International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

From Priming Modalities to Educational AI Chatbot Engagement: A Study of 67 Learners

Trang Xuan, Joni Salminen, Ilkka Kaate, Farhan Ahmed, Danial Amin, Rajat Patil, Soon-Gyo Jung, Jinan Y. Azem & Bernard J. Jansen

To cite this article: Trang Xuan, Joni Salminen, Ilkka Kaate, Farhan Ahmed, Danial Amin, Rajat Patil, Soon-Gyo Jung, Jinan Y. Azem & Bernard J. Jansen (06 May 2026): From Priming Modalities to Educational AI Chatbot Engagement: A Study of 67 Learners, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2654818

To link to this article: https://doi.org/10.1080/10447318.2026.2654818

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2654818

From Priming Modalities to Educational AI Chatbot Engagement: A Study of 67 Learners

Trang Xuan[a], Joni Salminen[a], Ilkka Kaate[b], Farhan Ahmed[a], Danial Amin[a], Rajat Patil[a], Soon-Gyo Jung[c], Jinan Y. Azem[c] and Bernard J. Jansen[c]

[a]School of Marketing and Communication, University of Vaasa, Vaasa, Finland; [b]Turku School of Economic, University of Turku, Turku, Finland; [c]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

ABSTRACT

Learner interactions with AI have become crucial in education. A study with sixty-seven learners in a within-subject quasi-experimental study was conducted. Participants were randomly assigned to counterbalance sequences of visual and audio learning materials about wild species, completing a pre-survey, two learning sessions with an educational AI chatbot called Cipherbot and post session experience surveys. This study focused on learner interactions with Cipherbot, specifically (1) how material modalities influenced recall; (2) chatbot engagement through behavioral patterns; (3) material modalities' influence on task completion; and (4) experiential interaction exploring how prior experiences moderated learning behaviors. Interactions were measured using surveys, audio-transcripts, and chatlogs. Results showed learners achieved higher knowledge recall with visual materials. Experienced AI-users interacted significantly less, whereas task-experienced learners interacted more after audio materials. The audio-butterfly sequence showed order effects, with learners spending more time producing more utterances. Chatbot-user guidelines with multiple modalities are recommended to enhance learning experience.

KEYWORDS

Learning; empirical studies for user behavior; multimodal interaction; artificial intelligence applications

1. Introduction

Many learners often assume that more information, such as media, modes, or stimulation, leads to better learning. However, cognitive research often shows the opposite: adding additional modalities can harm understanding by overwhelming working memory (Sweller, 2011). In video, audio, and interactivity environments, some students recall less in this multimodal content because multimodality competes for limited cognitive resources (Ginns, 2005). Prior work reports that learning improves when presentation reduces cognitive burden and directs attention to essential information (Mayer, 2024). This conundrum raises an important question for educators in the age of artificial intelligence (AI) education. As platforms become increasingly multimodal, the most effective learning strategy is a more targeted modality. Understanding how different modalities affect learning is central to designing effective AI-mediated learning experiences, services, and platforms.

AI chatbots have strongly influenced education in multiple aspects and have opened new venues for developing more innovative pedagogical strategies for learner experiences. Many applications have been studied in AI-assisted learning, from early education to higher education (Byrne, 2013; X. Chen et al., 2020; Deveci Topal et al., 2021; Fang & Tse, 2022). However, current research has not properly addressed how different priming modalities influence knowledge acquisition and retention, leaving educators without empirical guidance for integrating these technologies effectively in educational settings. Priming modalities refer to media that learners use in their study sessions (e.g., text, image, audio, video), which lead to changes in how they interact with the set-up learning environment. In this paper,

CONTACT Trang Xuan x9036166@student.uwasa.fi University of Vaasa, Wolffintie 34 65200 Vaasa, Finland

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2654818.

we focus on two main issues: (1) the priming modalities affecting learning experiences, and (2) testing the priming modalities affecting the usage of *educational AI chatbots (EAICs)*.

Regarding AI chatbots in an educational context, Yaacob et al. (2025) studied ChatGPTs perceived positively among learners as valuable and supportive of academic activities. Other studies have also shown that designing technology-supported learning environments can help teachers detect learners' learning situations and make suitable teaching decisions (Ait Baha et al., 2024; Alsobeh & Woodward, 2023; Han & Xu, 2024). However, among the general-purpose AI chatbots (e.g., ChatGPT, Gemini, Claude), many EAICs have been developed for academia-focused purposes (Fersch et al., 2023; Holderried et al., 2024; C.-C. Liu et al., 2022; Sharma et al., 2021). This is especially important when learners use EAIC to study. This paper uses the term *EAICs* to describe a growing type of AI chatbot that combines conversation with educational features, which is becoming increasingly important in digital learning. These chatbots help to create AI-assisted learning environments that focus on transforming learning experiences.

Regarding how the priming modality affects learners' experiences, students in clinical encounters were tested with different priming modalities to determine how they could find correct information about their patients (Stuart et al., 2019). Using the same psychological method, priming in human-computer interaction (HCI) has been studied to determine the influence of priming on creativity and idea generation (Lewis et al., 2011). Therefore, this goal is to determine the influence of priming on learning, which refers to the knowledge gained and the correct recall rate of the learners' answers.

To test the priming modalities affecting the use of EAIC, Lee et al. (2023) studied the role of AI in education, especially in the role of multimodality such as auditory, visual, kinesthetic, and linguistic modes. The adaptation of using different types of media in education has changed drastically, and with the integration of AI, we want to investigate the results of priming modalities in an AI-assisted learning environment, measured by the EAIC usability data.

In this study, Cipherbot was used as the main EAIC. Cipherbot, which was developed in 2022, is a chatbot tailored specifically for educators and learners. Cipherbot was set to only respond to queries based on the pre-uploaded materials from educators and could be designed so that learners could track citations, with recommended and follow-up questions based on specific topics. Our goal was to test the impact of priming modality on learners in a specific AI-assisted learning environment, focusing on the following research questions (RQs):

- *RQ1: How do modalities (visual and audio) influence learners' ability to recall knowledge?*
- *RQ2: How do learners' experiences (of the AI chatbot and task) influence learning experiences between modalities (visual and audio)?*
- *RQ3: How do materials modalities (visual and audio) influence task completion?*
- *RQ4: How does the order of learning contents influence learning behaviors?*

To address these RQs, we analyzed the learners' evaluations in each learning modality in terms of how they felt about the learning materials, the level of usability from their perspective about Cipherbot, and the overall learning experiences. Subsequently, we combined each learner's preference after experiencing both modalities and explained why they would prefer to choose the mode of learning together with the AI-enhanced learning environment.

The remainder of this paper is structured as follows: a review of the relevant literature on AI chatbots in education, the role of priming modalities in learning, and applicable learning theories, followed by an analysis of learner experiences in this study, and concluding with recommendations for integrating effective priming strategies into educational design.

2. Literature review

2.1. AI chatbots in learning

The continuous development of AI chatbots has created new opportunities to enhance and reshape the learning process (Robayo-Pinzon et al., 2024; S. Wang et al., 2025; Xue et al., 2025). According to

Labadze et al. (2023), AI chatbots can provide immediate support by answering questions, offering explanations, and providing additional resources. AI chatbots have significant benefits in education, for instance, reducing problems related to low teacher-student ratios and improving the overall learning experience. A low teacher-student ratio restricts and makes it difficult for students to receive immediate and interactive help (Y. Chen et al., 2023). Educational chatbots are emerging as solutions to these issues. Yin et al. (2024) mention that chatbots have shown promise in providing instructional support.

Additional benefits of using AI chatbots include, but are not limited to, learning basic content in a responsive, interactive, and confidential way (Y. Chen et al., 2023; Yin et al., 2024), immediate access and response (Belda-Medina & Kokošková, 2023; Labadze et al., 2023), individualized learning experiences, and addressing the diverse needs of students (Labadze et al., 2023; Yin et al., 2024). Homework and study assistance, personalized learning experience, and the development of various skills are also noticeable advantages (Jeon, 2023; Kim et al., 2019; Labadze et al., 2023). Moreover, AI chatbots support engagement and motivation through interactive features and real-time feedback, further enriching the learning experience (Belda-Medina & Kokošková, 2023; Fryer et al., 2017).

Chatbots are used in many fields, including academia and business. In business, for instance, they facilitate customer service by answering frequently asked questions and assisting in product choice, which helps cut expenditure (Feine et al., 2020). In other fields, such as healthcare, they help with appointment booking and basic advice, thus simplifying administrative burdens (Lucas et al., 2024). Chatbots are also employed in the programming field to generate code and provide detailed code explanations (Savelka et al., 2023). Chatbots are used in online education platforms to share course materials, conduct evaluations, and offer group work with timely and personalized assistance (Yin et al., 2024). The cases of their usage suggest that chatbots are effective in addressing problems in many areas.

Although chatbots, particularly in the pedagogical field, offer promising advantages, empirical research on the effectiveness of information systems in education remains scarce (Y. Chen et al., 2023). Further research is needed to understand students' learning processes when using these pedagogical AI chatbots, audio and video modality effects, and their impact on student perception and engagement behaviors. This gap is particularly relevant to the current study, as priming modalities (visual and audio) may influence how students perceive and interact with AI chatbots, such as Cipherbot, directly informing interaction quality and learning effectiveness.

2.2. Priming in learning

Priming is a key concept of cognitive psychology. This refers to the influence of a previously encountered object on future responses to similar objects (A. Wang et al., 2020; Wu et al., 2025). In the educational context, priming refers to any intervention to prepare learners for an educational experience or task (Stuart et al., 2019).

Priming can be used as a pedagogical technique that may improve learning outcomes by preparing students to interact with new content. *Visual Priming* occurs when the classification response is faster to a stimulus that is visually identical to a previous stimulus than to one that is identical only in name (Kroll & Schepeler, 1985). According to Xie et al. (2019), adding cues to parts of an image helps guide the learner's attention by emphasizing specific visual elements and directing focus to highlighted messages. In simple terms, if a part of an image is marked red, it will direct the reader's eye to the crucial section and reduce the search time. Although visual cues can help draw learners' attention, they may not always ensure better understanding. Many studies that used only visual or auditory cues failed to find a positive effect of cueing on learning outcomes (Xie et al., 2019). This suggests that guiding attention with single-modality cues does not guarantee a better integration of relevant words and pictures, especially when presenting complex multimedia materials. Presenting text auditorily instead of visually may improve learning when paired with visual elements, such as graphs, diagrams, or animations (Ginns, 2005).

Despite extensive research on the phenomenon of priming in cognitive psychology, its implications for education remain under-examined. There is a lack of studies on modality-specific priming, including visual and auditory inputs, in the context of Educational AI Experiences (EAIEs) in which EAICs

manifest. This literature study discusses three fundamental theories: the Cognitive Theory of Multimedia Learning (CTML) (Jiang et al., 2017; Mayer, 2024), Dual Coding Theory (DCT) (J. M. Clark & Paivio, 1991), and Cognitive Load Theory (CLT) (Van Merriënboer & Sweller, 2010), as well as the concept of modality effect (Ginns, 2005). These theories are selected for their emphasis on multimodal processing, which is central to understanding how visual and audio priming influences learning outcomes in AI-assisted environments, directly supporting the study's focus on interaction quality, engagement, and learning effectiveness.

2.3. Learning theories

2.3.1. Cognitive theory of multimedia learning

CTML offers a strong foundation for examining how modality priming affects learning outcomes and student engagement in AI-assisted settings. CTML is widely recognized in the field of educational psychology (Jiang et al., 2017; Mayer, 2024). Learning deepens when text pairs with graphics instead of text alone, as humans handle data via two distinct channels: one for visuals and the other for sound and words (Cavanagh & Kiersch, 2022; Yue et al., 2013). Each channel manages only a small amount at once, and meaningful learning depends on the students actively handling the material (Mayer, 2024).

Active processing is one of the key concepts of CTML, which states that meaningful learning requires learners to actively engage with content. As Mayer (2024) explains, meaningful learning occurs when a student actively processes lesson content by choosing pertinent details, arranging new information into a clear mental framework, and linking it to prior knowledge stored in long-term memory.

