P-14).

User Input & Errors: Students expected the chatbot to manage unclear or incorrect inputs. As one participant noted: “If I wrote the questions incorrectly, I think he [chatbot] may not know” (P-9). Misunderstandings and uncorrected errors were perceived as usability issues (P-11, P-12).

Addressing **RQ1**, students expect the chatbot to be easy to use, featuring a user-friendly interface and intuitive interactions. They value accurate, reliable answers, as errors hinder learning. Fast responses are essential to maintain focus, while human-like, conversational interactions boost engagement. Concise responses are preferred, as lengthy ones can be overwhelming.

2) *RQ2: How do middle-school students evaluate the usability of an AI chatbot?:* **Usability Score of the AI chatbot:** Using the post-session CUQ survey, we calculated the overall usability score (M = 71.44, SD = 16.28, median = 72.66) using Equation 1, and per-question averages (Table IV). Scores ranged from 31.25 to 98.43. Compared to the 68-point benchmark [17], the chatbot showed good usability. Although CUQ has no official threshold, we adopted 68 based on SUS conventions. The lowest score (31.25) came from a 9-year-old male with no daily online activity.

Usability Attributes: Our thematic analysis of student interviews identified key attributes in the CUQ usability cate-

gories: personality, navigation, understanding, responsiveness, and error handling [17]. *Personality* assessed human-likeness; *Navigation*, ease of use; *Understanding*, input comprehension; *Responsiveness*, reply helpfulness; and *Error handling*, how well mistakes were managed. Table III summarizes the themes, followed by detailed descriptions.

Personality: Answers’ Accuracy & Errors: Students generally found the chatbot’s answers efficient and helpful, with one noting: “Very straight away and effective answers” (P-8). Although some responses exceeded expectations (e.g., P-9), overly complex language sometimes challenged younger users’ understanding (e.g., P-13).

Knowledge Depth: Students appreciated the chatbot’s depth of knowledge, which enhanced its realism—“It knows a lot.” (P-6). However, limitations tied to educator-uploaded content made the experience feel less dynamic or human-like to some (P-18).

Interaction Quality: Many students highlighted the human-like quality of the chatbot’s dialogue, with one stating, “It seems like a real human with answers” (P-17), and others appreciating its ability to respond naturally to greetings (e.g., P-10, P-16). However, some found the interaction robotic—“It was like talking to a robot or AI” (P-3)—pointing to unnatural language, typing style, and overformality (P-4, P-5, P-7, P-9).

Answers Length: Students perceived non-realistic behavior because of the AI chatbot’s long answers, “long answers in a short period of time” (P-11); “not really because it gave me every single detail” (P-14); “long paragraph and it was like talking to a robot or AI.” (P-3).

Navigation: Speed of Response: Students found the AI chatbot easy to use because of its quick answers. Some examples are: “The answers here are quick” (P-2); “gave the answers right away” (P-10); “just I type the question and give me a full answer” (P-3).

Answers’ Accuracy & Errors: Students generally found it easy to locate relevant information using the chatbot, “I just type the question and [Cipherbot] gives me a full answer” (P-3), and appreciated its ability to handle long or misspelled queries (P-7, P-13). Responses were often described as “quick and easy to understand” (P-2), relevant (P-4), and accurate when questions were clearly phrased (P-6). Some students noted challenges when the chatbot failed to answer (P-10) or provided responses that were difficult to grasp (P-17).

User Input & Errors: Many students appreciated that the chatbot could interpret brief and informal input “I would write a bit and it would understand” (P-11), without needing proper punctuation or capitalization (e.g., P-12). Some found it challenging to frame questions clearly enough to get accurate responses. As one student noted, “The most challenging thing is that you must know what kind of question you should ask to reach your answer” (P-9). However, formulating questions was challenging for some, “I had to write two times. sometimes, the first time it said clarify the question.” (P-6); “Making up the questions properly to get correct answer” (P-8); “The most challenging thing is that you must know what kind of question you should ask so you can reach to your answer.” (P-9).

