applied system innovation

MDPI

Article

Design and Multi-Stakeholder Evaluation of the Perceived Pedagogical Value and Usability of an AI-Supported Educational Chatbot for Primary Education

Androniki Koutsikou [1,*], Nikos Antonopoulos [2] and Stamatis Papadakis [3]

[1] Department of Digital Media and Communication, Ionian University, 28100 Argostoli, Greece
[2] Department of Tourism, Ionian University, 49132 Corfu, Greece
[3] Department of Preschool Education, University of Crete, 74100 Rethymno, Greece
* Correspondence: akoutsikou@ionio.gr

Abstract

Artificial intelligence is included in several parts of education, and it serves as an important tool for it. In this study, we present the initial design and pre-deployment formative evaluation of an AI-powered educational chatbot, called TouriBot, from the perspective of different stakeholders. TouriBot is part of a gamified learning environment created specifically for students in primary education (ages 9–12). The application is intended to support potential engagement with critical literacy-related tasks in the setting of misinformation. The evaluation involved three stakeholder groups, namely, in-service primary school Information and Communication Technology (ICT) teachers (N = 112), ICT educational advisors (N = 13) serving as institutional pedagogical experts and user experience (UX) experts (N = 8). No student data were collected in this evaluation phase. Structured questionnaires assessed perceived usability and pedagogical value as well as behavioral intention to use the tool. The results show that stakeholders evaluated the chatbot positively regarding usability, pedagogical alignment and stated intention to use. ICT teachers indicated high perceived usefulness and ease of use and strong intention to use the system in their teaching environments. The user interface perception-based assessments from the UX experts were mainly positive, and the ICT educational advisors highlighted the chatbot's alignment with the intended pedagogical goals. However, TouriBot had limitations and drawbacks identified during this initial phase of the study. These stakeholder evaluations provide perception-based feedback that can inform further system refinement and guide student-centered evaluation in classroom settings.

Keywords: AI-supported educational chatbot; gamification; early-stage evaluation; intercultural communication; perception-based study; primary education; human–computer interaction; misinformation

Check for updates

Academic Editors: Teen-Hang Meen and Sheng-Tzong Cheng

Received: 4 August 2026
Revised: 6 September 2026
Accepted: 22 September 2026
Published: 29 September 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Recent progress in generative AI (GenAI) has reshaped how interactive learning tools are built and used. AI-powered conversational agents are becoming increasingly popular to facilitate interaction in learning environments [1]. Generative AI tools are conceptual generators and design assistants. Thus, educational activities are required to develop students' critical thinking and digital skills. As a result, students can identify the pitfalls and inaccuracies of artificial intelligence and respond critically rather than passively [2].

Chatbots are frequently used in learning environments to support structured interaction and self-directed learning. They function as conversational agents in diverse educational environments, providing opportunities for practice, feedback and increased student engagement [3]. They were also developed at the primary school level based on static databases of questions and answers and pre-defined knowledge bases. The emergence of GenAI has transformed these systems into active agents that can provide contextual feedback and more sophisticated forms of pedagogical scaffolding [4]. Large Language Models (LLMs) have made chatbots more advanced, capable of providing adaptive, context-sensitive responses and adapting to each learner's needs in real time [5]. Recent research indicates that the use of AI-supported chatbots in primary education depends on factors such as pedagogical guidance, response quality and interaction design. This means that user satisfaction and system usability should be evaluated systematically [6]. Nowadays, AI-based educational chatbots rely on machine learning to generate structured and guided interactions that meet teaching objectives, mainly in primary education [7]. These chatbots are used in class activities to help boost student engagement. They provide instant feedback, which improves communication and interaction in a learning environment [8]. Gamification in education can also increase motivation when aligned with lesson objectives, but it is most effective when the technology is well integrated and the instructional design is carefully planned [9].

Artificial intelligence-based conversational agents are progressively shaping the field of education. Ait Baha et al. [8] evaluated a chatbot as a virtual instructor for secondary students and concluded that interaction with the chatbot helps maintain student engagement. Previous studies of chatbots in different educational environments focused on user experience, interface design and thorough evaluation of AI-based systems [8,10]. Elmaadaway et al. studied a voice chatbot (Alexa) in primary schools and discussed the need for interactive support and user involvement in chatbot-based learning [11].

Some promising advancements have been made; however, the efficacy and value of AI-supported educational chatbots rely on aspects such as system architecture, adaptability and user experience. Previous work has suggested that better algorithms, better interfaces and expert judgment are needed to make chatbots more effective and usable [12]. A diversity of professional opinions may help us understand how these systems work and how they are acceptable for teaching, especially as more schools begin to adopt them.

Recent research explored general attitudes towards AI, such as the effect of knowledge about AI technologies on users' opinions. Most of the literature focuses on general attitudes instead of attitudes towards specific AI-based educational technology or the attitudes of professionals with technical or user-experience skills [13]. For example, some studies use methods such as the System Usability Scale (SUS) to assess usability, as in the Learning by Building Chatbot environment. They generally help teachers design chatbots, though. They do not analyze the satisfaction of users of a whole chatbot-based learning environment with different stakeholder groups [14].

Chatbots are used in different educational contexts, and the results vary depending on their design and how they are embedded in classroom practice. Research so far has been primarily on performance data in experimental settings. Yet their use in certain contexts and for particular design purposes remains poorly understood, notably in the case of gamified environments for the development of critical thinking and online literacy skills. Even fewer take a multi-perspective strategy that combines the thoughts of both UX experts and educators.

Another major issue is the absence of early assessment of AI-supported educational chatbots before they are deployed in schools. At this moment, few frameworks combine education, usability and system-level insights. Much of the present research does not

synthesize the perspectives of educators, UX experts and policy stakeholders in one analysis. This lack makes it difficult to comprehend how to create, refine and evaluate these systems prior to their usage in practical educational environments [3,12,14].

The present study addresses this gap through a multi-stakeholder evaluation of an AI-supported educational chatbot integrated into a gamified learning environment. This focuses on adult stakeholders' perceptions of its pedagogical value, usability and potential classroom applicability prior to subsequent evaluation with the intended users. This approach may help clarify how such systems can be designed and refined through early-stage stakeholder feedback before subsequent student-centered and classroom evaluation.

This need for early-stage, multi-stakeholder evaluation becomes even more critical when AI-supported tools are intended for use by young students who face increasing challenges when using digital content.

Young students are frequently exposed to vast amounts of visual and digital content. This can be particularly challenging given that their cognitive abilities and media literacy skills are still developing. Research shows that many students seem to have difficulty distinguishing reliable, accurate news and information from misleading content in the digital world [15–17]. Therefore, we can understand the importance of media education and digital literacy. Students should be able to evaluate and interpret digital content [18].

In this sense, the present work proposes the development of an educational application that integrates a chatbot within a gamified environment designed to support activities related to the identification and critical evaluation of misinformation. It will not provide an instant yes-or-no answer to whether information is misinformation or not but will guide students through questions and hints intended to encourage information cross-checking. The chatbot gives hints and feedback to students to help them engage in research, interact and reflect [19,20].

Thus, the interaction is not limited to basic dialog. It applies the Socratic (Maieutic) Method to turn the learner into an active citizen instead of a passive user of information. The chatbot is not an expert in providing correct answers but a trained conversation assistant who directs students with thoughtful questions and helps them to explore and evaluate information [21].

This study relies on the methodological framework of Human–Computer Interaction (HCI) Usability Evaluation and the Technology Acceptance Model (TAM) to perform a broad and multi-perspective evaluation. HCI research is concerned with user-centered design and transparent interaction in digital systems, principles that are particularly relevant for the evaluation of interactive educational technologies such as AI-supported educational chatbots. When designing and evaluating chatbot interactions, the focus should not only be on functionality but also on conversational clarity, perceived interaction style, transparency and the ability to support user control and error recovery [22,23].

Along with that, TAM constructs are used to capture teachers' attitudes towards the proposed application [24]. The evaluation tool is oriented by two established frameworks in educational game design: the Gaming Educational Balanced (GEB) Model, which helps ensure gamification alignment, and the E-Guess model, which focuses on pedagogical alignment [25,26].

Finally, this research follows a multi-stakeholder approach incorporating complementary but analytically distinct perspectives from ICT teachers, ICT educational advisors and UX experts. In primary school classes, ICT teachers usually teach 1 h per week in a computer lab nationwide. ICT educational advisors are a valuable resource for pedagogical advice and assistance with the implementation and evaluation of ICT-related practices in education. These roles empower both groups to evaluate the perceived pedagogical consistency and usability of educational technological innovations.

This study compares the views of three professional groups: ICT teachers, ICT educational advisors and UX experts. Each group considered a different part of the proposed system. The outcomes may contribute to a deeper understanding of how AI-supported educational chatbots are perceived by different stakeholder groups before being implemented in classrooms. These data can also inform the design cycle and early improvements to conversational agents and gamified learning environments. Based on a synthesis of theoretical perspectives relevant to technology-enhanced learning and user-centered design, the following research questions have been developed:

- **RQ1**: How pedagogically valuable and usable is the AI-supported educational chatbot application, according to ICT educational advisors?
- **RQ2**: How do UX experts perceive the usability and quality of interaction of the chatbot, based on Nielsen's usability heuristics?
- **RQ3**: How do ICT teachers evaluate the pedagogical value and usability of the chatbot?
- **RQ4**: Which variables are associated with ICT teachers' perceived intention to use the application?
- **RQ5**: What differences exist between ICT teachers and ICT educational advisors regarding their evaluation of the system?

2. Materials and Methods

2.1. Participants

The comprehensive evaluation of the proposed AI-supported educational chatbot was performed with three different groups of stakeholders: ICT educational advisors, UX experts and ICT teachers.

The sample consisted of 13 ICT educational advisors who specialized in computer science education, with an average of 21 years of teaching experience. Most ICT advisors were in the 50–59 age range, and a majority had advanced academic degrees, including a PhD. Most participants had prior experience with educational evaluation methods, which further emphasizes the professional relevance of this group to the study.

This study involved 8 UX experts from user experience research, university settings, and interface design. The group spanned a range of ages and included male and female respondents, although one participant did not disclose gender. Most participants were conducting research, followed by academics and user experience and digital design experts. Their professional experience varied and included extensive involvement in developing and evaluating digital systems.

Finally, this study enrolled 119 ICT teachers proficient in computer science for children aged 9 to 12 years. For each stakeholder group, there was a self-report consistency check at the end of the questionnaire to identify potentially unreliable responses. For the ICT educational advisors and UX experts, no cases were excluded, and all responses were retained for analysis. However, for the ICT teachers, four participants were excluded for self-reported insincerity on the honesty-check item, and three additional cases were excluded for incomplete data. The final dataset of ICT teachers comprised 112 valid responses (N = 112).

Primary schools received an email invitation outlining the study's aims, following ethical approval, which was then forwarded to ICT teachers. Participation was voluntary, and the sample was a convenience sample, so the findings are not statistically generalizable.