This study employed CTML as the primary theoretical framework for addressing *RQ1 (How do modalities (visual and audio) influence learners' ability to recall knowledge?)* and partially informing *RQ4 (How does the order of learning contents influence learning behaviors?)* CTML was selected because it provides the most comprehensive framework for predicting when and why different modality presentations (visual vs. audio) should differentially support learning outcomes, making it directly applicable to our investigation of visual (in text and image format) versus audio (in podcast format) priming materials in EAIC environments.

2.3.2. Dual-Coding theory

According to Clark and Paivio (1991) DCT, there are two separate but related cognitive systems: the verbal system, which processes language, and the non-verbal system, which processes images and visual content. Meaningful learning occurs when students engage in both systems simultaneously, enabling the formation of dual memory traces by offering several methods for retrieving information. These dual verbal and non-verbal representations improve memory and understanding (J. M. Clark & Paivio, 1991; Paivio, 2014). The DCT views cognition as a split between a verbal system for language and a non-verbal system for images and objects, with recall improving when both codes merge, as people in free-recall tasks quietly name items, creating dual memory marks (Paivio, 2014). This study employed DCT to address *RQ1 (How do modalities (visual and audio) influence learners' ability to recall knowledge)* by examining how visual (text and image) versus audio (podcast) priming materials support dual-code formation during EAIC interactions. DCT was selected because it provides specific predictions about how verbal and imaginal encoding systems differentially support memory and learning and is directly applicable to our comparison of text and image materials (engaging both systems) and audio materials (primarily engaging the verbal system).

2.3.3. Cognitive load theory

CLT is a key theory in educational psychology with significant applications in instructional design and learning sciences (Van Merriënboer & Sweller, 2010). CLT is based on a cognitive architecture that consists of limited working memory, with partly independent processing units for visual and audio information, which interacts with unlimited long-term memory (Kirschner, 2002). CLT shapes educational psychology, proposing that reducing excess mental health prevents overload and frees resources for meaningful learning (Skulmowski & Xu, 2022; Van Merriënboer & Sweller, 2010). The integration of CLT into modern technologies is a subject of ongoing studies, and its outcome is yet to be seen.

Skulmoski et al. (2022) proposed weighing how design factors alter cognitive load, with interactivity offering both gains and drawbacks. Matching the load to the learning tasks and evaluation methods reflects the design impact. This study employed CLT to address *RQ2 (How do learners' experiences (of the AI chatbot and task) influence learning experiences between modalities (visual and audio)?)* and *RQ3 (How do materials modalities (visual and audio) influence task completion?)*. CLT was selected because its constructs, intrinsic, extraneous, and germane cognitive load, provide a framework for understanding how learner characteristics and presentation sequences influence learning processes, directly applicable to our investigation of AI chatbot expertise and task experience as moderators.

2.3.4. Modality effect

The modality effect, based on CLT, emphasizes that learning effectiveness is improved when visual and auditory channels are employed simultaneously. This theory illustrates that learning outcomes are enhanced when graphical elements are visually shown while related textual information is communicated through auditory means (Ginns, 2005).

Previous researchers have mixed opinions regarding the application of the modality effect and its use in modern technologies. Research has shown that the modality effect can fade or shift in self-paced settings (Ginns, 2005; Tabbers et al., 2004). Ginns (2005) notes that modern tools do not always guarantee strong learning results. According to Mayer (2009), the Modality Principle favors pairing images with spoken words over printed text for deeper understanding. Applying the modality principle as a theory-based approach to support multimedia learning improves the outcomes for many learners, subjects, and formats (Moreno, 2006). Liu et al. (2022) stated that spoken text is more effective than written text in animation-based learning. This study employed the modality effect to generate alternative predictions for *RQ1 (How do modalities (visual and audio) influence learners' ability to recall knowledge?)* that potentially contradict the expectations of CTML.

In summary, CTML theory (Mayer, 2024) states that humans possess dual channels but have limited capacity. Active processing must be performed in each channel for deeper learning. This theory is relevant to the present study on effective learning, which depends on modality. DCT theory (J. M. Clark & Paivio, 1991) suggests that priming from audio, visual, or both can lead to better memory retrieval. CLT theory (Van Merriënboer & Sweller, 2010) explains how priming in different modalities can restrict overloading and aid decisions. Guided by these foundation theories, the present study investigated whether different types of priming, or priming modalities, visual or audio in formats, affect students' perceptions, engagement behavior, and learning outcomes in EAIE using Cipherbot. Although CTML (Mayer, 2024), DCT (J. M. Clark & Paivio, 1991), and CLT (Van Merriënboer & Sweller, 2010) have formed a basis for understanding how visual and audio priming shape the learning processes, their application to Educational AI chatbots, such as Cipherbot, is unexplored, making the current study crucial.

3. Methods

3.1. Research design

This study employed a within-subject quasi-experimental design to investigate the effects of priming modalities on learning experiences in an AI-assisted environment. The quasi-experimental approach was chosen because all participants experienced both experimental conditions (visual and audio modalities) instead of being randomly assigned to different control and treatment groups. This design follows a clear structure of pre-survey and post survey structure: learners started with completing a pre-survey to pre-assess their backgrounds and prior knowledge then engaged in two learning sessions (one with visual learning material and one with audio learning material) with counterbalance presentation order and completed learning experience survey after each session (Figure 1). This counterbalancing across four sequences (Table 1) controlled for potential order and carryover effects. The design increased statistical power by allowing each participant to serve as their own control with reducing individual difference confounds.

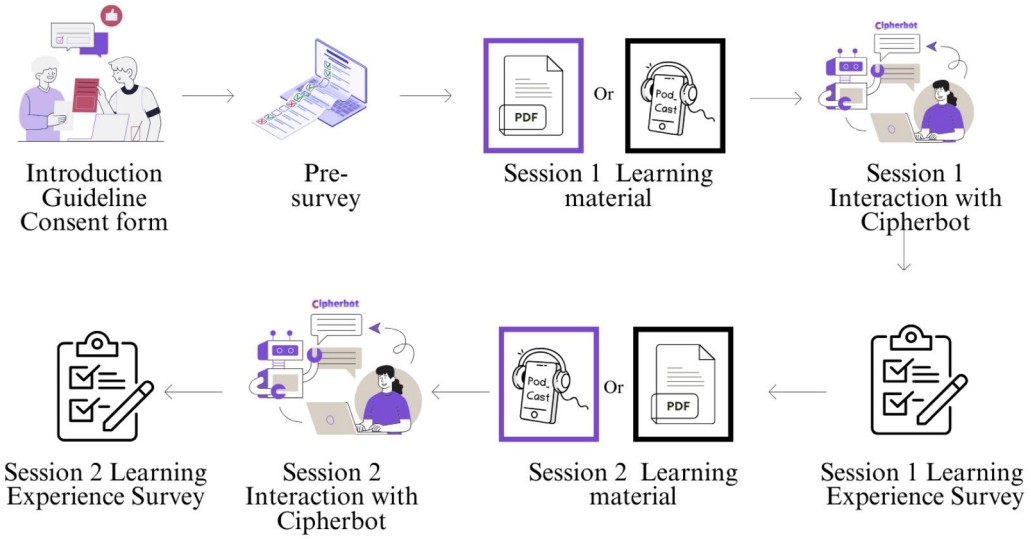


Figure 1. Visualization of the study flow. In the learning material, if session 1 were visual, then in session 2 would be audio and vice versa.

Table 1. The learner sequences' explanation.

Learning topics\types of materials	Visual – V (PDF file with text and image)	Audio – A (podcast)
Morpho butterfly – B	VB – Visual and butterfly	AB – Audio and butterfly
Poison dart frog – F	VF – Visual and frog	AF – Audio and frog

Each learner experiences 1 visual material, 1 audio material, 1 frog material, and 1 butterfly material.

3.2. Procedure

This study investigated the impacts of different priming modalities in an AI-assisted learning environment. The learning context of this study was to protect wild species (see Supplementary Appendix 1 for a detailed description). There were two learning sessions (Figure 1). In each learning session, learners were given learning material and five minutes to interact with Cipherbot to complete a learning task. Before the first session, they completed a pre-survey; after each session, they completed a learning experience survey. Overall, 67 learners volunteered to participate in this study. All the learners signed a consent form to share their data for research purposes. The learners were not financially compensated but were given a small gift for their participation.

Cipherbot is an EAIC that responds to queries based on pre-uploaded materials (Figure 2(a)) (Jung et al., 2025). Cipherbot helped educators create an AI-enhanced environment controlled only by class-code access. The learners were given class code access for the specific class they attended (Salminen et al., 2024). Educators can design classes to fit each level of learners, background, language, preferences, and the nature of the course (Figure 2(b)). Learners can use Cipherbot for multiple educational purposes: "Chat" for knowledge exchange, "Learn" for problem solving, and "Mentor" to enhance work quality. In this study, due to the purpose of the study, which is to find a solution for the learning task, we asked learners to use the "Learn" feature to interact with Cipherbot. Learners were randomly assigned different sequence combinations (Table 1).

There are four different types of learning materials (Table 1), which are later discussed as four sequences: Visual Butterfly (VB), Audio Butterfly (AB), Visual Frog (VF), and Audio Frog (AF). Visual materials included text and image, presented in a consistent format: title, explanations, image, and research of the species (see Figure 3(a) for VF and Figure 3(b) for VB) with similar lengths. The audio materials were created using Speeches features on Cipherbot (Figure 4(a)) in the same format: Podcast, 1 Female, 1 Male, 3–5 min, and in English. During the audio condition, learners had access to both the podcast audio and an accompanying transcript (see Figure 4(b) for AF and Figure 4(c) for AB), allowing them to follow along and review content as needed. The source material used for generation was the visual material of the species (e.g., AB generated from VB, AF generated from VF), allowing equivalent contents across both modalities.

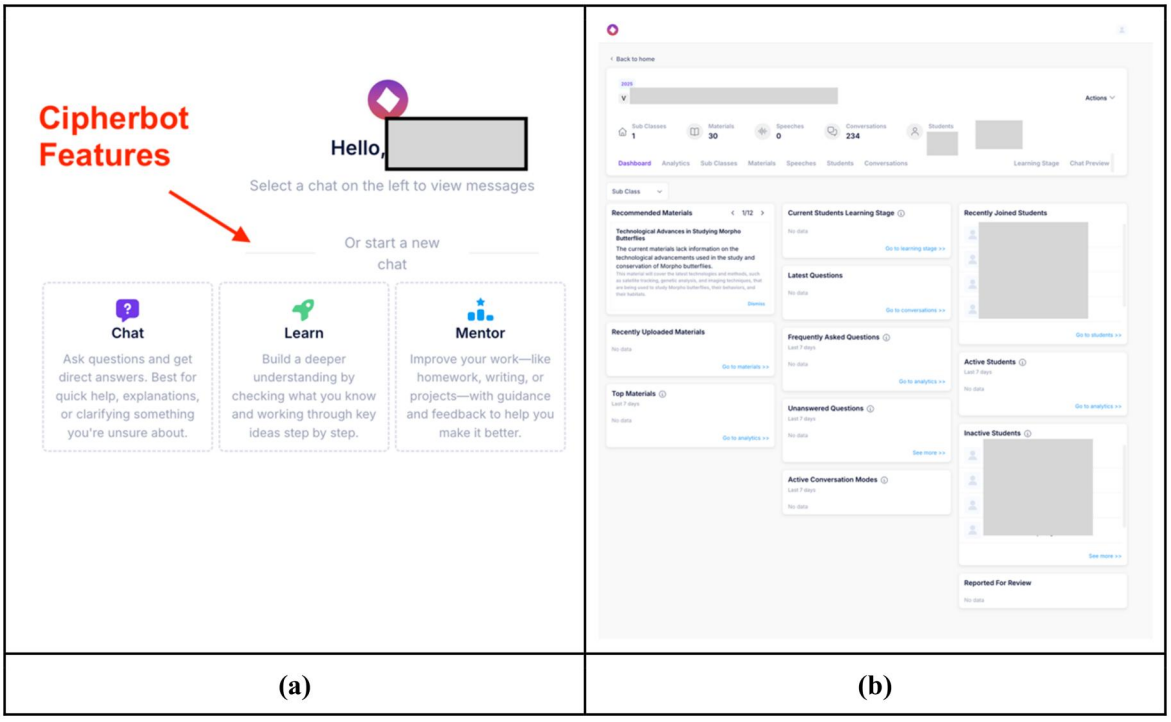


Figure 2. (a) Cipherbot interface with three learning features: Chat, learn, and mentor in the learner account (screenshotted in January 2025, before the study) and (b) cipherbot dashboard, which is managed by the teacher account (screenshotted in May 2025, after the study). (We encourage the reader to zoom in to increase text legibility.).