TABLE III
THE RESULTS OF THE THEMATIC ANALYSIS OF STUDENTS' INTERVIEWS.

| | Ease of Use | Answers' Accuracy & Errors | Speed of Response | Interaction Quality | Answers Length | User Input & Errors | Knowledge Depth | Learning Value |
|----------------------------|-------------|----------------------------|--------------------------------|-----------------------------|--------------------|---------------------|-----------------|----------------|
| User Expectations (RQ1) | X | X | X | X | X | X | | |
| Personality (RQ2) | | X | | X | X | | X | |
| Navigation (RQ2) | | X | X | X | X | X | X | |
| Understanding (RQ2) | | X | | | | X | | |
| Responsiveness (RQ2) | | X | X | X | X | X | | X |
| Learning experience (RQ3b) | | X | X | X | | | | X |
| | No Errors | Acknowledge uncertainty | Error Recovery & Understanding | Enhancing in Class Learning | Information Source | Homeworks & Exams | | |
| Error Handling (RQ2) | X | X | X | | | | | |
| Using CIPHERBOT (RQ3c) | | | | X | X | X | | |

TABLE IV

RESULTS ON A 5-POINT LIKERT SCALE SHOWING THE POSITIVE AND NEGATIVE FEATURES OF CIPHERBOT. THE HIGHER THE VALUE FOR POSITIVE FEATURES, THE BETTER, AND THE LOWER THE VALUE FOR NEGATIVE FEATURES, THE BETTER.

| Question | "Positive" features of the chatbot | M | SD |
|----------|--|------|------|
| 11 | Cipherbot responses were useful, appropriate, and informative. | 4.44 | 0.62 |
| 15 | Using Cipherbot was very easy. | 4.33 | 1.03 |
| 5 | Cipherbot did a good job explaining what it can do and its purpose. | 4.17 | 1.20 |
| 7 | Cipherbot was easy to navigate. | 4.06 | 0.80 |
| 13 | Cipherbot did a good job handling any mistakes or errors that came up. | 4.06 | 0.80 |
| 9 | Cipherbot understood what I was saying really well. | 3.89 | 1.08 |
| 3 | Cipherbot was welcoming during the initial setup. | 3.83 | 0.99 |
| 1 | Cipherbot felt like a real friend and was fun to talk to. | 3.56 | 1.10 |
| | "Negative" features of the chatbot | | |
| 2 | Cipherbot acted more like a robot than a buddy. | 3.33 | 1.24 |
| 8 | It was easy to get confused while using Cipherbot. | 2.50 | 1.25 |
| 14 | The chatbot seemed unable to handle any errors. | 2.44 | 1.10 |
| 10 | A lot of times, Cipherbot failed to understand my questions. | 2.28 | 1.07 |
| 12 | The answers Cipherbot said didn't match what I was asking. | 2.22 | 1.26 |
| 16 | Cipherbot was very complex to use. | 2.22 | 1.48 |
| 4 | Cipherbot seemed very unfriendly. | 2.06 | 0.87 |
| 6 | I couldn't tell what Cipherbot was supposed to help with. | 1.56 | 0.62 |

Answers length: Students appreciated the chatbot's short answers, "answers were not long" (P-4), "it easily answers long questions, not extra things that I do not need." (P-13). However, longer answers were found challenging for students, "The question: [...] it says a bunch of things, too much information." (P-13).

Interaction Quality: Many students found the chatbot easy and fun, likening it to normal texting. Examples: "Like how you text. Normally, you ask a question, and he answers." (P-8); "It was so easy. I just have to ask, and he already knows what I want to know [...] it's so much fun and beautiful." (P-9); "just I type the question and give me a full answer" (P-3). However, the robotic-like interaction is challenging when using the AI chatbot: "Just how he looked like a robot." (P-9)

Knowledge Depth: Since the chatbot knowledge was limited to one topic "Al Zubara Fort", students found it challenging when asking off-topic questions, "because I asked some questions and it kept on saying like sorry I don't know well." (P-1); "the requested information is not found" (P-10); "hard to find answers" (P-5).

Understanding: User Input & Errors: The chatbot generally understood students' questions, even with spelling errors: "Most of the time, it understood me. I purposely misspelled some stuff, but it still understood" (P-13). Some students

adapted their phrasing to be more natural for better comprehension (P-9). However, the chatbot struggled with poor grammar (P-5), occasionally misunderstood questions (P-6), and failed to respond to queries beyond its limited topic scope (e.g., Al Zubara Fort) (P-10).

Answers' Accuracy & Errors: Students felt that the chatbot understood their questions, providing accurate responses—"it gave me the exact answer I wanted" (P-11), and sometimes anticipating what was asked (P-9). One summed it up as, "It answered everything" (P-4).