ICT educational advisors tend to be a small but highly qualified professional group within the national education system. The official institutional framework [27] specifies their role as supporting the implementation of educational policy, providing pedagogical and didactic guidance to teachers, organizing professional development activities and contributing to the assessment and enhancement of instructional methods. ICT educational

advisors act as a bridge between educational policy and practice at the school level, helping schools adopt digital tools while encouraging innovative teaching approaches. At the time of this study, 29 ICT educational advisors served across the country. The invitation was sent to everyone surveyed, and 13 advisors responded, which is about 45% of all advisors and thus a significant percentage of the desired population. Their institutional responsibilities and system perspective make their feedback especially important for evaluating school-related best practices. To our knowledge, this study is one of the first to engage such a large percentage of ICT educational advisors at a national level in the evaluation of an educational technology application. The outcomes, however, should be viewed as perspectives informed by experts, not as statistically generalizable results.

An invitation email was also sent to UX experts. Participants were selected based on their background in UX/UI design, human–computer interaction or associated research areas. During the data collection phase, we sent a reminder email to all invited participants to encourage participation. The sample size for UX experts (N = 8) is small. The UX expert component of the present study consisted of a perception-based assessment structured around Nielsen's usability heuristics rather than a formal task-based heuristic inspection. Given the small sample, the findings are interpreted descriptively as the assessments of the participating experts rather than as statistically generalizable estimates.

The present study describes the first design and multi-stakeholder evaluation phase of the AI-supported educational chatbot. In this stage, no data were collected from students. The evaluation focused on adult stakeholder perspectives on the system's pedagogical design and usability and involved ICT teachers, ICT educational advisors and UX experts. Direct student-centered evaluation with the intended age group (9–12 years) represents a separate empirical direction that may complement the present findings by examining students' interactions with TouriBot, its usability in classroom settings and age-appropriate learning outcomes.

2.2. Research Design and Procedure

The tool used in this study was built with Genially [28], a cloud-based platform that supports the creation of interactive, gamified digital content. The intervention used an application called "Fake News Travels Fast" designed around practical problem-solving and interactive scenarios that help students stay involved. Furthermore, a specially designed AI-supported educational chatbot was developed using the Chatbase (https://www.chatbase.co/, accessed on 21 September 2026) and integrated into a simulated educational activity environment. The chatbot is called TouriBot. The application was designed as a gamified learning scenario aligned with critical literacy objectives among students aged 9–12. The "Fake News Travels Fast" application, as shown in Figure 1, was developed and assessed in its initial phase, incorporating core pedagogical scenarios and storytelling elements that have previously been evaluated in classroom settings in relation to students' engagement and critical thinking skills [29]. Our previous study examined the original gamified application in a primary school setting and reported on student-centered usability, experience and pre- and post-intervention outcomes for the original application, not the chatbot extension evaluated in the present study. Building on that foundation, the current study focuses on the chatbot extension and evaluates it through a multi-stakeholder lens involving ICT teachers, ICT educational advisors and UX experts.

The application was enhanced with TouriBot within the game narrative, as shown in Figure 2, which provides contextual prompts and feedback. In this scenario, students can interact with the application to explore situations involving the identification of misleading information and the evaluation of digital content. TouriBot aims to help users with the activities, so it provides guidelines, hints and feedback. The application has a variety

of activities, questions and interactive elements designed to encourage critical thinking, discussion and decision-making. We embedded game-like features such as challenges and progress bars to foster user engagement. In addition, TouriBot's content and responses were adjusted to the context of digital literacy and misinformation [30].



Figure 1. Example of an activity in which students are asked to evaluate whether an image is real or AI-generated. The original interface is in Greek. The task asks students to decide whether the photograph is authentic or AI-generated.

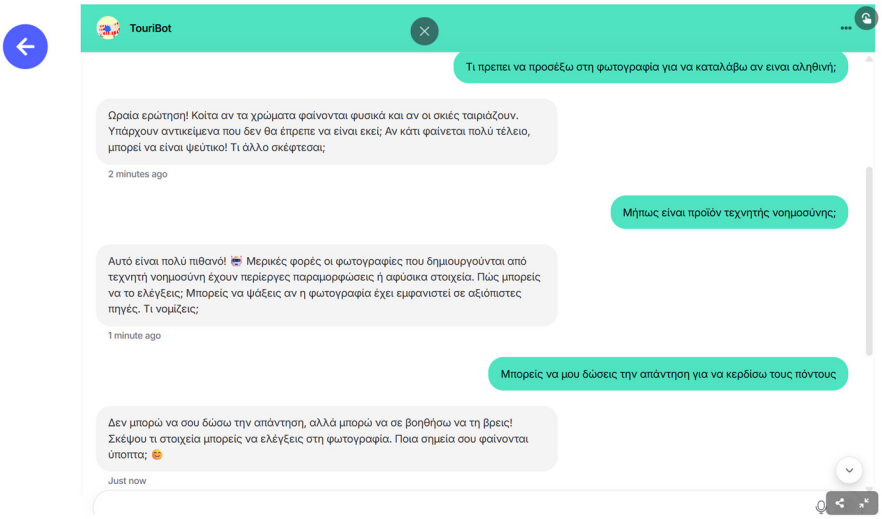


Figure 2. Example of the TouriBot conversational interface in one of the predefined activities. The original interface is in Greek. The conversation illustrates the chatbot's question-based guidance in an image-authenticity task, prompting the user to examine visual details and reflect on whether the image may be AI-generated rather than providing the answer directly. This conversation was created for illustrative purposes and does not contain any participant data. The application followed a strictly linear structure with no alternative pathways. All participants from the three stakeholder groups progressed through the same predefined activities in the same sequence, and completing the full sequence required approximately 15–20 min. Although exposure to the game activities was therefore consistent across participants, interaction with TouriBot was not standardized. Participants were not provided with predefined prompts or required to ask specific questions. Instead, TouriBot remained available throughout the activities, and participants could choose whether and when to use the chatbot during each activity and what questions to ask; however, they could not alter the predefined sequence of activities. Therefore, all participants completed the same activities in the game sequence, but the interaction with the chatbot depended on the individual choices. The procedure therefore standardized participants' exposure to the application but not the frequency, timing or content of their interactions with TouriBot.

Furthermore, they were asked to use a desktop or laptop computer because the application was designed for computer workstations in a school computer lab and not for mobile devices. All participants used the same fully integrated version of the application, including TouriBot, to ensure consistency in the system version available for evaluation. Participants accessed the application remotely via a shared link and completed the exploration individually, at a time and location of their own choosing. After this exploration, participants completed the evaluation questionnaire via Google Forms. This procedure was designed to give participants personal experience with the system before reporting professional perceptions.

2.3. System Architecture and TouriBot Design

TouriBot's conversation engine is supported by the application programming interface (API) of the GPT-4o Mini large language model (LLM). GPT-4o Mini was selected as the underlying language model for the conversational component of TouriBot. The system uses a retrieval-augmented generation (RAG) architecture that bases its responses on a tailored knowledge base, rather than relying exclusively on general pre-trained data. The system uses controlled prompts, retrieved passages and predefined interaction rules. These features were intended to keep the dialog focused on the educational task. Figure 3 presents the system architecture.

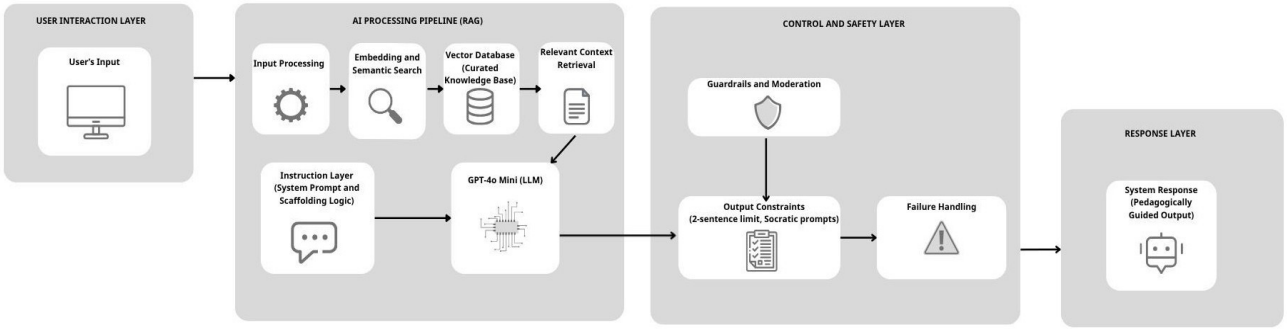


Figure 3. The figure shows TouriBot's architecture. Specifically, it illustrates the retrieval-augmented generation (RAG) workflow, the constraints on the system's prompts and the safety mechanisms designed to support focused and pedagogically structured interaction.

2.3.1. Knowledge Base and Retrieval Pipeline

The functionality of TouriBot was based on a carefully selected collection of domain-specific documents (N = 9) concerning institutional frameworks, custom-made pedagogical materials and game-specific resources in alignment with the narrative and logic of the gamified learning environment. The institutional part contains official European Commission documents concerning online misinformation. The pedagogical materials contain educational content on misinformation, fake news and image manipulation.

TouriBot uses semantic retrieval to identify passages in the knowledge base that are relevant to a user query. The language model receives these passages as context when generating a response. This process may improve contextual relevance. It does not establish that every response is accurate or suitable for children aged 9–12.

2.3.2. Prompt Engineering and Socratic Scaffolding

TouriBot's intended conversational behavior was specified through a structured system prompt that defined its role as a learning assistant rather than a direct problem solver. It is expected to have a friendly tone that is pedagogically tailored to the students' age. It is designed to work as a learning assistant. The prompt was designed around a scaffolding approach. The chatbot was instructed to avoid providing direct solutions and instead to

use questions and step-by-step hints. The prompt also included instructions to provide brief, constructive guidance when users indicated difficulty or frustration. Furthermore, the output constraints were designed to keep the responses short, accessible and appropriate for students aged 9–12. Finally, all generated content stayed focused on the learning goal of evaluating online information, so that the chatbot always supports its educational purpose (Table 1).

Table 1. Key system prompt rules and intended response constraints.

Category	Description
Role Definition	The chatbot acts as a friendly educational assistant for students aged 9–12
Pedagogical Strategy	Socratic scaffolding through reflective questioning
Output Constraints	Maximum of two short sentences per response
Answer Restriction	The chatbot is instructed not to provide direct answers
Domain Focus	Restricted to misinformation, fake news and critical evaluation
Off-topic Handling	Redirects unrelated queries to the learning objective
Failure Handling	Refers students to the teacher when no relevant answer is available

2.3.3. Guardrails and Moderation

We added multiple levels of interaction restrictions and filtering to reduce off-topic responses and support alignment with the intended educational context. First, a strict topic restriction mechanism limits TouriBot's responses to the topic being discussed. If students raised off-topic subjects or general questions, the system redirected them back to the main learning objective. These constraints were intended to maintain topic alignment and encourage responses appropriate to the intended age group. A mechanism that handles off-topic questions also has predefined alternative answers, so that users are directed back to the topic at hand.