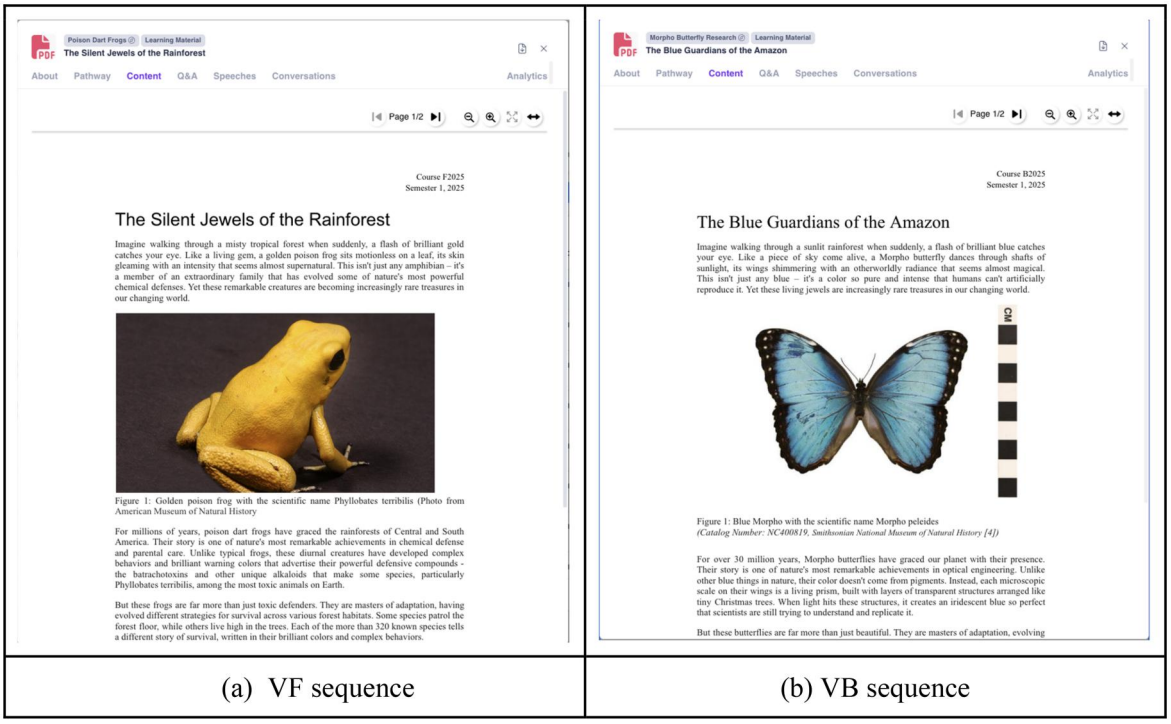
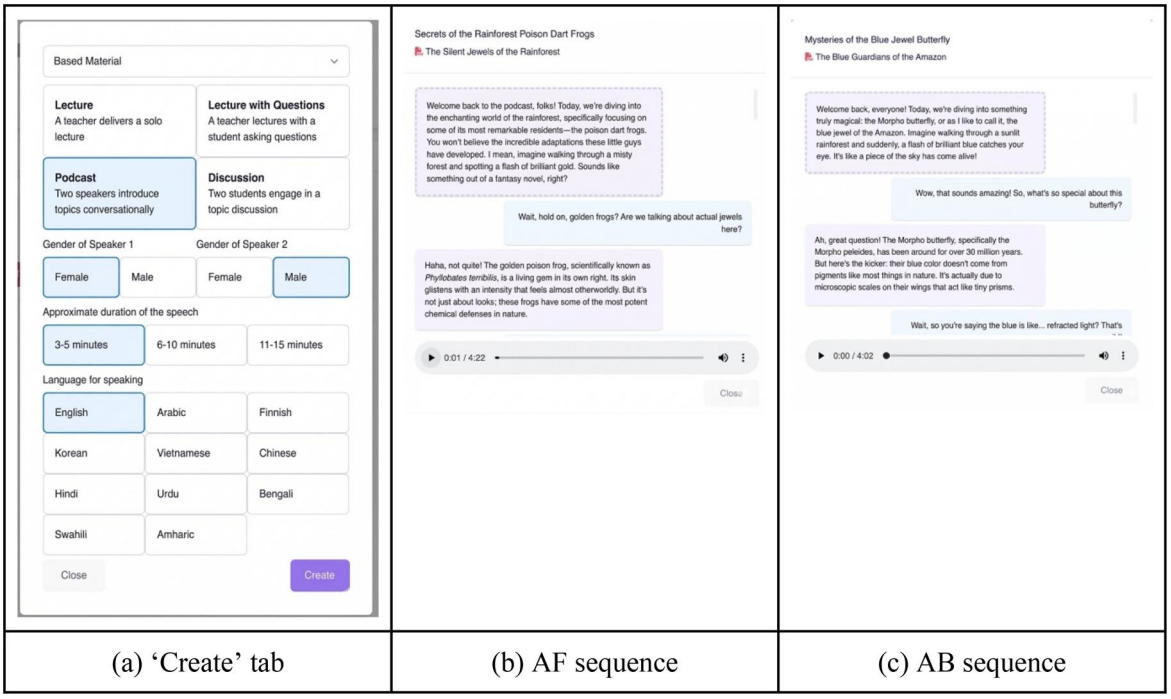


Figure 3. Visual interface from the learners' view (student account): (a) VF sequence and (b) VB sequence. Further details on the materials can be found in Supplementary Appendix 1.

3.3. Participants

The sample consisted of 67 participants: 40 males (59.7%) and 27 females (40.3%). The learners were aged between 18 and 70 years ($M = 36.72$, Mdn = 36.00, SD = 13.17). Their education levels were



(a) 'Create' tab | (b) AF sequence | (c) AB sequence

Figure 4. Speech feature of cipherbot, which lets both the teacher and student choose a specific material to create an audio. (a) "create" tab, which lets users choose what characters they want in the audio, and this is the option we chose in this study: Podcast, 1 Female, 1 Male, 3–5 min, and in English. (b) The AF sequence in play for learners. (c) The AB sequence in play for learners.

doctoral ($n = 30$, 44.8%), master's ($n = 18$, 26.9%), bachelor's ($n = 15$, 22.4%), and high school ($n = 4$, 6.0%). Learners represented 38 nationalities spanning six continents, with the largest groups being Pakistan ($n = 6$, 9.0%), Jordan ($n = 5$, 7.5%), and India ($n = 5$, 7.5%).

The participants' professional backgrounds showed the largest groups coming from computing and technology fields, including Computer Engineering/Science/Software ($n = 12$, 17.9%), AI/Machine Learning ($n = 6$, 9.0%), Cybersecurity ($n = 2$, 3.0%), and Data Science/Mining ($n = 2$, 3.0%). Education and research professionals ($n = 10$, 14.9%) and engineering fields ($n = 6$, 9.0%) formed the next largest group. Other backgrounds included Business & Communications ($n = 6$, 9.0%), Biological Sciences ($n = 4$, 6.0%), Physical and Environmental Sciences ($n = 4$, 6.0%), Economics ($n = 2$, 3.0%), and individual representatives ($n = 1$, 1.5%) of Healthcare, Design, Project Management, and Translation.

Before the study, the learners reported their experiences with AI and Cipherbot. Learners reported using AI primarily for work-related tasks ($n = 56$, 83.6%), followed by information searches ($n = 45$, 67.2%), and learning/education ($n = 40$, 59.7%). Entertainment was reported least frequently ($n = 9$, 13.4%).

3.4. Data collection

The study was conducted over a week in two locations (a research institute facility and a library), comprising five parallel sessions with eight moderators. All moderators followed a standardized script to ensure consistency in instructions, prompts, and interactions across sessions. Moderators were trained prior to data collection to minimize variability in delivery. The datasets we collected in this study included (1) pre-survey, (2) experience survey, (3) Cipherbot chatlogs, (4) audio recordings of the study, and (5) answers to the learning tasks. Both surveys were collected using a 5-point Likert scale and included open-ended questions for qualitative analysis (see survey items in Supplementary Appendix 2).

First, the pre-survey was conducted after stating the purpose of the study and before the start of the first learning session. The pre-survey included background information, AI usage, Cipherbot experiences, pre-knowledge regarding the learning tasks, and knowledge about wild species (Morpho Butterfly

and Poison Dart Frog). We chose the Morpho Butterfly and Poison Dart Frog because we would like to choose some topics that are not very popular, and the participants would actually need to learn something new from the learning sessions.

Second, a learning experience survey was conducted twice after each learning session. The survey included learners' evaluation of the learning material, Cipherbot usability, and learning tasks. In the second session's survey collection, we also asked the learners about their learning material preferences and their knowledge of wild species after the sessions.

Third, Cipherbot chatlogs are question-answer pairs of interactions between learners and Cipherbot ($n = 807$, Visual = 402, Audio = 405). This dataset shows the contents of their interactions, question themes, and the time of each interaction. Fourth, the audio recordings included learners' think-aloud data ($n = 67$). Fifth, answers to learning tasks included mainly the social media posts that learners were asked to create to trigger their networks to take action to help the wild species, two multiple-choice questions, and an open-ended question recalling the knowledge they gained from the learning materials.

3.5. Measures

This study measured four main forms of data: background variables, experimental conditions, recall rate, behavior engagement variables, and learners' preferences (Table 2).

3.6. Data analysis

This study employed a mixed-methods analytical approach to examine learners' experiences with priming modalities in an AI-assisted learning environment. The dataset included inputs from 67 learners collected through multiple instruments: a pre-survey, experience surveys (administered twice, once after each modality condition), behavioral data extracted from recorded think-aloud interactions, and chat logs with Cipherbot across counterbalanced sequence conditions. The data preparation involved removing duplicate pilot entries and merging all datasets into an unified dataset with aligned participant IDs to enable cross-referencing across data sources. Audio recordings from think-aloud sessions were transcribed using Microsoft Word online, which automatically generated timestamps and speaker labels. These transcripts were then computationally processed using Python in Google Colab to transform the word-formatted transcripts into a structured Excel format and filter conversations specifically occurring during Cipherbot interactions for subsequent behavioral coding and analysis.

Quantitative analyses included paired-samples t-tests to compare recall performance between modalities (RQ1), Pearson correlations to examine relationships between prior experience (AI chatbot expertise and task familiarity) and behavioral engagement metrics (RQ2), chi-square tests, Wilcoxon signed-rank tests to analyze task completion patterns and overlap between chatbot interactions and social media posts (RQ3), and Mann-Whitney U tests to investigate order effects across counterbalanced sequences (RQ4), with statistical significance set at $\alpha = .05$. Behavioral metrics extracted from the transcripts and chat logs included chatbot interaction count, moderator talk time, learner talk time, recording time/duration, utterance counts, and social media post length.

Qualitative analysis involved thematic coding of think-aloud transcripts and open-ended survey responses to identify patterns in learners' experiences, strategic approaches, and perceptions of the two modalities, with representative participant quotes extracted to illustrate key findings and triangulate them with quantitative results.

4. Findings

4.1. RQ1: How do modalities influence learners' ability to recall knowledge?

Prior to analyzing the modalities influencing recall knowledge, we looked into the learners' preferences for which modality they preferred after both learning sessions to understand the personal preference in case it could potentially influence their behaviors. The results showed that Visual was the preferred

Table 2. Explanations and sources of data for each group of variables that were used for analysis.