Responsiveness: Interaction Quality: The interaction was perceived as natural and human-like, with one student saying, "I kept on asking questions and it kept telling me stuff" (P-1). The chatbot provided correct answers even when students felt their own questions were unclear (P-7). However, a limitation was noted in the input method, as the study allowed only keyboard input, prompting suggestions for multimodal input such as voice: "Use only some things so I did not need to write" (P-18).

Answers' Accuracy & Errors: The students found the chatbot answers helpful because of their accuracy, "they give specific answers" (P-5), "give me a very specific and correct answer about what I need." (P-8). However, one student found the AI chatbot responses not helpful when they are wrong ("when it gave me the wrong ones", P-6).

User Input & Errors: The students found the AI chatbot answers helpful because the AI chatbot "understood despite typos" (P-7) on their questions.

Speed of Response: The students found the AI chatbot answers helpful because they are fast, "They [Cipherbot] gave the answer directly." (P-3); "found the answers I wanted. It was fast." (P-4); "It gave a straight answer" (P-18).

Answers Length: Some students appreciated the chatbot's informative and relevant answers, while others found responses too long or including unnecessary details, for example, "[Cipherbot] gave me the answer, they gave you a bunch of other things. I didn't need that" (P-3), and "Sometimes it was a little too long and not very convenient. Maybe shorter answers would be better" (P-8).

Learning Value: Students found the chatbot's answers helpful and educational, noting its role in enhancing their

knowledge: “*It made me learn a lot of things*” (P-6); “*It was very helpful and improved my knowledge about the topic I was learning*” (P-8); and “*All was helpful. Like I can read it and understand everything he [Cipherbot] said*” (P-9).

Error Handling: No Errors: Some students perceived no errors while using the AI chatbot, “*there was no mistakes*” (P-16); “*nothing went wrong*” (P-17); “*I don’t think I made a mistake*” (P-12). **Acknowledgment of uncertainty:** Students noted that the chatbot admitted uncertainty, replying with “*I’m sorry I don’t know*” (P-1) or asking for clarification when it did not understand (P-18). **Error Recovery and Understanding:** Students found the chatbot provided correct answers despite errors: “*it gave the right answers always*” (P-7); “*I made a lot of typos, and it understood the question correctly without me fixing mistakes*” (P-8); and “*It worked well even with typos or grammatical errors*” (P-13).

Addressing **RQ2**, MS students rated the chatbot’s usability at 71.44%. Their evaluations highlighted quick, accurate responses; understanding despite input errors; sufficient knowledge; human-like interaction; educational value; and effective error handling.

3) **RQ3: How does interacting with an AI chatbot influence middle-school students’ knowledge acquisition, learning experience, and engagement?:** **Knowledge Acquisition** Addressing knowledge acquisition, the results of the paired t-test indicated a significant increase in students’ knowledge about AI Zubara Fort from pre-session ($M = 2.33, SD = 1.08$) to post-session ($M = 3.11, SD = 1.08$), $t(17) = -3.50, p = .003$, suggesting that interaction with the AI chatbot significantly improved students’ learning.

Learning experience: In describing how Cipherbot supported their learning, students highlighted its educational value, human-like interaction, and ability to deliver fast, accurate answers.

Learning Value: Students reported gaining knowledge about the topic (e.g., “*it helped me know about Al Zubara Fort*” (P-4); “*help me learn*” (P-6)) and about using AI tools—“*helped me learn more about Cipherbot*” (P-7).

Interaction Quality: The chatbot’s human-like, contextual responses enhanced understanding—“*It gave answers with context... so I could understand more*” (P-3); “*It was informative and effective*” (P-8).

Speed of Response and Answers’ Accuracy & Errors: Students valued the quick, correct, and detailed responses: “*The answers came up quickly... the correct answer*” (P-2); “*helpful and detailed*” (P-17), and noted its ability to interpret imperfect questions (P-13).

Student Engagement Interest Student engagement was measured through the interview question: “*Are you interested in using Cipherbot for learning?*” Most students (15, 83.3%) responded “*Yes*”, showing strong potential for ongoing use. Two (11.1%) were unsure: one would use it for specific questions, the other preferred not to talk to a bot; another student (5.6%) found it boring and was not interested. Interested students highlighted three different use cases:

- **Enhancing In-Class Learning:** Using Cipherbot to clarify misunderstandings or ask class-related questions especially when misunderstanding in class, “*If I didn’t understand properly in class*” (P-7), students would use the AI chatbot to improve their understanding.
- **Information Source:** Quickly obtaining specific information with less effort than traditional search engines (e.g., “*Google does not give me the answers that I want*”, P-9).
- **Homeworks & Exams:** Helping with homework questions and exam preparation by memorization and answering support (P-3, P-6, P-11).