2.3.4. Failure Handling and Response Control

The system uses methods to handle uncertainty and reduce the chance of providing answers that are not supported by reliable information. TouriBot does not provide hypothetical answers if the retrieval process does not find relevant information. Instead, it gives general advice or suggests reaching out to the ICT teacher. The Control and Safety Layer was designed to reduce the likelihood of unsupported or incorrect outputs, including hallucination-like responses. The system was configured to use material retrieved from the selected knowledge base. When retrieval did not return relevant material, the chatbot was instructed to give general guidance or refer the student to the teacher. These controls may reduce unsupported responses, but they cannot prevent them in every case. The design includes topic restrictions, short-response rules and predefined handling of off-topic questions. These technical mechanisms were implemented as design safeguards. However, their effectiveness was not independently evaluated through a systematic benchmark of retrieval accuracy, hallucination frequency, response robustness or age appropriateness in the present study. They should therefore be interpreted as system-design features rather than empirically validated performance characteristics.

2.3.5. Example Interaction

Table 2 presents a few examples of TouriBot's interaction based on the implemented Socratic scaffolding approach.

Table 2. Examples of chatbot interaction based on the implemented Socratic scaffolding approach.

Student Input	TouriBot Response
"Is this picture real?"	"What do you notice about the source of the image? Can you find who created it?"
"Give me the answer."	"I can't give you the answer, but I can help you think about it. What clues can you use to decide?"
"This looks fake."	"What makes you think it might be fake? Can you find something that supports your idea?"

TouriBot combines a selected knowledge base, a structured system prompt and predefined interaction rules. The present study examined adult stakeholders' views of this design. Direct testing with students remains necessary. The next stage should test the chatbot's responses, usability and educational use with students and teachers in classroom settings.

2.4. Evaluation Instruments and Measures

We used structured questionnaires to examine stakeholders' views of the application. The questionnaires drew on established evaluation frameworks, with items adapted to the TouriBot context. Thus, for assessing educational games, pedagogical suitability and acceptance of the technological system, we used, in the first section of the questionnaire included questions based on the Gaming Educational Balanced (GEB) model, which examines key dimensions of educational game design, such as attractiveness, perceived engagement, appropriateness of difficulty levels, functionality of game mechanics and interface quality [25]. The second section of the questionnaire was adapted from the dimensions proposed in the E-Guess model to check game usability. This section evaluates pedagogical consistency, alignment with the cognitive domain of critical literacy and the perceived potential for students to apply and self-assess their knowledge [26]. Finally, the section used the measures of the Technology Acceptance Model (TAM) of Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) to identify the teachers' perceptions of the practicality and ease of use of the application. The final section of the questionnaire measured participants' stated intention to use or recommend the application. These parts were rated using a five-point Likert scale. The questionnaires produced ratings of user experience, pedagogical value and stated intention to use the application [24]. The questionnaire was distributed to the ICT teachers and the ICT educational advisors to elicit their evaluation of the system. Questionnaire items were adapted from the theoretical frameworks above (GEB, E-Guess, TAM) and tailored to the context of AI-supported educational chatbots. Each construct was implemented using multiple items measuring engagement, pedagogical relevance, usability, perceived usefulness and behavioral intention. The complete list of questionnaire items is provided in Appendix A (Table A1).

The evaluation was conducted with adult stakeholders rather than the intended student users, so items in the questionnaire about engagement, difficulty, learning relevance or critical literacy were interpreted as the participants' professional judgments about the expected characteristics and appropriateness of the design. These ratings did not function as direct measures of student experiences, usability, engagement, learning or critical literacy skills. Thus, the questionnaire data reflect stakeholder perceptions that are used in a pre-deployment formative evaluation and should not be interpreted as student-level evidence.

To further evaluate the interface, UX experts completed a Likert-scale questionnaire structured around Nielsen's ten usability heuristics. Rather than performing a task-based inspection to discover and record specific usability issues, experts expressed their level of agreement with statements based on each heuristic, providing a perception-based expert assessment of the interface in relation to established usability principles [31]. This approach

offers a way to gather structured expert impressions of the interface's strengths and potential weaknesses, without requiring extensive user testing. It addresses key design areas, including the perceived visibility of system status, consistency with common interface expectations and the perceived flexibility and freedom of interaction. The assessment also covers expert perceptions of error prevention, text legibility, interaction flexibility and the clarity and simplicity of the visual design. Together, these design criteria provide a structured, perception-based overview of the interface's perceived quality [32,33]. The complete assessment tool used in this study is provided to readers in Appendix A (Table A2).

In addition to the perception-based usability assessment informed by Nielsen's heuristics, the UX evaluation also considered dimensions related to conversational interaction quality, TouriBot's persona, transparency and dialog repair [12,31,34,35]. A questionnaire was developed to collect UX experts' assessments of TouriBot's interaction and conversation design. This instrument is based on the principles of conversation design and existing frameworks for chatbot evaluation that identify conversational quality, social presence, transparency and dialog repair as key dimensions of successful human–AI interaction [34]. The full evaluation instrument for TouriBot's interactions is presented in Appendix A (Table A3).

In general, TouriBot was assessed using a multi-stakeholder approach involving ICT teachers, ICT educational advisors and UX experts. Each group provided a different perspective on the system, including pedagogical, strategic and usability perspectives. The structure of the multi-stakeholder evaluation design is summarized in Figure 4.

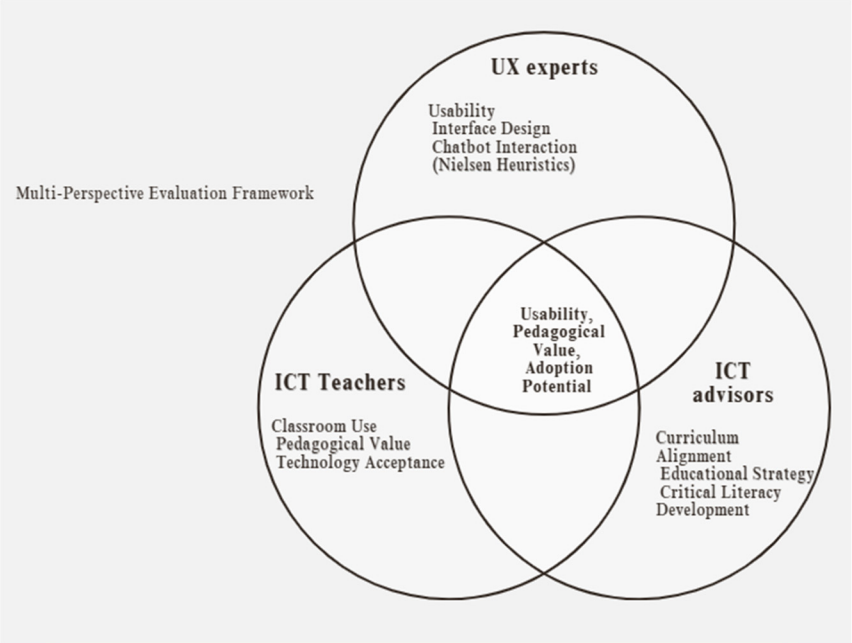


Figure 4. The figure shows the multi-stakeholder evaluation approach with the roles of ICT teachers, ICT educational advisors and UX experts in assessing the application's pedagogical value, usability and stated intention to use.

We analyzed the findings from each stakeholder group based on their specific evaluation focus. We then discussed them together in the Discussion section. This multi-stakeholder approach gave different but complementary perspectives on TouriBot. These perspectives were used to examine different parts of the system, not to fully validate the results.

3. Results

Because this study is exploratory and perception-based, the results are reported as descriptive and associative findings. These findings reflect stakeholder perceptions and do not constitute evidence of actual classroom adoption, instructional effectiveness, or student learning outcomes.

3.1. Evaluation of the ICT Educational Advisors on the Application (RQ1)

To answer RQ1, we calculated the means, standard deviations and 95% confidence intervals to examine how ICT educational advisors rated TouriBot themselves. The results indicate generally positive self-reported perceptions across all dimensions.

Since the number of ICT educational advisors was rather small (N = 13), we present 95% confidence intervals with the means to give a sense of the precision of these estimates. As Table 3 and Figure 5 suggest, the scores on the five-point scale ranged from 4.19 to 4.38, which indicated high self-ratings across all variables across the board. Behavioral intention was rated highest (M = 4.38, SD = 0.49). This may indicate that participants had a high intention level to use the application. Furthermore, ease of use was rated positively (M = 4.35, SD = 0.67), which shows that participants found the system easy and accessible to use. The pedagogical content was also well rated (M = 4.32, SD = 0.46), which shows that participants thought the educational content was suitable and appropriate. The gamification elements were also rated positively (M = 4.29, SD = 0.44), indicating that the game-based elements were perceived as engaging and appropriate for the learning activities for which they were created. Perceived usefulness was rated the lowest (M = 4.19, SD = 0.60) among all constructs, although it indicates a generally positive perceived value of the application.

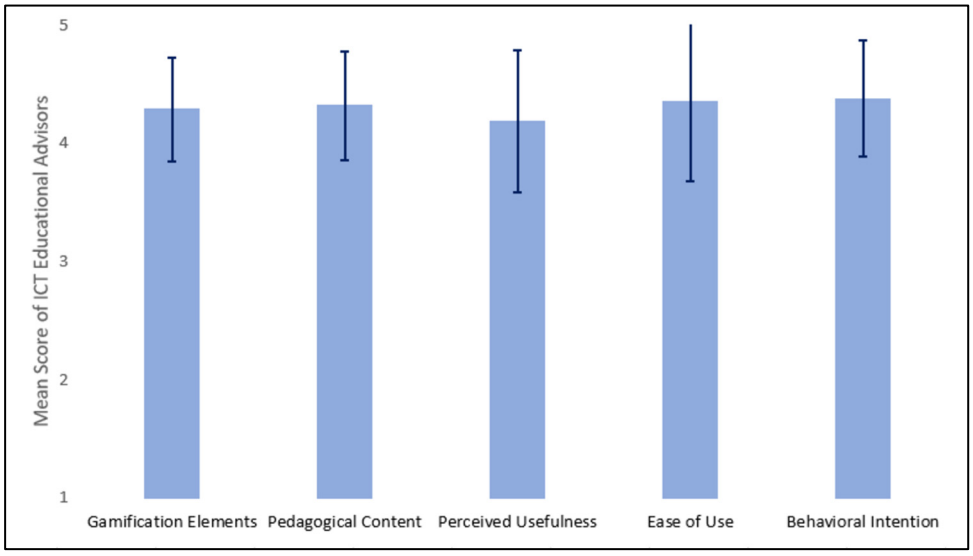


Figure 5. Mean scores of ICT educational advisors' evaluations across the five constructs. Error bars indicate standard deviations.

The ICT educational advisors had a positive opinion of the usability, pedagogical fit and intention to use TouriBot. However, these are perception-based evaluations from a limited expert sample, and the perceived expert sample should not be seen as evidence of actual system effectiveness or real-world adoption.

Table 3. Descriptive statistics and 95% confidence intervals for ICT educational advisors' evaluation (N = 13).