Group of variables	Explanation	Measure	Source of data
Background	Self-reported data from the pre-survey: age, gender, educational level, nationality, AI chatbot expertise, AI chatbot usage patterns, and prior task experience with social media post creation.	Likert 1–5 Open-ended	Pre-survey
Experim5ntal conditions	• Systematically manipulated factors: • Modality (Audio and Visual) • Learning content/task (Butterfly and Frog topics) • Presentation order/sequence (AB, AF, VF, VB) • Order of sessions (First and Second sessions)		Experience survey
Results	• Recall rate: the percentage of information correctly remembered after each session, measured through recall questions, which include two multiple-choice questions. The open-ended recall question was collected for supplementary qualitative insights but was not included in the quantitative recall rate analysis. The audio transcript of the recall part after the learning session was also taken into account when further investigating the results • Self-reported task completion (writing social media post) • Overlap between social media post and Cipherbot chatlog • Social media post topics and content types	Percentage Transcript Self-reported	• Answers to the learning tasks • Experience survey • Cipherbot chatlog
Behavioral engagement	• Chatbot interaction: chatbot interaction count (total number of messages sent to Cipherbot, counted from the chatbot log), record count (frequency of recordings) • Learner engagement: learner talk time (total duration in minutes/seconds, calculated from transcribed audio recording), learner utterance count (number of speaking turns) • Moderator interaction: moderator talk time (total duration in minutes/seconds, calculated from transcribed audio recording), moderator utterance count (number of speaking turns) • Session transcript: total length of session duration (minutes/seconds), total utterance count (combined speaking turns), recording time/duration (total interaction time per session), extracted from audio file metadata • Task output quality: social media post length (character/word count of created posts) created by learners, extracted from task output	Count Record	Audio recordings of the study
Learners' preference	Modality preference (Audio vs. visual learning materials).	Multiple choices	Experience survey

format for 37 learners (55.2%), whereas Audio was preferred by 30 learners (44.8%). This distribution suggests a relatively balanced preference between the two formats, with the slight majority favoring traditional format materials over audio content. The near-even split in preferences (55.2 vs. 44.8%) indicates that both formats serve an important role in meeting diverse user needs and learning styles, with text maintaining a modest advantage of approximately 10 percentage points over audio.

A paired-sample t-test was conducted to compare the recall rates between the audio and visual modalities. The results revealed a significant main effect of modality on recall performance ($t(66) = 2.34$, $p = 0.022$, Cohen's $d = 0.41$, 95% CI [0.06, 0.76]). Learners demonstrated higher recall rates in the visual condition ($M = 0.82$, SD = 0.27) than in the audio condition ($M = 0.68$, SD = 0.39) with a mean difference of 0.14, 95% CI [0.02, 0.26]. This represents a 20.6% improvement in recall for visual presentations with a medium effect size according to Cohen's (1988) conventions (Figure 5).

To understand the mechanisms underlying the observed recall advantage, the learners' transcripts were reviewed for qualitative themes. After learning from the visual modality material, learners consistently referenced their ability to control pacing and review the content. For example, one participant explicitly noted the review capability ("You don't only go back and look," P01, Participant), while another sought to access the material again ("Can I go back to look at the learning material?," S06, Participant). Learners also reported positive engagement with materials ("Yeah. The material from the

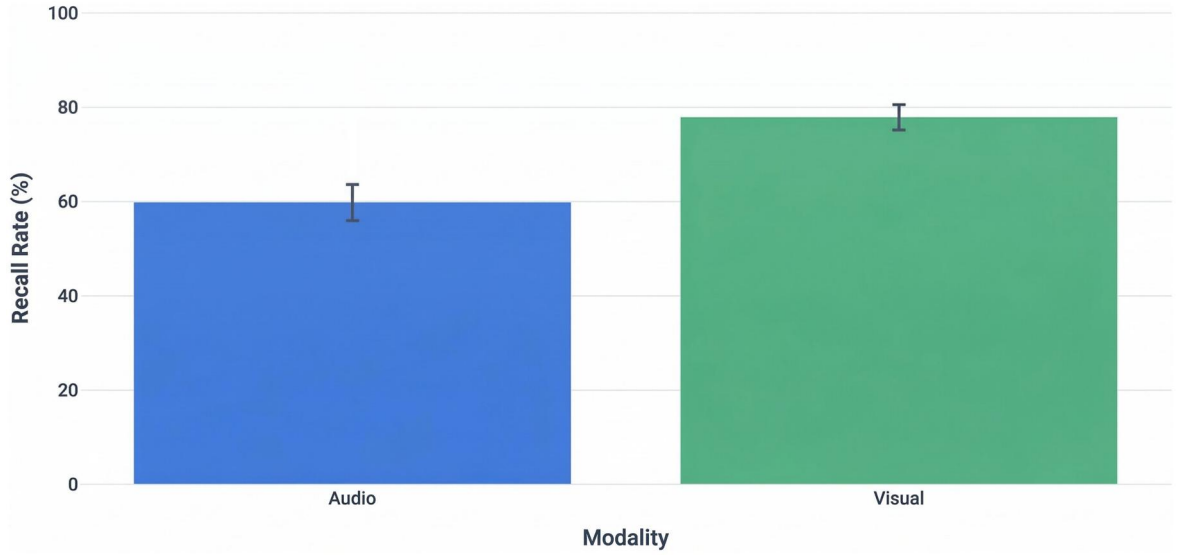


Figure 5. Mean recall rates (%) with standard error bars. Visual modality ($M = 82\%$) shows significantly higher recall than audio ($M = 68\%$), $t(66) = 2.34$, $p = 0.022$, $d = 0.41$.

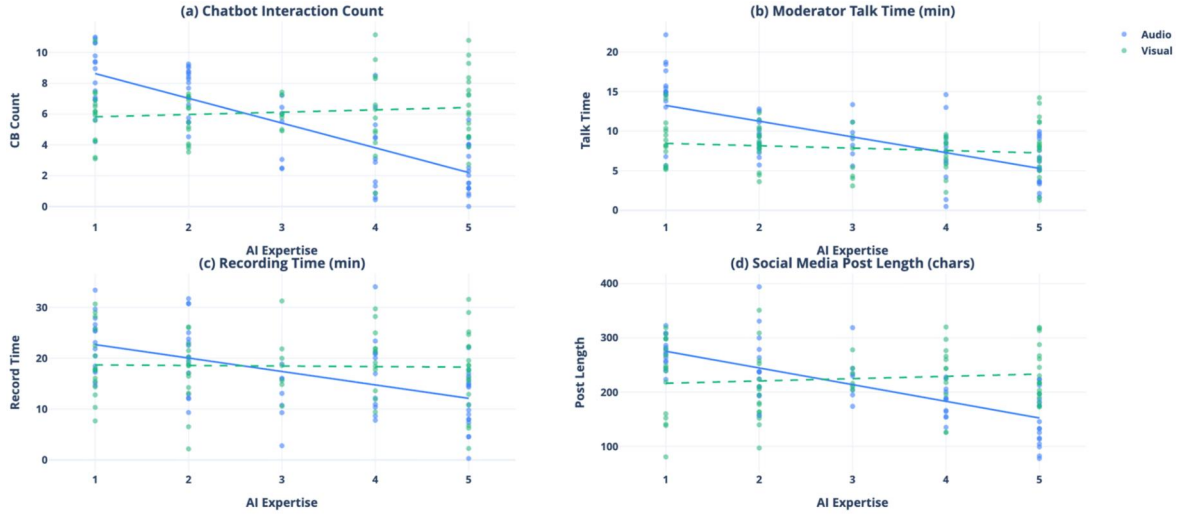


Figure 6. AI expertise influenced by modalities across different interactions. (a) with chatbot interaction count, which refers to the number of messages sent to cipherbot during each session of each modality, (b) with moderator talk time, which refers to the duration that the moderators need to spend to interact, encourage or explain to the learner during the interaction with cipherbot, (c) with total recording time, which refers to the total interaction time with cipherbot per session, (d) with social media post length, which refer to the length of post created by learner in each session.

article or book was very enjoyable reading. I enjoyed the reading. It was inspiring. The material helped me to ask," S08, Participant).

In contrast, after learning from the audio modality material, learners acknowledged the memory-dependent constraints imposed by the temporal nature of the audio presentation. Learners recognized that they had to rely on recall ("You need to recall what you remember," P01, Participant). Another learner expressed difficulty in retention ("Okay. I can't remember anything about this race in captivity, so I could do this from memory. I can't go back and ask this question," P42, Participant).

4.2. RQ2: How do learners' experiences (of the AI chatbot and task) influence learning experiences between modalities (visual and audio)?

4.2.1. Previous experiences of using EAIC

As an exploratory measure, learners were asked to give their AI chatbot background knowledge on how their AI chatbot expertise was and what they mainly used the AI chatbot for.

To examine whether AI expertise moderated the relationship between modality and learning experience, we conducted Pearson correlation analyses between AI expertise levels (Likert 5) and four key behavioral metrics: chatbot interaction count (Figure 6(a)), moderator talk time (Figure 6(b)), recording time (Figure 6(c)), and task completion (Figure 6(d)) in the social media post length data. These analyses were exploratory in nature and were used to explore and understand the main findings of the RQ.

In the audio modality, learners with higher AI expertise engaged in significantly fewer chatbot interactions ($r = -0.345$, $p = 0.004$), spent less time in moderator discussions ($r = -0.377$, $p = 0.002$), recorded for shorter durations ($r = -0.267$, $p = 0.030$), and produced briefer social media posts ($r = -0.327$, $p = 0.007$). These correlations ranged from small to medium in magnitude, explaining 7.1 to 14.2% of the variance in engagement behaviors. In contrast, low AI expertise was less visible. In contrast, learners with low AI expertise showed considerably fewer significant results, with most engagement metrics failing to demonstrate statistically significant associations with the expertise level.

In contrast, the visual modality showed no significant correlations between AI expertise and any behavioral metric (all $p > 0.05$, all $R^2 < 0.03$), with negligible strength. This pattern suggests that visual presentations may provide more equitable learning experiences across different expertise levels.

Overall, the differential impact of AI expertise across modalities was further confirmed through Pearson correlation analyses of the difference in scores (audio minus visual). AI expertise negatively predicted the difference in chatbot interactions between audio and visual conditions, with moderate strength ($r = -0.373$, $p = 0.002$, $R^2 = 0.139$), indicating that low-expertise learners relied substantially more on chatbot support in audio (M_diff = 0.42), whereas high-expertise learners showed the opposite pattern (M_diff = −0.58). Similar patterns emerged for moderator talk time differences, with moderate strength ($r = -0.377$, $p = 0.002$, $R^2 = 0.142$).

To further understand the influences of learners' AI experiences, we further investigated the audio transcripts to determine the different experiences of learners after learning from audio and visual. In the audio modality, experienced learners employed efficient strategies ("I interact quite rapidly and move as quickly as possible," S133, Participant), while novices struggled with the transient format ("If I was thinking about something that they were saying, then I might miss a little chunk of the audio because I couldn't pause it or rewind," S41, Participant). By contrast, the visual format provided equitable support. One learner explained, "When I read, I like to read a part and then I go back, like if I didn't capture the information and go back and read, but without it, just like it's fleeting" (S09, Participant).

4.2.2. Task experiences

We then further investigated how prior task experience (in terms of social media post-creation knowledge) influenced learning experiences across modalities. In the audio condition, task experience showed significant negative correlations with chatbot interaction count ($r = -0.308$, $p = 0.011$) (Figure 7(a)), moderator talk time ($r = -0.284$, $p = 0.020$) (Figure 7(b)), and social media post length ($r = -0.308$, $p = 0.011$) (Figure 7(d)) all representing medium effects, except the recording time (Figure 7(c)). Learners with low task experience (ratings 1–2) averaged 7.27 chatbot interactions compared to 3.67 for high-experience learners (ratings 4–5), a 49.5% reduction.

The visual condition demonstrated a weaker pattern, with task experience correlating significantly only with social media post length ($r = -0.255$, $p = 0.039$). Other engagement metrics showed no significant relationship with task experience in the visual modality.

Overall, the analysis of modality differences revealed that task experience significantly predicted chatbot interaction differences ($r = -0.299$, $p = 0.014$), with low-experience learners using chatbot more in audio (+0.84 interactions), while high-experience learners reversed this pattern (−1.40 interactions), $t(43) = 2.73$, $p = 0.009$, $d = 0.80$. This crossover interaction suggests that modality preference shifts as a function of task familiarity.