In summary, our investigation into the AI chatbot’s influence on MS students provided valuable insights. Our study found that interacting with the AI chatbot significantly improved students’ knowledge about AI Zubara Fort (**RQ3a**). The chatbot’s educational value, human-like interaction, and accurate, fast answers enhanced the learning experience (**RQ3b**). Most students (83.3%) expressed interest in ongoing use, highlighting its potential to support learning both in class and independently, promoting sustained engagement with educational technology among MS students (**RQ3c**).

VI. DISCUSSION

This study examined the usability and educational impact of the AI chatbot Cipherbot through three key questions: expectations (**RQ1**), usability assessment (**RQ2**), and effects on learning and engagement (**RQ3**).

The two studies, one with MS teachers and one with MS students, provide complementary and interconnected insights into Cipherbot’s usability and effectiveness. Educators emphasized accurate, age-appropriate, and contextually relevant content with necessary oversight, which aligns with students’ feedback on the need for clear, concise, accurate answers, human-like interaction, and effective error handling.

Regarding **RQ1**, students expected quick, accurate responses in an easy-to-use system. Most student expectations align with the CUQ survey [17]. For instance, the students’ preference for ease of use is captured by CUQ questions on intelligence (Q15–16) and navigation (Q7–8), while input comprehension (Q9–10), error handling (Q13–14), and chatbot personality (Q1–4) reflect students’ focus on human-like, responsive interaction. Although students valued concise answers, CUQ only measures response time and accuracy (Q11–12), not satisfaction with answer length or speed. Future surveys could include these aspects.

The only misalignment with students’ expectations was the onboarding process (questions 5-6) [17], which evaluates how well the AI chatbot explained its capabilities and purpose. This discrepancy in the CUQ results is possibly due to the overly complex or unengaging initial greeting of the chatbot. Adapting onboarding content for younger users is a potential area of future work.

For **RQ2**, the usability score shows the AI chatbot is effective but could improve personality and response time. MS educators rated usability higher than students, suggesting age and expertise influence how users perceive the chatbot’s

ease of use. The CUQ score of 71.44% (± 16.28) indicates that on average, students found the AI chatbot to be usable. However, the large standard deviation suggests a wide range of individual experiences. This variation is likely influenced by factors beyond the chatbot's performance, such as a student's prior digital literacy and their expectations of how an AI chatbot should interact. For instance, students with higher digital literacy are more likely to find the chatbot easier to use than those with lower literacy [57]. This interpretation is supported by further investigation of the lowest CUQ score (31.25%), which came from a male student (aged 9) who reported spending no time online on a typical day, suggesting that minimal digital exposure can significantly affect perceptions of usability.

For **RQ3**, our findings align with prior research [16], [23], [58], [59], confirming that AI chatbots support MS students' learning and outcomes. All students used Cipherbot for Q&A to complete their assigned task, although success varied based on digital literacy and self-regulated learning.

The findings suggest that AI chatbots are valuable tools for MS students, who valued the chatbot for in-class support, information access, and exam preparation. Effective use requires educator involvement to upload relevant material and guide ethical use, such as paraphrasing to avoid plagiarism and monitoring interactions [60].

Usability issues included mixed preferences for answer length, which stresses the need for a "shorten" or "lengthen" option. Even though accuracy built trust, some students struggled with complex language, suggesting a "simplify" button. Although the chatbot adjusted language by grade level (see Table I), individual differences still affected comprehension. The AI chatbot's ability to understand questions despite grammar or spelling issues enhanced usability and realism. Adding features such as related questions or clarification prompts (e.g., "Did you mean...?") could further aid comprehension.

Knowledge depth also shaped students' perceptions of usability, including navigation and answer relevance. Though the chatbot acknowledged off-topic queries, students often asked unrelated questions, likely influenced by experiences with open-ended tools like ChatGPT. Since access was limited to uploaded materials to ensure accuracy [61], clearly communicating this scope and adapting the CUQ items to the study context is recommended.

Implementing AI chatbots in real-world classroom settings involves integrating conversational agents to support both teaching and learning processes [62]. Beyond answering student queries, these conversational agents can facilitate skill development, support collaborative learning, and streamline administrative tasks, all while dynamically adapting to individual learning levels [63]. However, successful implementations depend on the selection of AI technology with robust natural language processing capabilities and on integration with existing learning management systems and alignment with pedagogical goals [64].