Measure	Mean	SD	95% CI (Lower)	95% CI (Upper)
Gamification Elements	4.29	0.44	4.02	4.56
Pedagogical Content	4.32	0.46	4.04	4.59
Perceived Usefulness	4.19	0.60	3.83	4.55
Ease of Use	4.35	0.67	3.94	4.75
Behavioral Intention	4.38	0.49	4.09	4.68

Note. Confidence intervals are based on the t-distribution due to the small sample size.

3.2. Perception-Based Usability Assessment by UX Experts (RQ2)

To answer RQ2, UX experts completed a Likert-scale questionnaire structured around Nielsen's ten usability heuristics. The perception-based opinions of the UX experts were expressed as averages and standard deviations. Interpretation levels were based on ranges of mean scores (1–5) to provide an informative classification of the replies: 1.00–2.99 (low), 3.00–3.99 (moderate), 4.00–4.49 (high), and 4.50–5.00 (very high). The categories are descriptive and are not intended to be viewed as objective performance benchmarks. These statistics are provided only as descriptive summaries of the heuristic assessments of the participating UX experts and should not be interpreted as population-level estimates given the small expert sample (N = 8).

The perception-based expert assessment (as shown in Table 4 and illustrated in Figure 6) showed generally positive UX expert ratings across most of the dimensions structured around Nielsen's usability heuristics. "Recognition rather than recall" (M = 4.88, SD = 0.35), "help and documentation" (M = 4.75, SD = 0.46) and "match between the system and the real world" (M = 4.63, SD = 0.52) received high scores. The ratings indicate generally positive expert perceptions of the interface in terms of recognition, guidance, and correspondence with familiar interaction patterns. Moreover, "user control and freedom" received M = 4.63 (SD = 1.06) and "consistency and standards" received M = 4.50 (SD = 0.76). The ratings also show positive expert ratings of user control and consistency with common interface norms. However, the scores were lower for "flexibility and economy of use" (M = 2.63, SD = 1.41) and "aesthetic and minimalist design" (M = 3.50, SD = 1.41).

Table 4. Results of the perception-based expert assessment structured around Nielsen's ten usability heuristics (UX expert ratings).

Heuristic	Mean	SD	Interpretation
Visibility of system status	4.13	0.99	High
Match between the system and the real world	4.63	0.52	Very High
User control and freedom	4.63	1.06	Very High
Consistency and standards	4.50	0.76	Very High
Error prevention	4.25	0.71	High
Recognition rather than recall	4.88	0.35	Very High
Flexibility and efficiency of use	2.63	1.41	Low
Aesthetic and minimalist design	3.50	1.41	Moderate
Error recognition and recovery	4.13	1.13	High
Help and documentation	4.75	0.46	Very High

Note. The interpretation categories are based on predefined Likert scale ranges (1–5) and should be considered indicative rather than absolute evaluations.

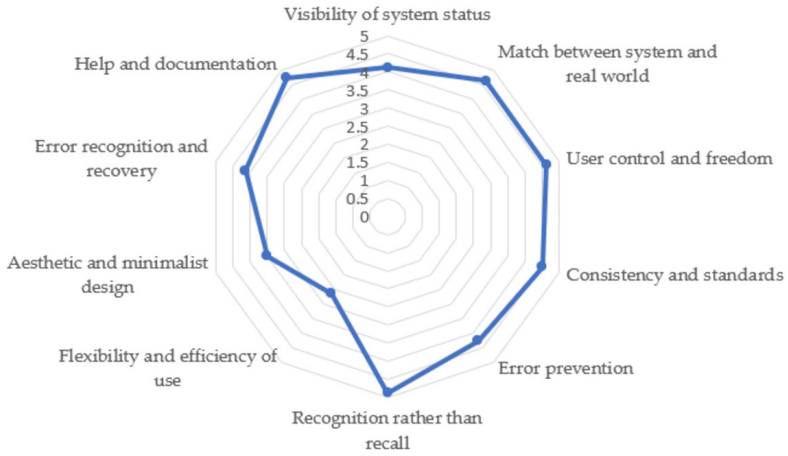


Figure 6. Mean UX expert ratings on questionnaire dimensions structured around Nielsen's ten usability heuristics.

UX experts gave the system comparatively low ratings for flexibility and efficiency of use (M = 2.63), suggesting that the system was relatively rigid and linear and may limit its ability to support diverse user tasks or interaction styles. The wide standard deviation (SD = 1.41) indicates a range of opinions among experts, with some indicating that the guided design was suitable for younger students, while others suggested that more experienced users may find the limited interaction to be restrictive. Likewise, the moderate rating for "aesthetic and minimalist design" (M = 3.50) indicates more mixed expert perceptions of this aspect of the interface. The relatively large standard deviation (SD = 1.41) indicates substantial variation in expert ratings of the visual layout's clarity and simplicity.

The UX expert ratings of TouriBot interactions were generally positive across the evaluated dimensions, as shown in Table 5. Given the relatively small number of UX experts (N = 8), we report 95% confidence intervals to determine the precision of the mean values.

Table 5. Descriptive statistics and 95% confidence intervals for UX expert ratings (N = 8).

Dimension	Mean	SD	95% CI (Lower)	95% CI (Upper)
Conversational clarity	4.00	0.97	3.19	4.81
Social presence and persona	3.75	1.08	2.85	4.65
Transparency and user control	3.88	0.88	3.14	4.61
Error recovery and dialog repair	4.63	0.38	4.31	4.94

Note. The confidence intervals are based on the t-distribution due to the small sample size.

The highest mean rating was obtained for error recovery and dialog repair (M = 4.63, SD = 0.38), which suggests that, in general, the participants perceived the chatbot as supportive when errors occurred or clarification was needed. The narrow confidence interval further suggests consistency in these opinions. Conversational clarity was also rated high (M = 4.00, SD = 0.97). The responses show that TouriBot was, in general, regarded as being clear and understandable. However, slightly lower scores were observed for transparency and user control (M = 3.88, SD = 0.88) and for social presence and persona (M = 3.75, SD = 1.08). These findings suggest that, while TouriBot interaction is perceived as functional and supportive, there is room to enhance its perceived personality and interaction style. Moreover, due to the small sample size (N = 8), we interpret these results as expert-informed assessments rather than statistically generalizable findings on usability.

3.3. ICT Teachers' Evaluation of the Application (RQ3)

Descriptive statistics (frequencies and percentages) were calculated to summarize the demographic characteristics of the participants. The demographic profile of the sample is presented in Table 6.

Table 6. Demographic characteristics of ICT teacher participants (N = 112).

Variable	Category	n	%
Gender	Male	41	36.6
	Female	71	63.4
Age	20–29	1	0.9
	30–39	16	14.3
	40–49	57	50.9
	50–59	34	30.4
	60+	4	3.6
Teaching Experience	0–5 years	16	14.3
	6–10 years	23	20.5
	11–15 years	12	10.7
	16–20 years	21	18.8
	21+ years	40	35.7
Education Level	Bachelor's Degree	38	33.9
	Master's Degree	69	61.6
	PhD	4	3.6
	Post-doc/Other [1]	1	0.9
Prior Evaluation Experience	Yes	45	40.2
	No	67	59.8

[1] Note: The "Other" category includes participants with postdoctoral research status.

To evaluate the internal consistency of the questionnaire scales, we conducted a reliability analysis on the full dataset of N = 112 participants (ICT teachers) using Cronbach's alpha (α) and McDonald's omega (ω) coefficients. The results indicated acceptable to high internal consistency across all dimensions, with Cronbach's α values ranging from 0.877 to 0.944 and McDonald's ω values ranging from 0.879 to 0.945. These internal consistency estimates are broadly consistent with the reliability coefficients reported in previous TAM research [36]. As shown in Table 7, all the questionnaire scales have strong internal consistency. These internal consistency estimates should not be interpreted as evidence of construct or discriminant validity.

Table 7. The reliability analysis of the questionnaire scales.

Evaluation Measure	Items	Cronbach's α	McDonald's ω
Gamification Elements	1–5	0.877	0.879
Pedagogical Content	6–12	0.910	0.913
Perceived Usefulness	13–16	0.944	0.945
Ease of Use	17–20	0.932	0.934
Behavioral Intention	21–23	0.912	0.917

Descriptive statistics (means and standard deviations) were calculated to analyze ICT teachers' perceptions of the application across the five aspects of the questionnaire in order to answer RQ3.

As shown in Table 8, all evaluation measures received high ratings. The mean of the ratings ranged from 4.22 to 4.44 on a five-point Likert scale, suggesting generally positive self-reported evaluations of the program across the constructs. The 95% confidence

intervals (CIs) were calculated to ensure the precision of these estimates. The confidence intervals estimate the precision of the sample means. They do not correct for self-selection, response bias or the use of a convenience sample.

Table 8. Descriptive statistics and 95% confidence intervals for ICT teachers (N = 112).

Evaluation Measure	Mean	SD	95% CI (Lower)	95% CI (Upper)
Gamification Elements	4.41	0.61	4.30	4.52
Pedagogical Content	4.36	0.64	4.25	4.48
Perceived Usefulness	4.22	0.79	4.07	4.37
Ease of Use	4.44	0.66	4.32	4.57
Behavioral Intention	4.29	0.79	4.14	4.44

Ease of use (M = 4.44, SD = 0.66) was evaluated as the highest, followed by gamification elements (M = 4.41, SD = 0.61) and pedagogical content (M = 4.36, SD = 0.64). Behavioral intention also received a high evaluation (M = 4.29, SD = 0.79), while perceived usefulness showed similarly positive responses (M = 4.22, SD = 0.79). The observed negative skewness across constructs indicates a tendency toward higher ratings. This pattern may result from response tendencies inherent in self-reported data, such as acquiescence bias or novelty effects associated with evaluating a newly introduced system. These factors should be considered when interpreting the generally high mean scores.

Given that the instrument was adapted for a context-specific application, the following analysis is strictly exploratory; its results should be interpreted as preliminary indications of the instrument's internal structure, rather than as evidence of construct validity. To explore the instrument's preliminary internal structure in this sample, an exploratory factor analysis (EFA) was conducted on the responses of ICT teachers (N = 112) using the minimum residual extraction method with oblimin rotation.

The suitability of the data for factor analysis was confirmed by the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (KMO = 0.925), indicating excellent sampling adequacy, and Bartlett's test of sphericity (χ^2 = 2486, df = 253, $p < 0.001$), which was statistically significant. A five-factor solution was specified based on the theoretical constructs of this study and the examination of the scree plot, collectively explaining 72.8% of the total variance. Factor loadings and item communalities (h^2) were examined. Although several items showed substantial primary loadings, others exhibited weak primary loadings or cross-loadings. Communality values ranged from 0.406 to 0.874.