To further investigate the reason behind learners' behavior based on their task experience level, we examined their conversations between audio and visual modalities. In audio, more experienced learners showed interest in how information needed to be validated before writing the post ("So this audio has a bunch of claims which don't make sense to me. So I'm actually gonna check them first. Okay. Just to

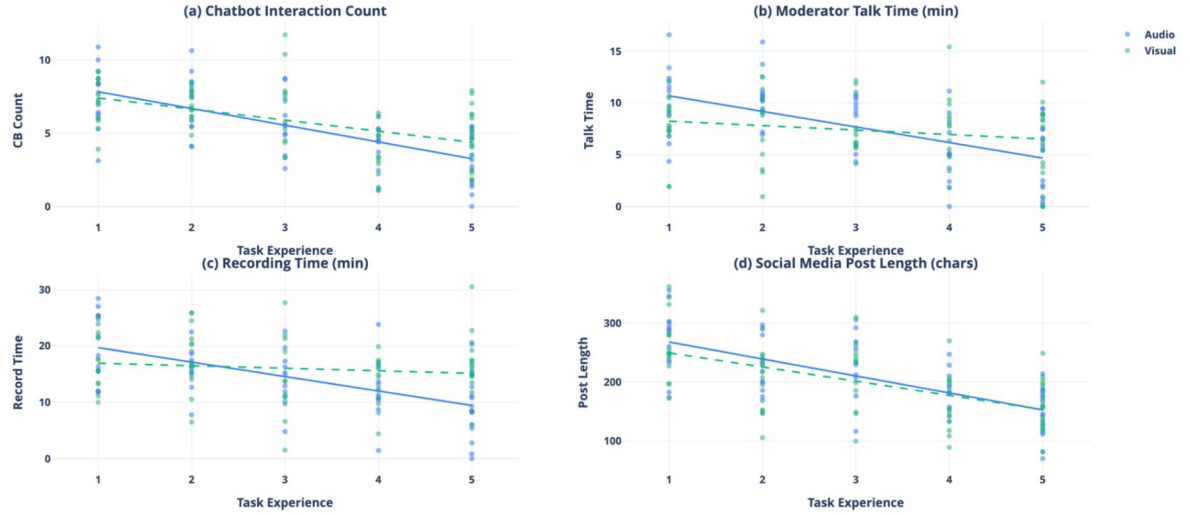


Figure 7. Learner's experiences with the task (creating a social media post) influenced by modalities across different interactions, (a) with chatbot interaction count, which refers to the number of messages sent to cipherbot during each session of each modality, (b) with moderator talk time, which refers to the duration that the moderators need to spend to interact, encourage or explain to the learner during the interaction with cipherbot, (c) with total recording time, which refers to the total interaction time with cipherbot per session, (d) with social media post length, which refer to the length of post created by learner in each session.

Table 3. Summary of self-reported data on how learners thought they wrote the post after each learning session.

Learners' process of writing the post	Audio count	Audio (%)	Visual count	Visual (%)
Wrote all of it by myself	25	37.3	20	29.9
Copied some of it from Cipherbot but manually edited the copied text	25	37.3	27	40.3
Copied all of it from Cipherbot	17	25.4	20	29.9
Total	67	100.0	67	100.0

make sure that if I'm sharing something, it has some basis in it," S12, Participant). However, less experienced learners expressed the need for explicit guidance on parameters, leading to more questions and, thus, higher interaction counts ("How can I know? I'm not using social media? I never write posts for social media. What are the criteria?," S11, Participant). In visual modality, learners with higher task experience wrote shorter social media posts ("Unless I'm starting with Cipherbot to learn something from scratch. That'll be, I think, a different case. But because I already learned the material, I went to Cipherbot, but with a different mindset. I have a task. I want to achieve the task. So I need to not really learn, but just immediately get information about certain things to get the task done," S03, Participant). Consequently, low-experience learners' difficulty with defining the post's structure and purpose requires multiple explicit interactions and revisions to produce the post, ("Okay, I don't like the structure. Why? Because I don't think that's the purpose of my post. So, the purpose of my post very clearly is that I need to share some information about the frogs because I want people, because my call to action is that people save these animals," S54, Participant).

4.3. RQ3: How do materials modalities (visual and audio) influence task completion?

To investigate how the learning material modalities influence task completion, we looked at the self-reported data of how learners thought they wrote the post *(How did you create the post?)* (Table 3). In the audio modality, learners demonstrated the highest rate of complete original content creation, with 37.3% ($n = 25$) reporting that they wrote all content by themselves. This proportion was equaled by learners who copied some content from Cipherbot but manually edited the copied text, which also accounted for 37.3% ($n = 25$) of the audio content. Direct copying from Cipherbot without modification constituted the smallest proportion (25.4%, $n = 17$). In terms of visual modality, the patterns differed significantly. The majority of the learners engaged in partial copying with manual editing, representing 40.3% ($n = 27$) of the visual content. The remaining content was nearly equally distributed

between completely original writing (29.9%, $n = 20$) and complete copying from Cipherbot (29.9%, $n = 20$). These findings indicate that, while both modalities incorporated substantial amounts of adapted content, the audio modality demonstrated greater reliance on original content creation, whereas the visual modality showed a stronger tendency toward editing externally sourced material rather than generating purely self-authored content.

After that, we conducted a series of analyses between the **questions that learners ask** about the EAIC after each modality and the final **social media post**. First, to test for overall differences between the modalities, a Wilcoxon signed-rank test was conducted. The analysis revealed no significant difference in the rate of overlap between the questions learners asked and social media post between Visual ($Mdn = 2$, $M = 1.81$) and audio ($Mdn = 2$, $M = 1.85$) modalities, $W = 255.00$, $p = 0.622$, with a negligible effect size ($r = 0.06$). Both modalities predominantly elicited "Little overlap" ratings, with 53.7% of the visual ratings and 67.2% of the audio ratings falling in this category.

Exact agreement between modalities was observed in 50.7% of the participants ($n = 34$). Among the 49.3% who showed different ratings ($n = 33$), there was no significant directional bias, with 60.6% rating audio higher and 39.4% rating visual higher ($p = 0.296$, sign test).

However, a chi-square test examining the symmetry of rating transitions revealed significant asymmetry in the pattern of changes between modalities, χ^2 (9, $N = 67$) = 21.89, $p = 0.009$. Examination of the contingency table showed that among participants who were initially found in Visual as "No overlap" ($n = 25$), 68% subsequently had Audio as "Little overlap," while only 32% maintained the "No overlap" for Audio (Figure 8). Conversely, among those who wrote posts in Visual as "Little overlap"

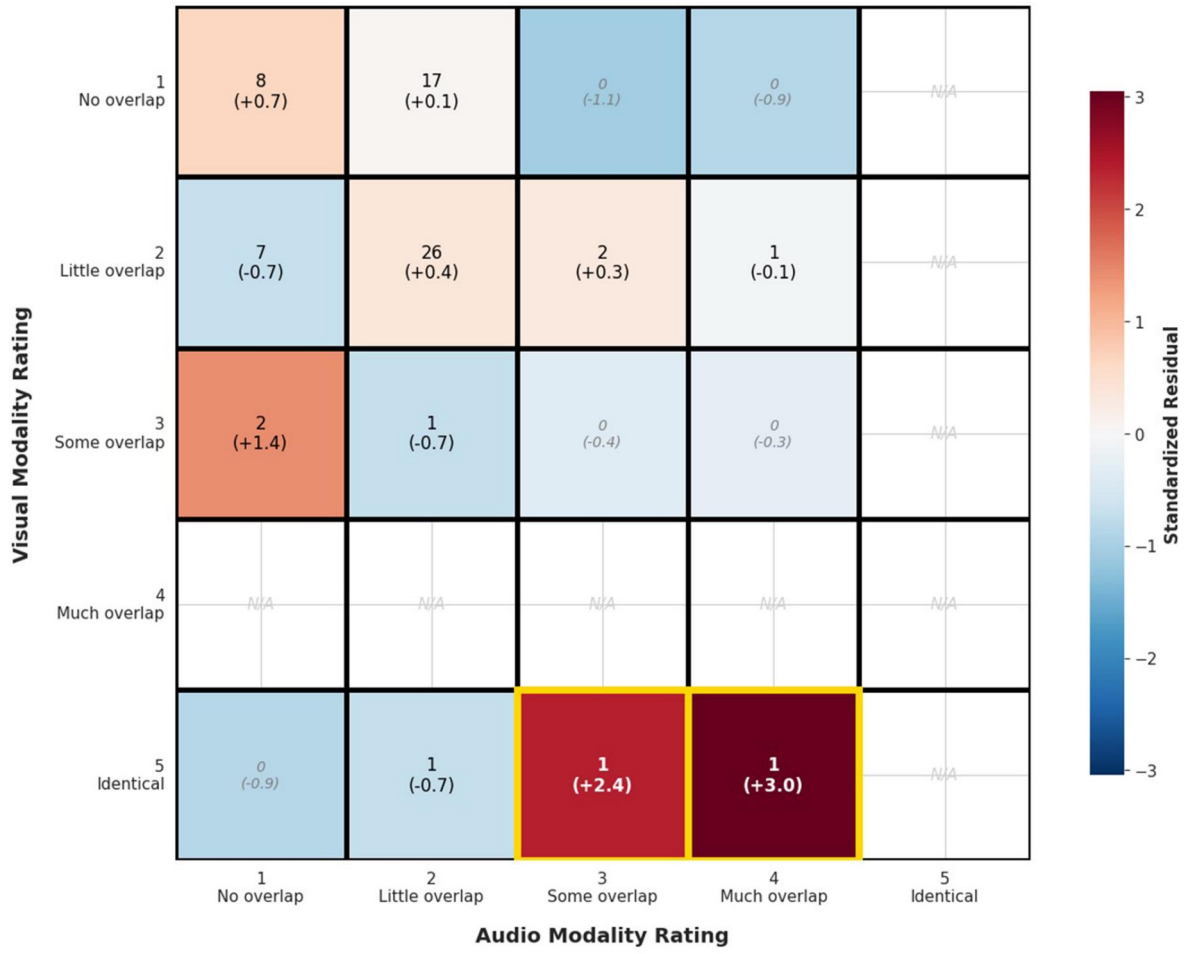


Figure 8. Heatmap association between visual and audio modality in the rate of overlap between questions learners asked and social media posts. *Note*: Values observed count with standardized residual in parentheses. Bold text and gold border indicate \residual\> 1.96 ($p < 0.05$). Blue = fewer than expected, red = more than expected, N/a = no data collected.

Table 4. Themes in terms of topics in social media posts written by learners after each learning session.

Topic group	Visual count ($n = 39$)	Visual (%)	Audio count ($n = 33$)	Audio (%)
Butterfly Focus	22	56.4	4	12.1
Amphibian focus (behavior & conservation)	4	10.3	15	45.5
Amphibian focus (habitat & adaptation)	12	30.8	12	36.4
Mixed/General focus	1	2.6	2	6.1
Total	39	100.0	33	100.0

Each learner used both learning material modalities (visual and audio) and wrote two social media posts, one after each session.

Table 5. Summary of social media post content types between modalities.

Social media post content type	Visual count	Visual (%)	Audio count	Audio (%)
Conservation/Urgency/Persuasive	125	57.1	69	43.4
Emotional/Awe-inspiring	75	34.2	18	11.3
Factual/Objective/Informative	19	8.7	41	25.8
Other	0	0.0	31	19.5
Total	219	100.0	159	100.0

($n = 36$) with what they asked Cipherbot, 72.2% had the same behavior for Audio. This asymmetric pattern suggests that while the overall rating distributions were similar between modalities, the specific transitions between rating categories differed systematically.

Following the rating analysis, a thematic qualitative coding approach was employed to investigate the actual topics present in each answer (social media posts) through topic and semantic analyses.

First, topic analysis revealed 72 segments across both sessions for all learners ($n = 67$) (Table 4). In terms of visual modality, the Butterfly Focus category constituted the largest portion of the visual content at 56.4% ($n = 22$). This dominance suggests that the visual channel was primarily used to showcase the aesthetic qualities, forms, and colorful patterns of butterflies. The secondary focus, Amphibian Focus (Habitat & Adaptation) at 30.8% ($n = 12$), was also consistent with visual content, covering aspects such as camouflage, mimicry, and rainforest environments. In contrast, the audio modality primarily focused on amphibians (Behavior & Conservation) at 45.5% ($n = 15$), followed closely by amphibians (Habitat & Adaptation) at 36.4% ($n = 12$). The relatively low percentage of butterfly content in the audio modality (12.1%) indicates that auditory content addressed only the non-visual and informational aspects of butterflies, such as evolution or symbolism, rather than their visual characteristics.