A. Implications

This research offers important theoretical and practical insights for AI chatbots, particularly for MS students, HCI, and usability. Theoretically, evaluating a knowledge-limited chatbot reveals the need to clarify the chatbot's limited knowledge base, and more class-specific CUQ questions. Since MS students prefer brief responses, chatbot evaluations should be adjusted to reflect this. Likewise, the CSQ should be adapted to better reflect student expectations and educational contexts.

Practically, this research points to key design improvements for improving chatbot usability for MS students: adding a "simplify" button for clearer responses, providing answer length customization, and improving question comprehension through detecting errors, acknowledging them, and suggesting corrected or related questions. These features would better align the AI chatbot with student needs. **HCI researchers**, should prioritize tailored usability studies for user-centered design. **AI engineers** can improve usability with features like personalized feedback. **Educators** can use AI chatbots to provide instant support and increase student engagement.

B. Limitations and Future Research

Limitations include a small sample size limiting generalizability and focus on a single topic restricting understanding of broader performance. Future research should involve larger, diverse groups and explore other subjects. Reliance on educator-provided content poses scalability challenges; automated, quality-controlled content curation is needed. Adding multimodal inputs like voice could improve accessibility for younger students. Enhancing natural language understanding to better handle complex queries and conducting comparative studies with other learning methods would further improve usability and clarify AI chatbots' classroom value.

VII. CONCLUSION

Cipherbot demonstrates the benefits of an AI educational chatbot. Our study with 18 MS students found it effective for knowledge acquisition, learning experience, and engagement. Students valued ease of use and answer accuracy, that influenced their evaluations. Findings stress the significance of understanding student expectations from AI chatbots, which strongly affects learning experience and usability assessments. This offers valuable insights for improving AI chatbots to better serve MS students' needs.

REFERENCES

- [1] K. A. Harper, E. Etkina, and Y. Lin, "Encouraging and analyzing student questions in a large physics course: Meaningful patterns for instructors," *Journal of Research in Science Teaching*, vol. 40, no. 8, pp. 776–791, 2003.
- [2] T. Wambganß, L. Haas, and M. Söllner, "Towards the design of a student-centered question-answering system in educational settings." in *ECIS*, 2021.
- [3] T. L. Good, R. L. Slavings, K. H. Harel, and H. Emerson, "Student Passivity: A Study of Question Asking in K-12 Classrooms," *Sociology of Education*, vol. 60, no. 3, pp. 181–199, 1987.
- [4] C. Chin and J. Osborne, "Students' questions: a potential resource for teaching and learning science," *Studies in Science Education*, vol. 44, no. 1, pp. 1–39, Mar. 2008.