Table 9 gives limited support to the proposed five-factor structure. The items for perceived usefulness formed the clearest group. Several other items had weak primary loadings or loaded on more than one factor. The five constructs were therefore not clearly separated in this sample. This may reflect overlap among the questionnaire domains, although common method variance may also have contributed. The analysis is exploratory and does not validate the instrument. We retained all items because they represented parts of the theoretical framework, and most had adequate communalities. Some primary loadings were below 0.40 [37]. Due to these substantial cross-loadings and the theoretical overlap among domains, the extracted factors cannot be considered empirically independent latent constructs in this sample. Consequently, the reliance on composite scores based on the a priori theoretical framework introduces inherent limitations for subsequent inferential analyses. These items require further examination in a larger, independent sample.

Table 9. Factor loadings and communalities (h^2) from exploratory factor analysis (EFA).

Item	Gamification	Ease of Use	Perceived Usefulness	Pedagogical Content	Behavioral Intention	h^2
Q14		0.153	**0.817**	0.119	−0.139	0.870
Q13			**0.795**	0.140		0.810
Q15			**0.791**			0.812
Q16			**0.790**		0.139	0.804
Q12	0.291		**0.454**			0.406
Q11	0.174	0.139	**0.327**	0.305		0.617
Q18		**0.996**				0.874
Q19		**0.825**				0.851
Q20		**0.801**				0.768
Q17		**0.671**	0.177		0.193	0.730
Q5	**0.695**	0.149	0.106	−0.108		0.666
Q1	**0.675**			0.134	0.182	0.655
Q2	**0.579**			0.286		0.616
Q3	**0.572**	0.186	0.167		−0.235	0.626
Q7	**0.410**	0.109	0.216	0.242	−0.235	0.623
Q4	**0.383**	0.143	0.366		0.115	0.635
Q9			0.170	**0.768**		0.811
Q6	0.108	0.144	−0.114	**0.736**		0.671
Q10			0.281	**0.597**		0.734
Q8	0.333		0.143	**0.596**	−0.181	0.822
Q21	0.126	0.182	0.353	0.118	**0.450**	0.823
Q23	0.162	0.241	0.214	0.120	**0.423**	0.714
Q22	0.246	0.118	0.270	0.256	**0.347**	0.813
SS Loadings	3.36	4.05	4.92	3.34	1.07	
% Variance	14.6%	17.6%	21.4%	14.5%	4.65%	72.8%

Note. Minimum residual extraction with oblimin rotation was used. All factor loadings with absolute values ≥ 0.10 are reported for transparency. h^2 represents communalities (1—uniqueness).

Because all questionnaire constructs were measured through self-report items collected from the same respondents at one time point, Harman's single-factor test was conducted as a limited post hoc diagnostic check. An unrotated EFA including all teacher-questionnaire items showed that the first factor accounted for 58% of the total variance. This result indicates a substantial general factor and means that common method variance cannot be ruled out. However, Harman's test does not distinguish method variance from conceptual overlap among the measures or a general positive response tendency. The result is therefore interpreted as an additional reason for caution, rather than as a quantitative estimate of common method bias.

Despite these limitations, the original composite scores were retained for subsequent analyses because they were specified a priori based on the theoretical frameworks that informed the instrument design (GEB, E-Guess and TAM), rather than derived from this sample-specific factor analysis. Consequently, the correlation and regression analyses should be interpreted as exploratory analyses based on theoretically defined composite measures, not as evidence regarding empirically independent latent constructs.

The results, as shown in Table 10, indicate moderate to relatively positive relationships among the extracted factors, ranging from 0.156 to 0.694. These findings indicate overlap among some of the factors extracted in this sample. The constructs were defined as

conceptually distinct during the development of the instrument. However, the EFA results indicate only partial empirical separation.

Table 10. Inter-factor correlation matrix derived from exploratory factor analysis (oblimin rotation). The matrix summarizes how the extracted factors relate to each other based on the analysis.

	Gamification	Pedagogical Content	Perceived Usefulness	Ease of Use	Behavioral Intention
Gamification		0.646	0.522	0.694	0.310
Pedagogical Content			0.627	0.566	0.259
Perceived Usefulness				0.503	0.189
Ease of Use					0.156
Behavioral Intention					

Note. Values represent correlations between factors extracted using minimum residual extraction with oblimin rotation. All correlations are positive.

Accordingly, these findings should be interpreted as exploratory and perception-based, emphasizing the relationships identified within the dataset rather than substantiating distinct or independent latent constructs.

3.4. Variables Associated with ICT Teachers' Reported Behavioral Intention to Use the Application (RQ4)

To address RQ4, Pearson correlation analysis was conducted to examine the relationships among the theoretically defined measures. This analysis was performed before the regression analysis to explore the associations among variables.

As shown in Table 11, all variables were positively and significantly correlated. Perceived usefulness ($r = 0.789$, $p < 0.001$), ease of use ($r = 0.748$, $p < 0.001$), pedagogical content ($r = 0.755$, $p < 0.001$) and gamification features ($r = 0.730$, $p < 0.001$) were found to correlate strongly with behavioral intention. These positive correlations indicate that higher ratings on these characteristics are associated with higher behavioral intention. However, caution is warranted in interpreting these connections, as all variables were assessed using self-report instruments within a single measurement context, which may have inflated the correlations due to common method bias. Because the EFA showed limited factor separation, these correlations should be viewed as associations among theoretically defined measures, not as relationships between empirically validated independent constructs.

Table 11. Pearson correlation matrix of the theoretically defined measures.

Evaluation Measure	Gamification	Pedagogical Content	Perceived Usefulness	Ease of Use	Behavioral Intention
Gamification	—				
Pedagogical Content	0.762 ***	—			
Perceived Usefulness	0.677 ***	0.818 ***	—		
Ease of Use	0.727 ***	0.705 ***	0.694 ***	—	
Behavioral Intention	0.730 ***	0.755 ***	0.789 ***	0.748 ***	—

Note. *** $p < 0.001$. The em dash (—) indicates diagonal elements for which correlations are not reported.

A regression analysis was performed to examine the patterns of association among the variables, not to determine causal relationships. The multiple regression analysis was regarded as exploratory since the EFA did not clearly separate the elements, as demonstrated in Table 12. Behavioral intention was used as the dependent variable, and gamification elements, pedagogical content, perceived usefulness and perceived ease of use were used as theoretically defined measures.

The purpose of the analysis was to examine how these variables relate in this sample, not to prove predictive validity or to show relationships between truly independent latent constructs. Given the empirical overlap and strong bivariate correlations among the theoretically defined measures ($r > 0.67$), the multiple regression analysis was treated

strictly as an exploratory analysis of the patterns of association. The resulting regression coefficients may be influenced by shared variance among overlapping measures and should therefore not be interpreted as isolated, independent effects of distinct latent constructs.

Following that, key regression assumptions were examined. The independence of residuals was confirmed by a Durbin–Watson statistic of 1.86 ($p = 0.420$). Although the Shapiro–Wilk test was significant ($p < 0.001$), visual inspection of the Q-Q plot (Figure 7) indicated that the residuals followed the normality line closely enough to satisfy the assumption for a sample of this size (N = 112). Linearity and homoscedasticity were also verified through residual plot analysis.

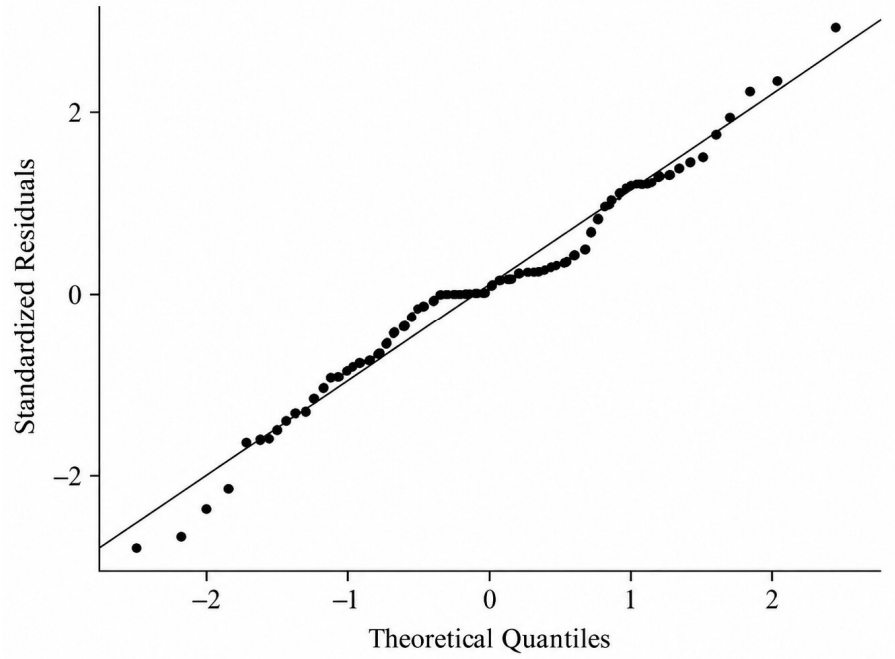


Figure 7. Q-Q plot of standardized regression residuals. Black dots represent the observed standardized residuals and the diagonal line represents the theoretical normal reference line. The regression model was statistically significant, F (4, 107) = 70.3, $p < 0.001$. The model accounted for 72.4% of the variance in reported behavioral intention ($R^2 = 0.724$). All variables were collected from the same respondents using the same self-report questionnaire at a single time point. Therefore, the relatively high proportion of explained variance ($R^2 = 0.724$) should be interpreted cautiously, as conceptual overlap and common method variance may have inflated the observed associations. Accordingly, the regression results should be viewed as exploratory associations among overlapping perceptions within this sample and not as evidence of predictive validity, actual system adoption or educational effectiveness. Within this exploratory model, perceived usefulness showed the largest standardized regression coefficient ($\beta = 0.390$, $p < 0.001$), followed by ease of use ($\beta = 0.262$, $p = 0.002$) and gamification elements ($\beta = 0.199$, $p = 0.023$). The coefficient for pedagogical content was not statistically significant when the other measures were included in the model ($\beta = 0.100$, $p = 0.328$). However, given the limited empirical separation among the measures, differences among the individual coefficients should not be interpreted as evidence of distinct or independent effects. In addition, the Variance Inflation Factor (VIF) values ranged from 2.56 to 4.04, remaining below commonly used thresholds for severe multicollinearity. Although the VIF values ranged from 2.56 to 4.04 and remained below commonly used thresholds for severe multicollinearity, they do not eliminate concerns about conceptual overlap and limited empirical separation among the measures identified in the EFA. Consequently, differences among the individual standardized coefficients should not be interpreted as evidence of differential importance among distinct constructs.

Table 12. Exploratory multiple regression analysis with reported behavioral intention as the dependent variable.

Evaluation Measure	β	*p*	VIF
Gamification	0.199	0.023	2.88
Pedagogical Content	0.100	0.328	4.04
Perceived Usefulness	0.390	<0.001	3.29
Ease of Use	0.262	0.002	2.56

3.5. Comparison of ICT Teachers and ICT Educational Advisors (RQ5)

In response to RQ5, Welch's independent-samples *t*-tests were carried out to examine whether there are any differences between ICT teachers and ICT educational advisors on the five evaluation dimensions. Descriptive statistics were also calculated to accurately represent the characteristics of the groups.