In addition, a semantic analysis was conducted to examine the distribution of content types across visual and audio modalities. The analysis revealed distinct patterns in how content was categorized within each modality, based on 219 visual segments and 159 audio segments (Table 5). In the visual modality, the Conservation/Urgency/Persuasive category constituted the largest portion (57.1%, $n = 125$), suggesting that visual content was predominantly utilized to convey conservation messages and persuasive appeals. The secondary category, Emotional/Awe-Inspiring, accounted for 34.2% ($n = 75$) of the visual content, indicating a substantial emphasis on evoking emotional responses through visual imagery. The Factual/Objective/Informative category represented a smaller proportion (8.7%, $n = 19$), while no content was classified as Other (0.0%). In contrast, the audio modality demonstrated a more distributed pattern across categories. Conservation/urgency/persuasive content remained the dominant category at 43.4% ($n = 69$), although proportionally lower than that in the visual modality. The actual/objective/informative category accounted for 25.8% ($n = 41$), representing a substantially higher proportion than the visual content. Emotional/Awe-Inspiring content comprised only 11.3% ($n = 18$), while the other category represented 19.5% ($n = 31$) of the audio content. These findings indicate that, while visual content emphasizes conservation messaging through emotional and persuasive approaches, audio content incorporates a more balanced distribution of factual information alongside persuasive elements.

4.4. RQ4: How does the order of learning contents influence learning behaviors?

To investigate how the order of content influences learning behaviors, a series of Mann–Whitney U tests was conducted to examine order effects across counterbalanced sequences. The analysis revealed significant order effects in 12 of the 40 behavioral measures (30%). These effects were distributed across

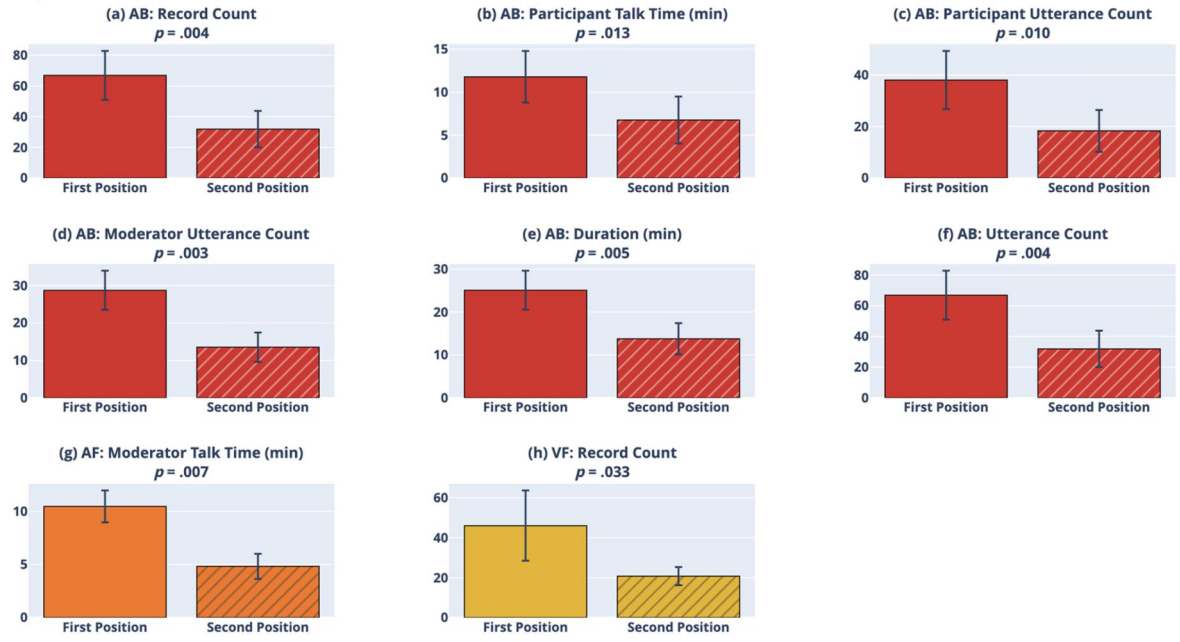


Figure 9. Comparing means between first and second session orders for each significant order effect ($p < 0.05$) with error bars representing standard errors. First, we have AB sequence effects showing declines in second session when interacting with cipherbot(a) record count, which refers to time of recording; (b) learner talk time, which refers to the duration that learner talks; (c) learner and (d) moderator utterance count, which refers to number of speaking turns for learners and moderator's; (e) duration, which refers to the total time; and (f) utterance count, which refers to total speaking turn between learner and moderator. Second, (g) AF sequence effect on moderator talk time shows reduced moderator engagement in the second session. Lastly, (h) VF sequence effect on record count demonstrates visual modality order influence.

sequences in the following order: Audio-Butterfly (AB) showed eight significant effects (67%) (Figure 9(a–f)), Audio-Frog (AF) (Figure 9(g)) and Visual-Frog (VF) (Figure 9(h)) showed two effects (17%), and Visual-Butterfly (VB) showed no effects. The concentration of order effects in the AB sequence, particularly for duration and interaction count measures, suggests a systematic carryover when audio presentation precedes the butterfly task. The significant results are shown in Figure 9.

The audio butterfly (AB) sequence exhibited the most pronounced order sensitivity, accounting for eight of the 12 significant effects. Learners who completed AB in the first position ($n = 15$) demonstrated significantly higher recording frequencies ($M = 66.87$, SD = 61.49) than those who completed it in the second position ($n = 17$, $M = 31.88$, SD = 48.84), $U = 203.5$, $p = 0.004$, representing a mean difference of 34.98 recordings. This pattern suggests that prior audio exposure influenced the recording strategies during the butterfly task. This difference reflects not just more recordings, but also a more active information-capture strategy, where learners actively engage with the material. For instance, learners were observed actively filtering and capturing key facts for later use ("It's just trying to keep me on the Blue Morpho. But I want the context first. I'm fascinated. It's amazing. I didn't know about this prismatic thing. I learned really cool stuff," S07, Participant).

Learner engagement patterns varied significantly across the AB sequences. Specifically, learner talk time showed significant differences when measured in both minutes (first position: $M = 11.79$, SD = 11.63; second position: $M = 6.76$, SD = 11.28; $U = 194.0$, $p = 0.013$) and seconds (first position: $M = 707.20$, SD = 697.91; second position: $M = 405.41$, SD = 676.96; $U = 194.0$, $p = 0.013$), with a mean difference of 5.03 min or 301.79 s. The learner utterance counts also differed significantly, with first-position learners producing more utterances ($M = 38.07$, SD = 43.59) than second-position learners ($M = 18.35$, SD = 33.43), $U = 196.5$, $p = 0.010$, indicating altered verbal engagement patterns following the sequential task combination. This activity resulted from the learners constantly adapting to conversational flow with EAIC ("Why? Every time it gives me. Sorry. Certainly. Great question. Why? It's not directly answer the question. Why? What? Why? It needs to give me that. Like I'm with the

teacher," S11, Participant). An explanation for this result could be from the "Learn" feature of Cipherbot, which actively encourages users to learn like educational assistants.

Moderator interaction patterns were similarly affected by AB sequence. Moderator utterance count varied significantly across conditions, with first-position sessions showing higher moderator contributions ($M = 28.73$, SD = 20.24) than second-position sessions ($M = 13.53$, SD = 16.24), $U = 206.5$, $p = 0.003$, potentially reflecting adaptive responses to learner behavior influenced by task order. Furthermore, the total session duration demonstrated significant order effects when measured in both minutes (first position: $M = 25.07$, SD = 17.54; second position: $M = 13.76$, SD = 15.00; $U = 202.0$, $p = 0.005$) and seconds (first position: $M = 1506.47$, SD = 1040.66; second position: $M = 819.18$, SD = 897.18; $U = 206.0$, $p = 0.003$), and the overall utterance count across both learners and moderators also showed significant variation (first position: $M = 66.87$, SD = 61.49; second position: $M = 31.88$, SD = 48.84; $U = 203.5$, $p = 0.004$).

These converging effects suggest that the Audio-Butterfly sequence creates systematic differences in engagement intensity, interaction pacing, and session structure, with first-position sessions consistently demonstrating higher levels of behavioral activity across all measured domains. Extended duration and moderator effort were often required to manage highly engaged verbal learners ("Yeah. Because it is an educational chatbot. That's the personality that we are trying to test here. The idea is actually that it is to be used in an educational environment," S11, Moderator).

The audio frog (AF) sequence exhibited only two significant order effects, both isolated to moderator behavior. Moderator talk time differed significantly across AF sequence conditions when measured in minutes, with first-position sessions ($n = 17$, $M = 10.47$, SD = 6.17) showing substantially higher moderator talk time than second-position sessions ($n = 18$, $M = 4.83$, SD = 5.02), $U = 234.5$, $p = 0.007$, representing a mean difference of 5.64 min. This effect was confirmed with second-level measurement precision (first position: $M = 629.12$, SD = 376.53; second position: $M = 296.39$, SD = 302.25; $U = 232.0$, $p = 0.010$), with a mean difference of 332.73 s. The isolation of these effects on moderator contributions suggests that the audio-frog combination specifically influenced instructor behavior rather than learner engagement patterns, indicating a more limited scope of carryover effects compared to the AB sequence ("You can ask questions to figure out interesting facts, or you can start with the recommendation questions on the screen… But you can always ask questions, so it describes," P01, Moderator).

The visual frog (VF) sequence demonstrated two significant order effects affecting distinct behavioral domains. Recording frequency varied significantly across VF sequence conditions, with first-position learners ($n = 17$, $M = 46.00$, SD = 72.50) showing higher recording counts than second-position learners ($n = 15$, $M = 20.73$, SD = 17.68), $U = 184.5$, $p = 0.033$, though this effect (mean difference = 25.27) was more modest in magnitude than the comparable effect observed in the AB sequence ("I'm trying to gather facts right now that are going to be easy to digest for individuals that would be on a social media platform that might be casually scrolling," S132, Participant). Additionally, total session duration showed a significant difference when measured in seconds, with first-position sessions ($M = 991.59$, SD = 534.80) being longer than second-position sessions ($M = 692.47$, SD = 633.66), $U = 180.0$, $p = 0.050$, representing a mean difference of 299.12 s and constituting the weakest significant effect observed across all sequences. The limited number and marginal strength of these effects suggest that the VF sequence produces relatively stable behavioral patterns with minimal carryover influence ("How come, uh, I'm asking him a question. He cannot ask me a question. This is. This is strange. I'm asking the questions," S11, Participant).

The visual butterfly (VB) sequence exhibited no significant order effects across any of the 40 behavioral measures examined. Although first-position learners ($n = 18$) showed numerically higher means across several metrics compared to second-position learners ($n = 17$), including recording count ($M = 43.78$, SD = 37.39 vs. $M = 25.00$, SD = 20.21, $U = 201.5$, $p = 0.113$), learner talk time in minutes ($M = 9.26$, SD = 5.25 vs. $M = 6.82$, SD = 7.11, $U = 197.0$, $p = 0.151$), and total duration in minutes ($M = 21.00$, SD = 20.26 vs. $M = 13.88$, SD = 12.46, $U = 201.0$, $p = 0.117$), none of these differences reached statistical significance. The limited number and marginal strength of these effects suggest that the VB sequence produces relatively stable behavioral patterns with minimal carryover influence. This

stability was rooted in the learner's perception of the interaction as a purely task-oriented exercise, limiting exploratory behavior ("It treated my questions as a learning kind of task. Although my goal was to complete a task, not learn because I already learned by consuming the podcast." S03, Participant).

5. Discussion

5.1. Theoretical implications

Overall, the results support the major theory regarding priming modalities in AI-assisted learning environments, in which we expanded CTML theory (Mayer, 2005), CLT (Sweller, 2011) and modality effect (Ginns, 2005). Across 67 learners in higher education, visual materials demonstrated superior recall performance, whereas learner experience with AI and task familiarity moderated engagement patterns across modalities. Order effects emerged primarily in the AB sequence, while learner preferences showed a balanced distribution between modalities.