- [5] L. M. Goldstein, "Questions and answers about teacher written commentary and student revision: teachers and students working together," *Journal of Second Language Writing*, vol. 13, no. 1, pp. 63–80, Mar. 2004.
- [6] M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Education and Information Technologies*, vol. 28, no. 1, pp. 973–1018, Jan. 2023.
- [7] H. B. Essel, D. Vlachopoulos, A. B. Essuman, and J. O. Amankwa, "ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs)," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100198, Jun. 2024.
- [8] D. Song, E. Y. Oh, and M. Rice, "Interacting with a conversational agent system for educational purposes in online courses," in *2017 10th International Conference on Human System Interactions (HSI)*, Jul. 2017, pp. 78–82.
- [9] S. Diederich, A. Brendel, S. Morana, and L. Kolbe, "On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research," *Journal of the Association for Information Systems*, Jan. 2022.
- [10] P. Zhang and G. Tur, "A systematic review of ChatGPT use in K-12 education," *European Journal of Education*, vol. 59, no. 2, 2024.
- [11] F. Martin, M. Zhuang, and D. Schaefer, "Systematic review of research on artificial intelligence in K-12 education (2017–2022)," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100195, 2024.
- [12] M. K. Wolf, "Interconnection between constructs and consequences: a key validity consideration in K–12 English language proficiency assessments," *Language Testing in Asia*, vol. 12, no. 1, p. 44, Oct. 2022.
- [13] K. Woodruff, J. Hutson, and K. Arnone, "Perceptions and Barriers to Adopting Artificial Intelligence in K-12 Education: A Survey of Educators in Fifty States," in *Reimagining Education - The Role of E-Learning, Creativity, and Technology in the Post-Pandemic Era*, S. Mistretta, Ed. IntechOpen, Sep. 2023.
- [14] R. F. Elmore, "Bridging the gap between standards and achievement: The imperative for professional development in education," *Secondary lenses on learning participant book: Team leadership for mathematics in middle and high schools*, pp. 313–344, 2002.
- [15] A. R. McAlister, S. Alhabash, and J. Yang, "Artificial intelligence and ChatGPT: Exploring Current and potential future roles in marketing education," *Journal of Marketing Communications*, vol. 30, no. 2, pp. 166–187, Feb. 2024.
- [16] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100033, 2021.
- [17] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. Mctear, "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*. BELFAST United Kingdom: ACM, Sep. 2019, pp. 207–214.
- [18] R. Wu and Z. Yu, "Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 10–33, 2024.
- [19] C. K. Lo, K. F. Hew, and M. S.-y. Jong, "The influence of ChatGPT on student engagement: A systematic review and future research agenda," *Computers & Education*, vol. 219, p. 105100, Oct. 2024.
- [20] R. Garg, H. Cui, S. Seligson, B. Zhang, M. Porcheron, L. Clark, B. R. Cowan, and E. Beneteau, "The Last Decade of HCI Research on Children and Voice-based Conversational Agents," in *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–19.
- [21] T. K. F. Chiu, Q. Xia, X. Zhou, C. S. Chai, and M. Cheng, "Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100118, Jan. 2023.
- [22] X. Zhai, X. Chu, C. S. Chai, M. S. Y. Jong, A. Istenic, M. Spector, J.-B. Liu, J. Yuan, and Y. Li, "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020," *Complexity*, vol. 2021, p. e8812542, Apr. 2021, publisher: Hindawi.
- [23] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachler, "Are We There Yet? - A Systematic Literature Review on Chatbots in Education," *Frontiers in Artificial Intelligence*, vol. 4, 2021.
- [24] S. Ruan, L. Jiang, J. Xu, B. J.-K. Tham, Z. Qiu, Y. Zhu, E. L. Murnane, E. Brunskill, and J. A. Landay, "QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland UK: ACM, May 2019, pp. 1–13.
- [25] R. Winkler, S. Hobert, A. Salovaara, M. Söllner, and J. M. Leimeister, "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*. Honolulu HI USA: ACM, Apr. 2020, pp. 1–14.
- [26] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, and R. Buyya, "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 19–23, Jan. 2024.
- [27] D. T. K. Ng, C. W. Tan, and J. K. L. Leung, "Empowering student self-regulated learning and science education through chatgpt: A pioneering pilot study," *British Journal of Educational Technology*, vol. 55, no. 4, pp. 1328–1353, 2024.
- [28] D. Teng, X. Wang, Y. Xia, Y. Zhang, L. Tang, Q. Chen, R. Zhang, S. Xie, and W. Yu, "Investigating the utilization and impact of large language model-based intelligent teaching assistants in flipped classrooms," *Education and Information Technologies*, vol. 30, no. 8, p. 10777–10810, Dec. 2024.
- [29] E. Svikhushina, A. Placinta, and P. Pu, "User Expectations of Conversational Chatbots Based on Online Reviews," in *Designing Interactive Systems Conference 2021*. Virtual Event USA: ACM, Jun. 2021, pp. 1481–1491.
- [30] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots," *International Journal of Human-Computer Studies*, vol. 151, p. 102630, Jul. 2021.
- [31] Z. Ashktorab, M. Jain, Q. V. Liao, and J. D. Weisz, "Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–12.
- [32] D. Lahoual and M. Frejus, "When Users Assist the Voice Assistants: From Supervision to Failure Resolution," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–8.
- [33] Y.-F. Lee, G.-J. Hwang, and P.-Y. Chen, "Impacts of an AI-based chatbot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation," *Educational technology research and development*, vol. 70, no. 5, pp. 1843–1865, Oct. 2022.
- [34] C.-C. Liu, M.-G. Liao, C.-H. Chang, and H.-M. Lin, "An analysis of children' interaction with an AI chatbot and its impact on their interest in reading," *Computers & Education*, vol. 189, p. 104576, Nov. 2022.
- [35] J. Jeon, "Exploring AI chatbot affordances in the EFL classroom: young learners' experiences and perspectives," *Computer Assisted Language Learning*, vol. 37, no. 1-2, pp. 1–26, Jan. 2024.
- [36] R. Meyer von Wolff, J. Nörtemann, S. Hobert, and M. Schumann, "Chatbots for the Information Acquisition at Universities – A Student's View on the Application Area," in *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, Nov. 2019, pp. 231–244.
- [37] Y. Sumikawa, M. Fujiyoshi, H. Hatakeyama, and M. Nagai, "Supporting creation of faq dataset for e-learning chatbot," in *Intelligent Decision Technologies 2019: Proceedings of the 11th KES International Conference on Intelligent Decision Technologies (KES-IDT 2019), Volume 1*. Springer, 2019, pp. 3–13.
- [38] S. I. Ch'ng, L. S. Yeong, and X.-Y. Ang, "Preliminary Findings of using Chat-bots as a Course FAQ Tool," in *2019 IEEE Conference on e-Learning, e-Management & e-Services (IC3e)*, Nov. 2019, pp. 1–5.
- [39] T. K. Chiu, B. L. Moorhouse, C. S. Chai, and M. Ismailov, "Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot," *Interactive Learning Environments*, vol. 0, no. 0, pp. 1–17, 2023.
- [40] J. Brooke, "SUS-A quick and dirty usability scale," in *Usability evaluation in industry*, 1996, vol. 189, pp. 4–7, publisher: London, England.
- [41] B. Laugwitz, T. Held, and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire," in *HCI and Usability for Education and Work*, A. Holzinger, Ed. Berlin, Heidelberg: Springer, 2008, pp. 63–76.