As shown in Table 13, both groups reported generally high self-reported evaluations across all dimensions. ICT teachers reported slightly higher mean ratings for gamification elements (M = 4.41) and ease of use (M = 4.44) than ICT educational advisors. Additionally, ICT teachers recorded marginally higher scores in pedagogical content and perceived usefulness, although the differences between the two groups were relatively small. In contrast, ICT educational advisors reported a slightly higher behavioral intention rating (M = 4.38). In general, the descriptive comparison suggests that both groups reported generally positive perceptions of the application within the context of this evaluation. However, the consistently high mean scores across both groups should be interpreted with caution, as they may reflect response tendencies associated with self-reported data, such as acquiescence bias or novelty effects.

Table 13. Descriptive comparison between ICT teachers and ICT educational advisors.

	Group	Mean	Median	SD	SE
Gamification Elements	ICT Teachers	4.41	4.50	0.608	0.0575
	ICT Educational Advisors	4.29	4.20	0.444	0.123
Pedagogical Content	ICT Teachers	4.36	4.43	0.636	0.0601
	ICT Educational Advisors	4.32	4.29	0.456	0.126
Perceived Usefulness	ICT Teachers	4.22	4.13	0.785	0.0742
	ICT Educational Advisors	4.19	4.25	0.596	0.165
Ease of Use	ICT Teachers	4.44	4.75	0.664	0.0628
	ICT Educational Advisors	4.35	4.50	0.673	0.187
Behavioral Intention	ICT Teachers	4.29	4.33	0.791	0.0747
	ICT Educational Advisors	4.38	4.33	0.488	0.135

Welch's independent-samples *t*-tests were conducted to examine differences between ICT teachers and ICT educational advisors across five evaluation dimensions: gamification elements, pedagogical content, perceived usefulness, ease of use and behavioral intention. The analysis included Welch's t statistic, degrees of freedom (df), significance level (*p*-value),

the mean difference between the two groups, the standard error of the mean difference and the effect size (Cohen's d).

As shown in Table 14, Welch's independent-samples *t*-tests indicated no statistically significant differences between ICT teachers and ICT educational advisors across any of the five evaluation dimensions ($p > 0.05$ for all comparisons). The calculated effect sizes (Cohen's d) were small across all comparisons. Consistent with these effect sizes, the estimated group differences were small. However, the ICT advisor sample included only 13 participants, so the estimates remain imprecise, and small differences may not have been detected.

Table 14. Comparison of ICT teachers and ICT educational advisors using Welch's independent-samples *t*-tests across the five evaluation dimensions.

	Welch's t	df	*p*	Mean Difference	SE Difference	Cohen's d
Gamification Elements	0.858	17.7	0.403	0.117	0.136	0.219
Pedagogical Content	0.329	17.9	0.746	0.046	0.140	0.083
Perceived Usefulness	0.146	17.2	0.886	0.026	0.181	0.038
Ease of Use	0.486	14.8	0.634	0.096	0.197	0.143
Behavioral Intention	−0.621	20.2	0.542	−0.096	0.155	−0.146

Note. Welch's independent-samples *t*-tests were used because of the unequal group sizes. Mean difference = ICT teachers-ICT educational advisors.

In light of the limited factor separation observed in the EFA, the original composite scores were retained solely as theoretically defined, context-specific summary measures. They should not be interpreted as empirically validated or discriminant latent constructs. Accordingly, the group comparisons reported here should be interpreted as exploratory comparisons of theoretically defined composite measures within this sample.

4. Discussion

We chose to perform a multi-stakeholder evaluation prior to directly evaluating TouriBot with its intended student users. This aligns with established instructional design approaches and shows that professional stakeholder assessments can provide early formative insights into the design and perceived usefulness of educational software [32]. This process provides an opportunity to examine TouriBot's interaction design and pedagogical scaffolding and to identify potential issues before direct evaluation in classroom settings [14,31]. Such an evaluation represents an initial formative step in examining the newly integrated AI component and builds on our previous evaluation of the baseline application with students [29]. Bringing Large Language Models (LLMs) into K-12 contexts creates new opportunities but also raises ethical and instructional problems. Thus, formative evaluation can help identify potential interaction, usability and pedagogical concerns before direct evaluation with young students [4].

Figure 4 presents the multi-stakeholder evaluation design and the distinct perspectives contributed by the three stakeholder groups. Each group of participants had its own perspective. They provide different but complementary insights into the overall operation of the system. ICT teachers mainly commented on usability and classroom practicality, but ICT educational advisors provided a broader pedagogical and strategic viewpoint. Instead, UX experts focused more on the design of the user interface and basic principles of usability.

The findings indicate generally positive participant perceptions of the system's usability and pedagogical value. However, this study did not directly assess student learning

outcomes. The results should therefore be interpreted as adult stakeholders' perceptions of TouriBot within a specific evaluation context and educational scenario, not as evidence of educational effectiveness, student usability or readiness for classroom use.

The findings are discussed in relation to each research question. Research Question 1 (RQ1) explores the perspectives of educational advisors in the field of ICT. The results showed that the ICT educational advisors evaluated the application positively in terms of its perceived pedagogical value and usability. Research Question 2 (RQ2) examines the perception-based usability assessment conducted with UX experts. Research Question 3 (RQ3) investigates ICT teachers' perspectives. Furthermore, Research Question 4 (RQ4) investigates the variables related to reported behavioral intentions, and Research Question 5 (RQ5) compares the evaluations of ICT teachers and ICT educational advisors across the five theoretically defined measures.

The results of RQ2 showed that UX experts provided generally positive ratings on most of the usability dimensions informed by Nielsen's heuristics, while flexibility and aesthetic design received comparatively lower ratings. ICT teachers in RQ3 reported rather high ratings on TAM-based measures, including perceived usefulness and perceived ease of use. The findings from RQ4 indicated that perceived usefulness, perceived ease of use and gamification components were positively associated with ICT teachers' expressed intention to use the system. Perceived usefulness showed the strongest adjusted association in the regression model ($\beta = 0.390$, $p < 0.001$). Finally, no statistically significant differences were detected in Research Question (RQ5) in the perceptions of ICT teachers and ICT educational advisors about the educational value and usability of TouriBot.

Nevertheless, the high ratings across all components and the negatively skewed distributions indicate the need for careful interpretation. The observed pattern may partly reflect response tendencies associated with self-report evaluation, including acquiescence bias, as well as possible novelty effects associated with participants' exposure to a newly introduced AI-supported educational chatbot. These potential influences were not measured directly and should therefore be regarded as possible sources of bias rather than established explanations for the high ratings. These issues are discussed further in the limitations section.

The pattern of the results is broadly consistent with TAM, in which perceived usefulness is expected to be associated with behavioral intention. In the exploratory regression model, perceived usefulness showed the strongest adjusted association with behavioral intention. However, this result should be interpreted carefully. The high correlations among ratings across the measured dimensions, the weak factor separation in the EFA, and the fact that all the data came from one self-report questionnaire at a single time point suggest that overlap among the theoretically defined measures and shared method variance may have influenced the observed associations.

Accordingly, the regression results should be understood as indicating adjusted associations rather than as evidence of causal relationships. The theoretically informed associations with behavioral intention are based on self-reported perceptions and do not provide evidence of actual system adoption. Instead, they reflect participants' reported willingness to use the system under the study's conditions. These findings should therefore not be interpreted as evidence of instructional effectiveness or learning impact, but rather, as indicative of perceived system characteristics within a pre-deployment evaluation context.

Notably, despite the low expert ratings for flexibility ($M = 2.63$) and moderate ratings for aesthetics ($M = 3.50$), both ICT teachers and ICT educational advisors reported relatively high stated behavioral intention. This pattern indicates a descriptive difference between the perspectives captured by the stakeholder-specific measures. However, because the stakeholder groups had different professional roles and completed different evaluation mea-

sures, and because the UX expert group was small, this pattern should not be interpreted as evidence of a systematic discrepancy between perceived usability and behavioral intention.

The relatively high ratings for perceived ease of use reported by ICT teachers can be considered alongside the UX expert ratings structured around Nielsen's usability heuristics. These assessments provide different but complementary perspectives: the TAM-based measures capture ICT teachers' perceptions of the usefulness and ease of use of TouriBot, whereas the heuristic-informed expert assessment provides descriptive information about specific aspects of the interface that may warrant further design attention. These frameworks are therefore treated as complementary evaluation perspectives rather than as an integrated model or as mutually validating measures.

Previous research on expert-based usability evaluation has shown that judgments may vary across evaluators [38]. The present study did not examine inter-evaluator agreement. Therefore, the UX ratings are interpreted descriptively as the assessments of the participating experts rather than as definitive estimates, consistent with a cautious approach to the interpretation and communication of quantitative HCI findings [39].

TouriBot was designed to provide guidance and hints without directly providing answers, reflecting the intended scaffolding strategy of the system. However, the effectiveness of this approach has not been empirically evaluated with students.

From a human–computer interaction perspective, usability may influence participants' perceptions of the system's potential educational value, as reported in this pre-deployment evaluation context. From a design perspective, reducing unnecessary cognitive demands may help learners focus on task-relevant information, while appropriate guidance can support structured engagement with learning activities. In this sense, these principles informed the design rationale of the interface. However, the present study did not measure students' cognitive load or engagement. Accordingly, whether these design features support engagement with critical literacy-related learning objectives requires empirical evaluation with the intended student users in classroom settings [40,41].

Among the dimensions informed by Nielsen's heuristics, recognition rather than recall received the highest UX expert rating (M = 4.88, SD = 0.35). This indicates that the participating experts generally perceived the interface as minimizing the need for users to rely on memory by making relevant options visible. In contrast, flexibility and efficiency of use received the lowest rating (M = 2.63, SD = 1.41), indicating that the participating experts perceived limited flexibility in accommodating different user levels or interaction needs.

The perception-based usability assessment revealed that two dimensions rated comparatively lower than the others, which may warrant attention before student-centered evaluation in classroom settings. First, low expert ratings for flexibility and efficiency of use (M = 2.63), and second, moderate ratings for aesthetic and minimalist design (M = 3.50). Particular attention should be given to the low scores for flexibility and efficiency of use, which suggest that the UX experts who participated perceived limitations in the system's ability to adapt to different interaction methods or levels of user experience. Given the perception-based nature of the assessment, this finding should be interpreted as an indication of a potential area for further design refinement rather than as evidence of an objectively established usability problem.

Additionally, the moderate aesthetic evaluation suggests that the environment needs improvement in areas such as organization and minimalism, thereby reducing the visual noise created by non-essential elements. Thus, these aspects can be regarded as indicators and directions for enhancement rather than mere weaknesses. Therefore, we may conclude that there is a need for a clearer interface, better visual consistency and more flexible interaction routes, while keeping the original design focus of the system for students aged 9–12.

In order to collect complementary but analytically separate stakeholder viewpoints in an early formative evaluation, this study used pedagogical evaluation measures, heuristic-informed UX assessments and TAM-based measurements. These frameworks served as complementary conceptual lenses for organizing the evaluation rather than as an integrated theoretical model tested by the present data.