Regarding RQ1 *(How do modalities (visual and audio) influence learners' ability to recall knowledge?),* the findings proved the significant influence of visual learning materials on actual learning progress. Traditionally, audio modalities have been discussed as capable of reducing cognitive load in certain learning contexts; however, in the AI-assisted learning environment, especially when using EAICs, the direction has changed. Visual learning material was found to produce higher recall capability across content types, suggesting that audio learners might have reconstructed mental representations from memory during chatbot dialogue, while visual learners can reference stable mental models from the actual learning materials. Specifically, in this study, the visual material includes text and a photo of the species; therefore, it expands the previous theories of multimodality, which include text and audio, and would work better than text only. This result can expand the definition of using multiple modalities for text and audio.

In addition, previously, Baddeley (2003) discussed the idea that working memory includes temporary storage and manipulation of information, which could be used for cognitive activities, and Fenesi et al. (2015) conceptualized this idea in educational research. Therefore, this result has expanded the arguments toward how working memory would work in the AI-assisted learning environment, showing that visual materials (including both text and image) help learners hold their working memory better than audio materials. Furthermore, the qualitative data of the visual modality showed a self-paced, reviewable format aligned with Mayer's (2005) cognitive theory of multimedia learning, which posits that learners benefit when they can control the pace of information presentation to manage their cognitive load. The audio modality showed challenges, which align with cognitive load theory (Sweller, 2011), as the inability to revisit audio content likely increased the extraneous cognitive load. The transient nature of auditory information means that learners must encode and store material in working memory without external support, whereas visual learners can offload this demand by re-referencing the text. This mechanistic difference provides a plausible explanation for the performance gap observed for RQ1.

Regarding RQ2 *(How do learners' experiences (of the AI chatbot and task) influence learning experiences between modalities (visual and audio)?),* our results indicate that both pre-knowledge of AI chatbot usage and task experiences influence how they think during the learning session with Cipherbot. Their think-aloud data showed that their thinking flows differed dramatically between learners with high and low levels of experience (1–3 as low and 4–5 as high in both variables). Between the two types of modalities used in this experiment, we found that audio modality advantages novices but disadvantages experts, whereas visual modality provides consistent, equitable experiences regardless of expertise or task familiarity. This suggests that modality selection should be personalized based on learner characteristics. In addition, with the qualitative results of AI expertise, we found that the behavior aligns with Mayer's (2021) cognitive theory of multimedia learning, which posits that learners benefit when they can control the pace of information presentation to manage their cognitive load. On the other hand, from the qualitative results of task experience, the findings showed that the behaviors when interacting with the EAIC helped novice social media writers compensate for their lower expertise while remaining useful for experienced learners. These patterns align with cognitive load theory: transient audio imposes

greater demands on novices lacking relevant schemas (Sweller, 2011), while persistent text provides scaffolding across expertise levels (Mayer, 2024).

Regarding RQ3 *(How do materials modalities (visual and audio) influence task completion?),* The findings expanded the DCT (J. Clark & Paivio, 1991) which originally focused on the verbal and non-verbal systems of learning, but now expanded in AI-assisted learning environments in the context of learning from different types of materials, influencing their memories and interactions with EAICs to finalize their learning tasks. In addition, the results of the task completions also expand the theories of asking questions with purposes in interacting with EAICs (Y. Chen et al., 2023; Lee et al., 2023; Salminen et al., 2024; Yin et al., 2024), which focus on how questions are asked for creative purposes, in this case, for business-related purposes.

Regarding RQ4 *(How does the order of learning contents influence learning behaviors?),* the current findings justify the importance of working memory in interacting with EAICs. The audio materials showed disadvantages, as the main recall of the audio works directly with the participants' memories after listening only, while the visual materials can be freely reviewed anytime, based on any order, with a purpose during the given time. In addition, the heavy focus of Audio-Butterfly's significant results shows the difficulty of "less known" content (butterfly) with "less preferred" modality (audio), establishing the behavior pattern. We suggest the additional theories of "adapting with difficulties," which means the learners with difficulty in the Audio-Butterfly sequence establish the adaptation toward learning to cover the knowledge missing from the starting point (after the learning material). Learners need to expand their interactions with Cipherbot and moderators to recover the missing knowledge they could not obtain from the materials and their pre-knowledge. This strategy expanded even when they continued to the second session, which was the VF sequence. On the other hand, the other sequence of the VB showed significant improvement in results with the advancement of visual material. Overall, these findings show the importance of having more useful material and higher pre-knowledge to influence learners' thinking when interacting with EAICs.

5.2. Practical implications

Based on our findings, we propose a guideline for using EAICs and applying multiple modalities to enhance the learning experience with EAICs.

Decide how EAICs would be used in the course: Support teaching, store materials, independent learning, or a complete mix of all. Support teaching refers to integrating the platform to replace the tasks of a teaching assistant, including keeping track of learning performance (e.g., the ability to recall knowledge and learners' experiences in learning activities). Store materials refer to internal knowledge databases that can be used to find, validate, and learn. Independent learning refers to creating an environment for learners to learn at their own pace and preferences (e.g., practice with a focused topic, creating AI-generated audio learning materials, and evaluating tasks).

Tailor the design according to *who* the learners are (e.g., backgrounds, expertise levels, prior experiences). Educators should pre-assess AI literacy and task familiarity before selecting modalities and deciding on support levels. For learners with varying AI expertise, visual materials may provide more equitable experiences, whereas audio formats require careful consideration, as they can create accessibility gaps between novice and expert users. The nature of audio content poses particular challenges for beginners, who need time to process information, whereas experienced users can navigate it efficiently. Similarly, task-specific experience determines the type of EAIC support that is required. For example, experienced learners typically focus on validation and fact-checking, requiring minimal guidance, whereas less experienced learners benefit from structured guides that clarify tasks and goals. Understanding learner backgrounds allows educators to decide on suitable instruction and offer visual materials as a foundation for all learners while using audio for more advanced and experienced students or providing supplementary support resources for the less experienced to attempt audio-based learning activities.

Evaluate *what* can influence learning experiences with EAICs (e.g., experiences, learning preferences, chatbot usage, modality effects, and sequence). Educators should assess how different modalities align with learning outcomes. While learner preferences may be balanced, effectiveness can vary significantly depending on whether the goal is knowledge retention or creative production. Understanding these

modality effects helps select suitable formats for different objectives. Additionally, monitoring task completion strategies reveals how different modalities shape learners' approaches, informing decisions about when to use visual versus audio materials. Sequence effects require particular attention. Educators should track engagement analytics to identify patterns in which prior activities influence subsequent performance. To maintain consistent engagement and avoid scheduling consecutive sessions with the same modality, we need to regularly review EAIC analytics, detect possible issues, and adjust course design proactively, ensuring that technology integration supports learners' experience.

Examine *why* the learners are learning with EAICs (e.g., knowledge retention goals, creative production objectives, skill development aims). Educators should match modality selection to primary learning objectives rather than assuming that one format suits all educational purposes. When the goal is knowledge retention and factual accuracy, the visual materials provide stronger support. Audio materials can provide better stimulation when prioritizing creative production and persuasive communication. Both modalities offer valuable learning opportunities for AI literacy and digital skill development, although the type and level vary based on learner expertise. In courses with multiple learning objectives, consider using mixed approaches: visual materials for foundational knowledge acquisition, followed by audio for creative applications.

5.3. Limitations and future research

This study has some limitations. First, limited modality, as we focused mainly on visual (text and image) and audio (podcast), might influence learners' experience with how they learn and interact with Cipherbot. For future expansion, we will include other types of educational materials, such as video, printed text, or PowerPoint slides.

Second, learners were given only five minutes of learning time, which means that participants only had five minutes to receive information from either their reading or listening. Future research should either expand or diversify timing based on learners' behavior. For example, slow readers could be given more time, or fast listeners could adjust their timing based on their preferences.

Third, there might be some confounding variables that influence learning behavior due to the design of Cipherbot. For the future expansion of this research, we recommend the use of several different types of EAICs for more impactful results.

Fourth, the combination of text and image that made up the visual material proved to work better than audio only, whereas in previous studies, the combination of text and audio proved to work better than text only. For future research, further expansion to test different combinations could be conducted to suit different types of learning experience.

Fifth, the audio modality was generated by Cipherbot, which can be considered robotic. From the audio transcript, we found some comments that mentioned AI-generated priming material. For example: "I like podcasts, but I *hated* this podcast," "it was fake" or "I knew it tried to sound natural, but it wasn't."

Lastly, most learners had higher educational backgrounds, with 44.8% being college graduates or higher, which could show an expected higher capability in using EAICs and working on intensive experiments like this study. This limits the generalizability of the results. Future studies should include a more diverse group of educational levels (e.g., K12, vocational schools) to repeat the study in different educational contexts.

6. Conclusion

Previous research has shown different influences between material modalities; however, we have always asked ourselves as researchers and educators: "Would that still matter in the age of AI?" This research focuses on investigating the actual impacts on how learners receive information, what imprint in their head, and how learners' working memories work in an AI-assisted learning environment. This study reveals that the learning progress and even interaction with EAIC were not actually influenced much by modalities but by how learners receive information, how they think, and how they discuss their knowledge. Visual modalities influence how they recall knowledge, whereas audio modalities have more conversational interactions.

Acknowledgements

The authors would like to thank all the participants who volunteered for this study and the moderators who facilitated the data collection sessions. This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical approval

This study was approved by the Institutional Review Board of Hamad Bin Khalifa University (approval reference number: HBKU-IRB-2025-60). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all the individual participants included in the study. All participants signed consent forms prior to participation, agreeing to share their data for research purposes. The participants received a small non-monetary gift for their participation, which was disclosed during the consent process. All the participants provided informed consent for the publication of their anonymized data in academic research outputs. The privacy rights of all human subjects have been observed. No personally identifiable information is included in this manuscript. All data presented is anonymized to protect participant confidentiality.

Author contributions

CRediT: **Trang Xuan**: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Writing – original draft; **Joni Salminen**: Methodology, Project administration, Supervision, Validation, Writing – review & editing; **Ilkka Kaate**: Data curation, Resources; **Farhan Ahmed**: Conceptualization, Data curation, Writing – original draft; **Danial Amin**: Data curation; **Rajat Patil**: Data curation; **Soon-Gyo Jung**: Software; **Jinan Y. Azem**: Data curation, Project administration; **Bernard J. Jansen**: Data curation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Trang Xuan http://orcid.org/0009-0006-0110-7530
Joni Salminen http://orcid.org/0000-0003-3230-0561
Ilkka Kaate http://orcid.org/0000-0002-9429-3566
Farhan Ahmed http://orcid.org/0009-0005-0803-9246
Danial Amin http://orcid.org/0009-0000-7597-2267
Rajat Patil http://orcid.org/0009-0001-5114-6494
Soon-Gyo Jung http://orcid.org/0000-0002-6130-8012
Jinan Y. Azem http://orcid.org/0009-0003-0490-6707
Bernard J. Jansen http://orcid.org/0000-0002-6468-6609

Data availability statement

The datasets generated during the current study are publicly available in the Open science framework (OSF) repository: https://osf.io/w7b68/. Owing to privacy restrictions protecting participant confidentiality, audio recordings, and personally identifiable information were excluded from the public dataset. The repository contains anonymized survey responses, anonymized chatbot interaction logs, variable codebook, and study materials (learning content and survey instruments).

References

Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2024). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, *29*(8), 10153–10176. https://doi.org/10.1007/s10639-023-12166-w

Alsobeh, A., & Woodward, B. (2023). AI as a partner in learning: A novel student-in-the-loop framework for enhanced student engagement and outcomes in higher education. *Proceedings of the 24th Annual Conference on Information Technology Education* (pp. 171–172).

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Belda-Medina, J., & Kokošková, V. (2023). Integrating chatbots in education: Insights from the chatbot-human interaction satisfaction model (CHISM). *International Journal of Educational Technology in Higher Education*, *20*(1), 62. https://doi.org/10.1186/s41239-023-00432-3

Byrne, B. M. (2013). *Structural equation modeling with EQS: Basic concepts, applications, and programming.* Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780203726532/structural-equation-modeling-eqs-barbara-byrne-barbara-byrne.