- [42] S. Borsci, A. Malizia, M. Schmettow, F. van der Velde, G. Tariverdiyeva, D. Balaji, and A. Chamberlain, "The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents," *Personal and Ubiquitous Computing*, vol. 26, no. 1, pp. 95–119, Feb. 2022.
- [43] F. Holderried, C. Stegemann-Philipps, L. Herschbach, J.-A. Moldt, A. Nevins, J. Griewatz, M. Holderried, A. Herrmann-Werner, T. Festl-Wietek, and M. Mahling, "A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study," *JMIR Medical Education*, vol. 10, no. 1, p. e53961, Jan. 2024.
- [44] K. Boyd, C. Potts, R. Bond, M. Mulvenna, T. Broderick, C. Burns, A. Bickerdike, M. Mctear, C. Kostenius, A. Vakaloudis, I. Dhanapala, E. Ennis, and F. Booth, "Usability testing and trust analysis of a mental health and wellbeing chatbot," in *Proceedings of the 33rd European Conference on Cognitive Ergonomics*, ser. ECCE '22. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 1–8.
- [45] M. Sharma, S. Yadav, A. Kaushik, and S. Sharma, "Examining Usability on Atreya Bot: A Chatbot Designed for Chemical Scientists," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, Dec. 2021, pp. 729–733.
- [46] College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, A. Alogayli, and H. Abdelhafez, "Intelligent Chatbot for Admission in Higher Education," *International Journal of Information and Education Technology*, vol. 13, no. 9, pp. 1348–1357, 2023.
- [47] S. Holmes, R. Bond, A. Moorhead, J. Zheng, V. Coates, and M. McTear, "Towards Validating a Chatbot Usability Scale," in *Design, User Experience, and Usability*, A. Marcus, E. Rosenzweig, and M. M. Soares, Eds. Cham: Springer Nature Switzerland, 2023, pp. 321–339.
- [48] J. Salminen, S.-g. Jung, J. Medina, K. Aldous, J. Azem, W. Akhtar, and B. J. Jansen, "Using CIPHERbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot," in *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. Atlanta GA USA: ACM, Jul. 2024, pp. 279–283. [Online]. Available: <https://dl.acm.org/doi/10.1145/3657604.3664690>
- [49] J. Salminen, S.-G. Jung, J. Medina, K. Aldous, J. Azem, W. Akhtar, E. Häyhänen, and B. J. Jansen, "Communication Design for an Educational AI Chatbot: Analyzing CIPHERbot's Communication Style and Challenges," in *Proceedings of the 27th International Academic Mindtrek Conference*. Tampere Finland: ACM, Oct. 2024, pp. 176–187. [Online]. Available: <https://dl.acm.org/doi/10.1145/3681716.3681727>
- [50] J. Ha, H. Jeon, D. Han, J. Seo, and C. Oh, "Clochat: Understanding how people customize, interact, and experience personas in large language models," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024.
- [51] L. Lin, "A quarter of U.S. teachers say AI tools do more harm than good in K-12 education," May 2024. [Online]. Available: <https://www.pewresearch.org/>
- [52] L. Nielsen, J. Salminen, S.-G. Jung, and B. J. Jansen, "Think-Aloud Surveys - A Method for Eliciting Enhanced Insights During User Studies," *Human-Computer Interaction - INTERACT 2021*, vol. 12936, pp. 504–508, Aug. 2021, publisher: Springer.
- [53] J. R. McColl-Kennedy, M. Zaki, K. N. Lemon, F. Urmetzer, and A. Neely, "Gaining Customer Experience Insights That Matter," *Journal of Service Research*, vol. 22, no. 1, pp. 8–26, Feb. 2019.
- [54] V. Braun and V. Clarke, *Thematic analysis*. American Psychological Association, 2012. [Online]. Available: <https://psycnet.apa.org/record/2011-23864-004>
- [55] J. Deeks, J. Dinnes, R. D'Amico, A. Sowden, C. Sakarovich, F. Song, M. Petticrew, D. Altman, I. S. T. C. Group, S. T. C. GROUP *et al.*, "Evaluating non-randomised intervention studies," *Health technology assessment*, vol. 7, no. 27, pp. 1–173, 2003.
- [56] C. Peterson, B. Jesso, and A. McCabe, "Encouraging narratives in preschoolers: An intervention study," *Journal of child language*, vol. 26, no. 1, pp. 49–67, 1999.
- [57] K. Boyd, C. Potts, R. Bond, M. Mulvenna, T. Broderick, C. Burns, A. Bickerdike, M. Mctear, C. Kostenius, A. Vakaloudis *et al.*, "Usability testing and trust analysis of a mental health and wellbeing chatbot," in *Proceedings of the 33rd European Conference on Cognitive Ergonomics*, 2022, pp. 1–8.
- [58] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 56, Oct. 2023.
- [59] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75 264–75 278, 2020, conference Name: IEEE Access.
- [60] G.-J. Hwang and C.-Y. Chang, "A review of opportunities and challenges of chatbots in education," *Interactive Learning Environments*, vol. 31, no. 7, pp. 4099–4112, Oct. 2023.
- [61] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea preprints*, vol. 1, no. 3, pp. 1–26, 2023.
- [62] Y. Chen, S. Jensen, L. J. Albert, S. Gupta, and T. Lee, "Artificial intelligence (ai) student assistants in the classroom: Designing chatbots to support student success," *Information Systems Frontiers*, vol. 25, no. 1, pp. 161–182, 2023.
- [63] D. H. Chang, M. P.-C. Lin, S. Hajian, and Q. Q. Wang, "Educational design principles of using ai chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization," *Sustainability*, vol. 15, no. 17, p. 12921, 2023.
- [64] N. F. Davar, M. A. A. Dewan, and X. Zhang, "Ai chatbots in education: challenges and opportunities," *Information*, vol. 16, no. 3, p. 235, 2025.