4.1. Limitations and Scope of the Evaluation

This study presents an early formative evaluation of an AI-supported educational chatbot. It focused on perceived usability, pedagogical value and potential classroom applicability. This research did not include in vivo classroom implementation and direct measurement of student learning outcomes. The interpretation of the findings should take into account several sampling-related limitations. The sample of ICT teachers (N = 112) was one of convenience and voluntary participation, which means that there is a possibility of self-selection bias. Participants with a higher interest in educational technology may have been more likely to participate in this study, and this potential self-selection should be considered when interpreting the consistently high ratings observed across the measured dimensions.

The expert samples were relatively small. They comprised 13 ICT educational advisors and 8 UX experts. The UX expert sample (N = 8) was used for the perception-based usability assessment. However, because of its small size, it can yield only limited quantitative questionnaire results. This implies that the reported means and standard deviations for the UX evaluation should be interpreted descriptively as summaries of the participating experts' assessments rather than as statistically generalizable estimates.

Although the ICT advisors' sample includes a large portion of the national population (approximately 45%), the findings are presented as informed expert opinions rather than evidence typical of a larger population. The findings are further limited by the use of self-reported data, which may be vulnerable to response biases such as social desirability and acquiescence. Consequently, the observed associations among the theoretically defined questionnaire measures may partly reflect shared method variance and conceptual overlap rather than relationships among clearly distinct constructs. Therefore, the high reported intention to use the system should be interpreted as an indicator of perceived acceptance in the research context rather than as evidence of actual classroom implementation.

The adapted questionnaire also has limitations regarding construct validity. As reported above, the EFA offered only limited support for the intended five-factor structure, with several items showing weak primary loadings or loading on more than one factor. In addition, the questionnaire has not been tested on a different group of participants. The post hoc Harman test identified a dominant general factor accounting for 58% of the item variance. Together with the high correlations among the theoretically defined measures and limited factor separation, this finding indicates that shared method effects, conceptual overlap and general evaluative response tendencies may have contributed to the observed associations. Given the limited factor separation, high correlations among the theoretically defined measures and potential shared method variance, the individual regression coefficients should not be interpreted as unique or empirically isolated effects of distinct latent constructs. The high proportion of variance accounted for by the exploratory regression model ($R^2 = 0.724$) should also be interpreted in light of these measurement limitations.

Furthermore, this study took place in a specific sociocultural and educational setting. This may limit the transferability of the findings to other educational systems, pedagogical techniques or technical systems.

In addition, the evaluation was based exclusively on adult stakeholder perceptions and did not include direct evaluation with students, who are the intended users of the

system. ICT teachers and educational advisors can provide professionally informed recommendations of pedagogical design and possible classroom application, but their judgments cannot substitute for the experiences of the intended student users. More specifically, ratings of the questionnaire items on expected levels of engagement, difficulty, usability, relevance to learning and support for critical literacy are professional perceptions and not direct evidence of suitability, experience or educational effectiveness at the student level. Thus, direct evaluation with primary school students is necessary to examine how the intended users interact with and experience TouriBot in practice.

While all participants followed the same predefined sequence of activities, their interaction with TouriBot was not standardized. Participants could choose whether and when to use the chatbot during each activity and what questions to ask, but they could not alter the predefined sequence of activities. This flexibility reflects TouriBot's intended role as an on-demand support tool, since it cannot be predetermined whether, when, or at what point an individual learner may require assistance. Nevertheless, this design may have resulted in differences in the frequency, timing and content of chatbot interactions and, consequently, in different levels of exposure to TouriBot before completing the evaluation questionnaire. This variability should therefore be considered when interpreting participants' evaluations.

A further limitation is the absence of a dedicated system-level technical evaluation of TouriBot. Although retrieval-augmented generation and response-control mechanisms were incorporated into the system design, the present study did not systematically assess the retrieval accuracy, hallucination frequency, prompt robustness, age appropriateness, or consistency of the intended Socratic interaction strategy. The present study focused on perception-based and UX evaluation rather than technical validation of the underlying AI mechanisms. Therefore, the reported stakeholder and UX assessments should not be interpreted as evidence of system-level technical performance.

4.2. Future Research

Building on the limitations discussed above, future research should incorporate more systematic documentation of participants' interactions with the chatbot, such as logs that record the frequency, timing and type of use, while maintaining TouriBot's intended on-demand support function.

Given the limited factor separation observed in the present study, further research should refine the questionnaire and examine its dimensional structure in larger, independent samples to determine whether the theoretically defined measures can be more clearly distinguished empirically.

In addition, future multi-stakeholder research should move beyond parallel assessments of separate stakeholder groups toward a more integrated evaluation design that brings together the perspectives of ICT teachers, ICT educational advisors, UX experts and students within a common analytical framework.

Dedicated system-level evaluation should also complement the present perception-based findings by systematically examining retrieval accuracy, hallucination occurrence, prompt robustness, age appropriateness and consistency of the intended Socratic interaction strategy.

Future research may also extend the present evidence through direct student-centered and classroom-based evaluation of TouriBot, including observational evidence and direct assessment of learning outcomes. Together, these complementary directions would address key limitations of the present study and provide a more comprehensive basis for examining TouriBot across methodological, technical and educational dimensions.

5. Conclusions

This study represents an initial formative evaluation based on the perspectives of ICT teachers, ICT educational advisors and UX experts. It examined perceived pedagogical value, usability and potential classroom applicability from these adult stakeholder perspectives.

Specifically, ICT teachers and ICT educational advisors generally reported positive perceptions across the evaluation measures. UX experts gave generally positive ratings across most aspects assessed through the heuristic-informed and conversational interaction measures. They also reported noticeably lower ratings for aspects such as interaction flexibility and aesthetic and minimalist design.

The principal contribution of this study is the multi-stakeholder formative evaluation of TouriBot, an AI-supported educational chatbot integrated into a gamified learning environment addressing misinformation in primary education, from three complementary professional perspectives: ICT teachers, ICT educational advisors and UX experts. By bringing together classroom-oriented, broader pedagogical and UX perspectives, this study provides a broader formative view of perceived pedagogical value, usability, potential classroom applicability and interface-related strengths and limitations. This multi-stakeholder approach does not constitute an integrated validation framework. Rather, it provides complementary formative evidence that can inform subsequent research and system refinement.

The present study does not provide evidence of student learning outcomes or classroom effectiveness. Although the findings derive from a single application evaluated within a single national and educational context, they identify specific areas for further refinement, particularly interaction flexibility and interface design.

Author Contributions: Conceptualization, A.K.; methodology, N.A. and S.P.; software, A.K.; validation, A.K.; formal analysis, A.K.; investigation, A.K.; data curation, A.K.; writing—original draft preparation, A.K.; writing—review and editing, N.A. and S.P.; visualization, A.K.; supervision, N.A. and S.P.; project administration, N.A. and S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The ethical approval for the study was obtained from Ionian University under Protocol No. 89/08-01-2026.

Informed Consent Statement: The study involved adult participants only and participation was voluntary and anonymous. No personally identifiable information was collected through the questionnaire.

Data Availability Statement: The data are not publicly available due to privacy and ethical restrictions related to the approved research protocol.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Questionnaire Instruments

Appendix A.1. Teacher and ICT Advisor Questionnaire

The questionnaire used in this study was developed by adapting items from the Gaming Educational Balanced (GEB) model, the E-Guess model, the Technology Acceptance Model (TAM) and the UTAUT framework.

The instrument was originally developed in Greek and translated into English for reporting purposes. All items were adapted to reflect the specific context of TouriBot embedded in a gamified learning environment, focusing on critical literacy. The questionnaire reflects participants' perceptions and does not represent or replace learning outcomes. Several changes and adjustments were made before distributing it to the experts to ensure its validity, clarity and alignment with the educational content and pedagogical objectives.

The instrument represents an adapted, context-specific questionnaire and should be considered appropriate for exploratory and formative evaluation purposes rather than as a fully validated measurement scale.

Table A1. Teacher and ICT advisor questionnaire items.

Code	Item Statement	Framework
Q1	The application is perceived as attractive and engaging for students.	GEB
Q2	The difficulty level is perceived as appropriate and supportive of student engagement.	GEB
Q3	The application is perceived as motivating students to participate actively.	GEB
Q4	The game elements are perceived as supporting the learning experience.	GEB
Q5	The interface is perceived as appropriate for students aged 9–12.	GEB
Q6	The pedagogical content is perceived as aligned with critical literacy principles.	E-Guess
Q7	The activities are perceived as appropriate for the intended learning objectives.	E-Guess
Q8	The application is perceived as supporting meaningful learning experiences.	E-Guess
Q9	The application is perceived as connecting learning with real-life contexts.	E-Guess
Q10	The application is perceived as allowing students to reflect on their understanding.	E-Guess
Q11	The learning activities are perceived as engaging and interactive.	E-Guess
Q12	The pedagogical content is perceived as appropriate for the target age group.	E-Guess
Q13	The application is perceived as useful for supporting critical literacy-related skills.	TAM
Q14	The application is perceived as helping students evaluate information.	TAM
Q15	The application is perceived as supporting the achievement of learning objectives.	TAM
Q16	The application is perceived as supporting the learning process.	TAM
Q17	The application is perceived as easy for students to use.	TAM
Q18	Learning to use the application is perceived as easy.	TAM
Q19	The application is perceived as clear and understandable.	TAM
Q20	Navigation within the application is perceived as easy.	TAM
Q21	I would recommend this application for classroom use.	UTAUT
Q22	I intend to use this application in my teaching practice.	UTAUT
Q23	I would encourage other educators to use this application.	UTAUT

Note. Questionnaire items adapted from the GEB model [25], E-Guess framework [26], TAM and UTAUT-based behavioral intention constructs [24]. All items were contextually adapted to evaluate TouriBot in primary education.

Appendix A.2. UX Experts' Evaluation Instrument

The perception-based usability assessment completed by UX experts was informed by Nielsen's ten usability heuristics. Each heuristic was operationalized as a perception-based statement and rated using a five-point Likert scale.

Table A2. Nielsen heuristics-based questionnaire.

Code	Heuristic	Item Statement
H1	Visibility of system status	The system is perceived as clearly communicating its current status to the user.
H2	Match with the real world	The language and symbols used are perceived as familiar and understandable.
H3	User control and freedom	The system is perceived as allowing users to undo, redo, or navigate freely between actions.
H4	Consistency and standards	The system is perceived as consistent and aligned with established design standards.
H5	Error prevention	The system is perceived as preventing errors before they occur.
H6	Recognition rather than recall	The system is perceived as minimizing memory load by making options visible.
H7	Flexibility and efficiency	The system is perceived as adaptable to both novice and experienced users.
H8	Aesthetic and minimalist design	The interface is perceived as clear and free from unnecessary elements.
H9	Error recognition and recovery	Error messages are perceived as clear and helpful.
H10	Help and documentation	Help and support are perceived as available and useful when needed.

Note. The questionnaire items represent perception-based adaptations of Nielsen's ten usability heuristics for TouriBot [31].