Cavanagh, T. M., & Kiersch, C. (2022). Using commonly-available technologies to create online multimedia lessons through the application of the cognitive theory of multimedia learning. *Educational Technology Research and Development*, *71*(3), 1–21. https://doi.org/10.1007/s11423-022-10181-1

Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers*, *25*(1), 161–182. https://doi.org/10.1007/s10796-022-10291-4

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Cohen, S. (1988). *Perceived stress in a probability sample of the United States.* https://psycnet.apa.org/record/1988-98838-002?email=RlFUZ1JKSTJwwwxtNFN4M05KRGU0MFdvbXJVcDZSZ1l3WUR5R0RxV21SMD0tLTlOSFd5Q2VacHFQSHg1-ME5ONVB6bXc9PQ%3D%3D–9dfeb6e1fdc172d071f735152115700206acc5db

Deveci Topal, A., Dilek Eren, C., & Kolburan Geçer, A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies*, *26*(5), 6241–6265. https://doi.org/10.1007/s10639-021-10627-8

Fang, C., & Tse, A. W. C. (2022). Quasi-experiment: Postgraduate students' class engagement in various online learning contexts when taking privacy issues to incorporate with artificial intelligence applications. *Proceedings of the 14th International Conference on Education Technology and Computers* (pp. 356–361). ACM. https://doi.org/10.1145/3572549.3572606

Feine, J., Morana, S., & Maedche, A. (2020). A chatbot response generation system. *Proceedings of the Conference on Mensch Und Computer* (pp. 333–341). ACM. https://doi.org/10.1145/3404983.3405508

Fenesi, B., Sana, F., Kim, J. A., & Shore, D. I. (2015). Reconceptualizing working memory in educational research. *Educational Psychology Review*, *27*(2), 333–351. https://doi.org/10.1007/s10648-014-9286-y

Fersch, M., Schacht, S., & Woldai, B. (2023). Exploring AI in education: A quantitative study of a service-oriented university chatbot. In *The Paris Conference on Education 2023 Official Conference Proceedings* (pp. 439–453). IAFOR. https://doi.org/10.22492/issn.2758-0962.2023.37

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior*, *75*, 461–468. https://doi.org/10.1016/j.chb.2017.05.045

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, *15*(4), 313–331. https://doi.org/10.1016/j.learninstruc.2005.07.001

Han, K., & Xu, J. (2024). Research on learner classroom behavior recognition based on YOLOv5 algorithm. *Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence, ICIEAI '23* (pp. 177–182). ACM. https://doi.org/10.1145/3660043.3660075

Holderried, F., Stegemann–Philipps, C., Herschbach, L., Moldt, J.-A., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., & Mahling, M. (2024). A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, *10*(1), e53961. https://doi.org/10.2196/53961

Jeon, J. (2023). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, *36*(7), 1338–1364. https://doi.org/10.1080/09588221.2021.1987272

Jiang, D., Renandya, W. A., & Zhang, L. J. (2017). Evaluating ELT multimedia courseware from the perspective of cognitive theory of multimedia learning. *Computer Assisted Language Learning*, *30*(7), 726–744. https://doi.org/10.1080/09588221.2017.1359187

Jung, S.-G., Medina, J., Aldous, K., Azem, J., Salminen, J., & Jansen, B. J. (2025). Cipherbot: A learning platform for AI-augmented education. *Proceedings of the Augmented Humans International Conference 2025* (pp. 478–481). ACM. https://doi.org/10.1145/3745900.3746106

Kim, N.-Y., Cha, Y., & Kim, H.-S. (2019). Future English learning: Chatbots and artificial intelligence. *Multimedia-Assisted Language Learning*, *22*(3), 32. https://doi.org/10.15702/mall.2019.22.3.32

Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and Instruction*, *12*(1), 1–10. https://doi.org/10.1016/S0959-4752(01)00014-7

Kroll, N. E. A., & Schepeler, E. M. (1985). Visual priming effects as a measure of short-term visual memory. *The American Journal of Psychology*, *98*(3), 449–468. https://doi.org/10.2307/1422629

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, *20*(1), 56. https://doi.org/10.1186/s41239-023-00426-1

Lee, G.-G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., Guo, S., Wu, Z., Liu, Z., Wang, H., Mai, G., Liu, T., & Zhai, X. (2023). Multimodality of AI for education: Towards artificial general intelligence. *arXiv: 2312.06037*. https://doi.org/10.48550/arXiv.2312.06037

Lewis, S., Dontcheva, M., & Gerber, E. (2011). Affective computational priming and creativity. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 735–744). ACM. https://doi.org/10.1145/1978942.1979048

Liu, C.-C., Liao, M.-G., Chang, C.-H., & Lin, H.-M. (2022). An analysis of children' interaction with an AI chatbot and its impact on their interest in reading. *Computers & Education*, *189*, 104576. https://doi.org/10.1016/j.compedu.2022.104576

Liu, L., Subbareddy, R., & Raghavendra, C. G. (2022). AI intelligence chatbot to improve students learning in the higher education platform. *Journal of Interconnection Networks*, *22*(Supp02), 2143032. https://doi.org/10.1142/S0219265921430325

Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*, *58*(11), 1276–1285. https://doi.org/10.1111/medu.15402

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 31–48). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.004

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press.

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, *36*(1), 8. https://doi.org/10.1007/s10648-023-09842-1

Moreno, R. (2006). Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. *Journal of Computer Assisted Learning*, *22*(3), 149–158. https://doi.org/10.1111/j.1365-2729.2006.00170.x

Paivio, A. (2014). Intelligence, dual coding theory, and the brain. *Intelligence*, *47*, 141–158. https://doi.org/10.1016/j.intell.2014.09.002

Robayo-Pinzon, O., Rojas-Berrio, S., Rincon-Novoa, J., & Ramirez-Barrera, A. (2024). Artificial intelligence and the value co-creation process in higher education institutions. *International Journal of Human–Computer Interaction*, *40*(20), 6659–6675. https://doi.org/10.1080/10447318.2023.2259722

Salminen, J., Jung, S., Medina, J., Aldous, K., Azem, J., Akhtar, W., & Jansen, B. J. (2024). Using Cipherbot: An exploratory analysis of student interaction with an LLM-based educational chatbot. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 279–283). ACM. https://doi.org/10.1145/3657604.3664690

Savelka, J., Agarwal, A., An, M., Bogart, C., & Sakr, M. (2023). Thrilled by your progress! Large language models (GPT-4) no longer struggle to pass assessments in higher education programming courses. *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1, ICER '23, 1* (pp. 78–92). ACM. https://doi.org/10.1145/3568813.3600142

Sharma, M., Yadav, S., Kaushik, A., & Sharma, S. (2021). Examining usability on Atreya Bot: A chatbot designed for chemical scientists. *2021 International Conference on Computational Performance Evaluation (ComPE)* (pp. 729–733). IEEE. https://doi.org/10.1109/ComPE53109.2021.9752288

Skulmowski, A., & Xu, K. M. (2022). Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, *34*(1), 171–196. https://doi.org/10.1007/s10648-021-09624-7

Stuart, E., Hanson, J. L., & Dudas, R. A. (2019). The right stuff: Priming students to focus on pertinent information during clinical encounters. *Pediatrics*, *144*(1), e20191311. https://doi.org/10.1542/peds.2019-1311

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (vol. 55, pp.37–76). Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Tabbers, H. K., Martens, R. L., & Van Merriënboer, J. J. G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *The British Journal of Educational Psychology*, *74*(Pt 1), 71–81. https://doi.org/10.1348/000709904322848824

Van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies: Cognitive load theory. *Medical Education*, *44*(1), 85–93. https://doi.org/10.1111/j.1365-2923.2009.03498.x

Wang, A., Wu, X., Tang, X., & Zhang, M. (2020). How modality processing differences affect cross-modal nonspatial repetition inhibition. *PsyCh Journal*, *9*(3), 306–315. https://doi.org/10.1002/pchj.332

Wang, S., Qiu, L., & Sun, C. (2025). Adaptive education system for drama education in college education system based on human-computer. *International Journal of Human–Computer Interaction*, *41*(3), 1872–1887. https://doi.org/10.1080/10447318.2022.2079169

Wu, X., Ma, J., Shen, Y., Feng, J., Ma, S., & Li, W. (2025). Effects of feedback types and modalities on university students' online test performance: An fNIRS-based study. *International Journal of Human–Computer Interaction*, *0*(0), 1–20. https://doi.org/10.1080/10447318.2025.2565388

Xie, C., Bagozzi, R. P., & Grønhaug, K. (2019). The impact of corporate social responsibility on consumer brand advocacy: The role of moral emotions, attitudes, and individual differences. *Journal of Business Research*, *95*, 514–530. https://doi.org/10.1016/j.jbusres.2018.07.043

Xue, L., Ghazali, N., & Mahat, J. (2025). A systematic review of UTAUT and UTAUT2 for AI adoption in education. *International Journal of Human–Computer Interaction*, *0*(0), 1–25. https://doi.org/10.1080/10447318.2025.2552867

Yaacob, Y., Mahmud, M. M., Nagasundram, U., Mustamam, N. I., Ahmad, R., & Mohd A'Seri, M. S. (2025). Navigating new norms: ChatGPT's Integration and its critical reception in higher education. *Proceedings of the 2024 16th International Conference on Education Technology and Computers, ICETC '24* (pp. 179–184). ACM. https://doi.org/10.1145/3702163.3702411

Yin, J., Goh, T.-T., & Hu, Y. (2024). Interactions with educational chatbots: The impact of induced emotions and students' learning motivation. *International Journal of Educational Technology in Higher Education*, *21*(1), 47. https://doi.org/10.1186/s41239-024-00480-3

Yin, J., Goh, T.-T., Yang, B., & Hu, Y. (2024). Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance. *IEEE Transactions on Learning Technologies*, *17*, 1378–1389. https://doi.org/10.1109/TLT.2024.3364015

Yue, C., Kim, J., Ogawa, R., Stark, E., & Kim, S. (2013). Applying the cognitive theory of multimedia learning: An analysis of medical animations. *Medical Education*, *47*(4), 375–387. https://doi.org/10.1111/medu.12090

About the authors

Trang Xuan is a doctoral candidate at the School of Marketing and Communication in University of Vaasa, Vaasa, Finland. Trang's current research focus is on the usability of AI chatbots in Marketing Education.

Joni Salminen is an Associate Professor (tenure track) at the School of Marketing and Communication in University of Vaasa, Vaasa, Finland. His research focuses on data-driven personas, along with overlapping topics such as Human-Centered AI, Quantitative UX, Interactive Systems, User Segmentation Algorithms, and so on.

Ilkka Kaate is a postdoctoral researcher in marketing at the University of Turku, Finland, with research interests in deepfake personas. Additionally, Ilkka is a digital marketing entrepreneur specializing in social media and search marketing. In his free time, Ilkka likes to act, sing (solo and choir), and play the guitar.

Farhan Ahmed is a doctoral student at the University of Vaasa researching AI chatbots and digital personas in education. He holds a Master of Business Administration from Mid Sweden University. Experienced in software development, cloud computing, and intelligent product development, he holds Azure AI and Oracle Cloud certifications.

Danial Amin is a doctoral candidate at the University of Vaasa, focusing on ethical AI and persona development for social good, with extensive experience in data science and AI development. His research leverages GenAI technologies to develop ethical and inclusive AI-generated personas that benefit marginalized communities in the Global South.

Rajat Patil is a Doctoral Student at the University of Vaasa, Vaasa, Finland. His research focuses on Data-Driven Personas for Sustainable Growth.

Soon-gyo Jung is a Full-Stack Software, AI, and Data Engineer at Qatar Computing Research Institute (QCRI), specializing in scalable data-driven systems, LLM applications, and research-oriented software engineering.

Jinan Y. Azem is a Researcher and UX/Product Designer at Qatar Computing Research Institute. Her work focuses on user-centered digital systems, with research interests in usability, user experience, accessibility, and the application of HCI methods to digital products, and explores how digital technologies can better serve diverse user populations.

Jim Jansen is a Principal Scientist at the Qatar Computing Research Institute, working on automatically visualizing user data. He is a West Point graduate with a PhD in computer science from Texas A&M University. Professor Jansen is editor-in-chief of Information Processing & Management and former editor-in-chief of Internet Research.