APPENDIX

THE CHATBOT SUITABILITY QUESTIONNAIRE

Instructions: The Chatbot Suitability Questionnaire (CSQ), an assessment survey consisting of ten questions to measure factors related to the participants' beliefs about an AI chatbot's suitability for deployment in a MS classroom. The CSQ is based on SUS [40], and all questions are measured using multiple Likert-style items along three constructs of Interface and Design (3 items) of CIPHERbot, Language and Communication (3 items) of the chatbot's dialogue, and Interest and Beneficial (4 items) of and for MS students. The CSQ average score is calculated (similar to SUS) by subtracting one from score of each question, summing these new scores, and multiplying by 2.5, with results in a number with one hundred as the maximum. For deployment, we suggest changing the phrase "the chatbot" to the actual name of the system being evaluated. Also, the term "MS" can be changed to any appropriate grade or grade level. **The Chatbot Suitability Questionnaire (CSQ) items are:**

- **Interface and Design**

- My K12 students would likely find the chatbot's interface easy to navigate and understand.
- My K12 students would likely find the chatbot's colors, fonts, and layout appropriate and engaging.
- My K12 students would likely find the chatbot interesting.

- **Language and Communication**

- My K12 students would likely find the chatbot language appropriate and understandable.
- The chatbot responds quickly enough for my K12 students.
- My K12 students would likely be satisfied with their interactions with the chatbot.

- **Interest and Beneficial**

- The chatbot provides educational benefits or supports learning objectives for my K12 students.
- The chatbot's interactions would be meaningful and helpful for my K12 students.