In addition to heuristic evaluation, UX experts assessed the perceived quality of TouriBot interactions using perception-based items grounded in conversational design principles.

Table A3. TouriBot interaction evaluation.

Code	Dimension	Item Statement
C11	Conversational Clarity	The chatbot responses are perceived as clear and understandable for users aged 9–12.
C12	Conversational Clarity	The chatbot responses are perceived as relevant to the context of each activity.
C13	Conversational Clarity	The chatbot is perceived as providing sufficient but not excessive information during interactions.

Table A3. *Cont.*

Code	Dimension	Item Statement
C14	Conversational Clarity	The interaction with the chatbot is perceived as coherent and naturally flowing.
C15	Social Presence and Persona	The chatbot is perceived as having a clear role and communication style.
C16	Social Presence and Persona	The chatbot's communication style is perceived as friendly and pedagogically appropriate for users aged 9–12.
C17	Social Presence and Persona	The chatbot persona is perceived as supporting user engagement and interest.
C18	Transparency and User Control	Users are perceived as understanding why the chatbot provides specific responses or guidance.
C19	Transparency and User Control	The user is perceived as maintaining a sense of control during interaction.
C20	Transparency and User Control	It is perceived as clear when the chatbot provides information, guidance, or support.
C21	Transparency and User Control	Users are perceived as able to recognize when the chatbot makes an error and how to proceed.
C22	Error Recovery and Dialog Repair	When users do not understand a response, the chatbot is perceived as providing appropriate clarification.
C23	Error Recovery and Dialog Repair	In cases of incorrect or inappropriate responses, the chatbot is perceived as not discouraging the user.
C24	Error Recovery and Dialog Repair	The chatbot is perceived as supporting the smooth continuation of interaction after errors.

Note. Items C11–C14 (Conversational Clarity) were informed by conversational communication principles and chatbot interaction research [12,34,35]. Items C15–C17 (Social Presence and Persona) were informed by media equation and social presence perspectives in human–computer interaction [12,35]. Items C18–C24 (Transparency and User Control; Error Recovery and Dialog Repair) were informed by established usability principles and the chatbot interaction literature [12,31]. All items were specifically developed and adapted to the educational context of an AI-supported educational chatbot, called TouriBot, for students aged 9–12 years and were intended for UX expert evaluation.

References

1. Kharrufa, A.; Johnson, I.G. The Potential and Implications of Generative AI on HCI Education. In *Proceedings of the 6th Annual Symposium on HCI Education (EduCHI '24)*; ACM: New York, NY, USA, 2024; pp. 1–8. [CrossRef]
2. York, E. Evaluating ChatGPT: Generative AI in UX Design and Web Development Pedagogy. In *Proceedings of the 41st ACM International Conference on Design of Communication*; SIGDOC '23; Association for Computing Machinery: New York, NY, USA, 2023; pp. 197–201. [CrossRef]
3. Chee, K.N.; Abdul Rahman, M.H.; Yahaya, N.; Ibrahim, N.; Abdul Razak, R.; Sugino, C. Exploring the Trend and Potential Distribution of Chatbot in Education: A Systematic Review. *Int. J. Inf. Educ. Technol.* **2023**, *13*, 516–525. [CrossRef]

4. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
5. Zhai, X. ChatGPT for Next Generation Science Learning. *XRDS Crossroads ACM Mag. Stud.* **2023**, *29*, 42–46. [CrossRef]
6. Yuan, Y. An Empirical Study of the Efficacy of AI Chatbots for English as a Foreign Language Learning in Primary Education. *Interact. Learn. Environ.* **2024**, *32*, 6774–6789. [CrossRef]
7. Škobo, M. Dual Chatbots for K–12 Literature Education: A Serbian Case Study of AI-Powered Dialogue Systems. In *Virtual Tutors and AI-Powered Instructional Tools in K-12 Settings*; IGI Global: Hershey, PA, USA, 2025; pp. 45–68. [CrossRef]
8. Ait Baha, T.; El Hajji, M.; Es-Saady, Y.; Fadili, H. The Impact of Educational Chatbot on Student Learning Experience. *Educ. Inf. Technol.* **2024**, *29*, 10153–10176. [CrossRef]
9. Dos Santos, L.F.; Oliveira, W.; Corrêa de Lima, A.; de Castro Junior, A.A.; Hamari, J. The Effects of Gamification on Learners' Engagement According to Their Gamification User Types. *Technol. Knowl. Learn.* **2025**, 1–21. [CrossRef]
10. Wollny, S.; Schneider, J.; Di Mitri, D.; Weidlich, J.; Rittberger, M.; Drachsler, H. Are We There Yet? A Systematic Literature Review on Chatbots in Education. *Front. Artif. Intell.* **2021**, *4*, 654924. [CrossRef] [PubMed]
11. Elmaadaway, M.A.N.; El-Naggar, M.E.; Abouhashesh, M.R.I. Improving Primary School Students' Oral Reading Fluency Through Voice Chatbot-Based AI. *J. Comput. Assist. Learn.* **2025**, *41*, e70019. [CrossRef]
12. Belda-Medina, J.; Kokošková, V. Integrating Chatbots in Education: Insights from the Chatbot-Human Interaction Satisfaction Model (CHISM). *Int. J. Educ. Technol. High Educ.* **2023**, *20*, 62. [CrossRef]
13. Škobo, M.; Šović, M. Teachers' Perspectives on Chatbots and AI Agents in Primary and Secondary Education. *Int. J. Stud. Educ.* **2025**, *7*, 410–442. [CrossRef]
14. Nikou, S.A.; Chang, M. *Learning by Building Chatbot: A System Usability Study and Teachers' Views About the Educational Uses of Chatbots*; Frasson, C., Mylonas, P., Troussas, C., Eds.; Springer Nature: Berlin/Heidelberg, Germany, 2023; pp. 342–351. [CrossRef]
15. Potter, W.J. *Theory of Media Literacy: A Cognitive Approach*; SAGE Publications: Thousand Oaks, CA, USA, 2010.
16. Loos, E.; Nijenhuis, J. Consuming Fake News: A Matter of Age? The Perception of Political Fake News Stories in Facebook Ads. In *Human Aspects of IT for the Aged Population. Technology and Society*; Springer: Cham, Switzerland, 2020. [CrossRef]
17. Stauch, L.; Renninger, D.; Rangnow, P.; Hartmann, A.; Fischer, L.; Dadaczynski, K.; Okan, O. Digital Health Literacy of Children and Adolescents and Its Association with Sociodemographic Factors: Representative Study Findings from Germany. *J. Med. Internet Res.* **2025**, *27*, e69170. [CrossRef] [PubMed]
18. Hall, S. Encoding and Decoding in the Television Discourse. In *Culture, Media, Language*; University of Birmingham Centre for Contemporary Cultural Studies, Ed.; Routledge: London, UK, 1980; pp. 128–138.
19. Vygotskij, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Cole, M., John-Steiner, V., Scribner, S., Souberman, E., Eds.; Harvard University Press: Cambridge, MA, USA; London, UK, 1978.
20. Quintana, C.; Reiser, B.J.; Davis, E.A.; Krajcik, J.; Fretz, E.; Duncan, R.G.; Kyza, E.; Edelson, D.; Soloway, E. A Scaffolding Design Framework for Software to Support Science Inquiry. *J. Learn. Sci.* **2004**, *13*, 337–386. [CrossRef] [PubMed]
21. Aleven, V.; Mclaren, B.; Koedinger, K. Towards Computer-Based Tutoring of Help-Seeking Skills. In *Help Seeking in Academic Settings: Goals, Groups, and Contexts*; Taylor & Francis Group: Abingdon, UK, 2006. [CrossRef]
22. Carroll, J. Human-Computer Interaction: Psychology as a Science of Design. *Annu. Rev. Psychol.* **1997**, *48*, 61–83. [CrossRef] [PubMed]
23. Doherty, K.; Doherty, G. Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv.* **2018**, *51*, 99. [CrossRef]
24. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef] [PubMed]
25. Martinez, K.; Menéndez-Menéndez, M.I.; Bustillo, A. A New Measure for Serious Games Evaluation: Gaming Educational Balanced (GEB) Model. *Appl. Sci.* **2022**, *12*, 11757. [CrossRef]
26. Silveira, A.; Martins, R.; Vieira, E. E-Guess: Usability Evaluation for Educational Games. *RIED Rev. Iberoam. Educ. A Distancia* **2021**, *24*, 245–263. [CrossRef]
27. Hellenic Republic. *Law 4823/2021: Upgrading the School, Empowering Teachers and Other Provisions*; Government Gazette: Athens, Greece, 2021; Volume 136/A. Available online: https://www.kodiko.gr/nomothesia/document/739038/nomos-4823-2021 (accessed on 12 March 2026).
28. Genially. The Easiest Way to Create Interactive Experiences. Available online: https://genially.com/ (accessed on 8 February 2026).
29. Koutsikou, A.; Antonopoulos, N. Gamification and User Experience in Fake News Detection on Tourism in Primary Education. *Electronics* **2025**, *14*, 2200. [CrossRef]
30. Chatbase. AI Chatbot Platform. Available online: https://www.chatbase.co/ (accessed on 8 February 2026).
31. Nielsen, J.; Mack, R.L. *Usability Inspection Methods*; John Wiley & Sons: New York, NY, USA, 1994.
32. Squires, D.; Preece, J. Predicting Quality in Educational Software. *Interact. Comput.* **1999**, *11*, 467–483. [CrossRef]

33. Tzamos, N.; Ioannou, D.; Katsanos, C. Wordinary: A Software Tool for Teaching Greek Word Families to Elementary School Students. In *Proceedings of the 26th Pan-Hellenic Conference on Informatics*; PCI '22; Association for Computing Machinery: New York, NY, USA, 2023; pp. 192–199. [CrossRef]
34. Radziwill, N.M.; Benton, M.C. Evaluating Quality of Chatbots and Intelligent Conversational Agents. *arXiv* **2017**, arXiv:1704.04579. [CrossRef]
35. Reeves, B.; Nass, C.I. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*; Cambridge University Press: New York, NY, USA, 1996.
36. King, W.R.; He, J. A Meta-Analysis of the Technology Acceptance Model. *Inf. Manag.* **2006**, *43*, 740–755. [CrossRef]
37. Hair, J.F.; Black, W.C.; Babin, J.B.; Anderson, R.E. *Multivariate Data Analysis: A Global Perspective*; Pearson Education: Upper Saddle River, NJ, USA, 2010.
38. Hertzum, M.; Jacobsen, N. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *Int. J. Hum.-Comput. Interact.* **2001**, *13*, 421–443. [CrossRef] [PubMed]
39. Dragicevic, P. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*; Robertson, J., Kaptein, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 291–330. [CrossRef]
40. Mayer, R.; Moreno, R. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educ. Psychol.* **2003**, *38*, 43–52. [CrossRef] [PubMed]
41. Wood, D.; Bruner, J.S.; Ross, G. The Role of Tutoring in Problem Solving. *J. Child Psychol. Psychiatry* **1976**, *17*, 89–100. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.