Chatbots for Active Learning: A Case of Phishing Email Identification

Sebastian Hobert , Asbjørn Følstad , Effie Lai-Chong Law

PII: S1071-5819(23)00117-9
DOI: https://doi.org/10.1016/j.ijhcs.2023.103108
Reference: YIJHC 103108

To appear in: *International Journal of Human-Computer Studies*

Received date: 22 August 2022
Revised date: 12 June 2023
Accepted date: 12 July 2023

Please cite this article as: Sebastian Hobert , Asbjørn Følstad , Effie Lai-Chong Law , Chatbots for Active Learning: A Case of Phishing Email Identification, *International Journal of Human-Computer Studies* (2023), doi: https://doi.org/10.1016/j.ijhcs.2023.103108

Highlights

- Educational chatbots can be designed to support varying levels of cognitive engagement and active learning.
- Chatbot interaction designs for enhanced cognitive engagement increase users' time spent within chatbot-based learning processes.
- Users’ perceived subjective learning outcome can be increased by chatbot interaction designs for enhanced cognitive engagement.
- Future research is required to determine the effects of chatbot interaction designs for cognitive engagement and active learning on user engagement.

Chatbots for Active Learning: A Case of Phishing Email Identification

Sebastian Hobert[a], Asbjørn Følstad[b]*, Effie Lai-Chong Law[c]

[a] University of Goettingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany, shobert@uni-goettingen.de

[b] SINTEF, Forskningsveien 1, 0373 Oslo, Norway, Asbjorn.Folstad@sintef.no

[c] Durham University, Department of Computer Science, Upper Mountjoy Campus, DH1 3LE Durham, U.K., lai-chong.law@durham.ac.uk

* Corresponding author

Abstract

Chatbots represent a promising approach to provide instructional content and facilitate active learning processes. However, there is a lack of knowledge as how to design chatbot interactions for active learning. In response to this knowledge gap, we conducted an experimental study (n = 164) comparing four modes for providing instructional content in chatbots, with varying demands for cognitive engagement. The four modes – passive, active, constructive, and interactive – were based on the ICAP framework of active learning. The learning content concerned identification of phishing emails and the four modes were distinguished by how the participants were invited to engage with the content during their chatbot interaction. The ICAP modes of higher cognitive engagement required participants to spend more time on the interaction and led to perceptions of higher subjective learning outcome. However, the effects of the different ICAP modes were not found to be significantly different in terms of user engagement, social presence, intention to use, or objective learning outcomes. The study represents an important first step towards understanding the design of chatbots for active learning.

Keywords: Chatbot Interactions; Educational Chatbots; Technology-Enhanced Learning; ICAP Framework

Author contributions

Sebastian Hobert: Conceptualization, Methodology, Software, Investigation, Writing – Original Draft.

Asbjørn Følstad: Conceptualization, Methodology, Formal Analysis, Investigation, Resources, Writing – Original Draft.

Effie Lai-Chong Law: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft.

Funding

This work was supported by the Stifterverband für die Deutsche Wissenschaft e. V. [grant number H120 5228 5008 32762] and the Research Council of Norway [grant number 270940].

Colour should not be used for any figures in print.

1. Introduction

In the information society, there is an increasing demand for acquisition of new knowledge and skills in an efficient and effective manner. Employees and consumers are frequently requested to engage with new material for learning and dissemination of essential information – for example, to learn skills required for safe and efficient use of digital services. To meet this demand, companies, educators, and government organizations are investing heavily in online learning tools and -contents, making online learning increasingly available and important to users. A 2020 industry report on workplace learning found that the surveyed organizations had substantially increased budgets for online learning while reducing budgets for instructor-led training (LinkedIn Learning, 2020 & 2022). Reflecting this trend, the global education technology market is forecasted a compound annual growth rate of more than 15% towards 2030 (Grand View Research, 2022). Furthermore, while online learning content is typically distributed as documents, screencasts, or videos (Taylor & Hung, 2022), there is increasing interest in exploring innovative tools and formats, such as providing content via chatbots (LinkedIn Learning, 2022).

An active learning process is important to the learning outcome (Chi & Wylie, 2014). Such a process can well be supported by conversational approaches that are increasingly deployed for online learning – specifically by the use of chatbots, that is, automated agents for access to information and services in everyday language (Følstad et al., 2021). Chatbots have, for example, been used for language learning (Huang, Hew, & Fryer, 2022), to convey microlearning content (Yin et al., 2021), to provide scaffolding during educational process (Winkler, Hobert, Salovaara, Söllner, & Leimeister, 2020), and to serve as teaching assistants (Hobert, 2021). When using chatbots for learning purposes, users engage with learning content through conversation rather than through the passive content reception typical for courses based on instructional articles or videos. In consequence, chatbots as part of learning processes have been found to be beneficial for the users' self-efficacy, engagement, and learning (Chang, Hwang, & Gao, 2021). Hence, it seems justified to assume that chatbots may represent a valuable complement to currently much used approaches to online learning such as delivering content with pre-recorded videos (Hansch et al 2015).

Despite the potential benefits of chatbots as conveyors of learning content, there is a lack of knowledge on how chatbot interactions should be designed to enable active learning and good learning outcomes. While several studies presented how chatbots were used to support learning processes (e.g., Fryer, Nakao, & Thompson, 2019), or presented design processes leading to the implementation of chatbots for learning (e.g. Hobert, 2019), there are, to the best of our knowledge, no studies that have systematically compared different chatbot interaction designs for online learning to foster active learning and cognitive engagement. This lack of knowledge may potentially obstruct the successful development and wider uptake of chatbots for online education, in part due to insufficient exploration of how educational theories and principles may be implemented in a chatbot user interface, and in part due to the lack of requisite knowledge available for designers of educational chatbots.

Motivated to bridge this knowledge gap, we have conducted a study to investigate chatbot interaction designs to support active learning and cognitive engagement. Grounded in Chi and Wylie's (2014) framework for active learning, we designed chatbot interactions that reflected learning strategies with different levels of cognitive engagement: passive, active, constructive, and interactive strategies. To increase cognitive engagement, different types of reflection tasks were integrated in the chatbot interactions: a multiple-choice quiz (active), quiz with a request to formulate answers in one's own words (constructive), and quiz with follow-up questions also to be answered in one's own words (interactive). In the passive mode, the learner was not invited to engage in any reflection task. In the active mode, the learner was engaged in simple manipulation of the learning material. In the constructive mode, the learner describes the learning content in their own words. And in the interactive mode, the learner engaged in additional turn-taking with the chatbot when reflecting on the learning content. While higher levels of cognitive engagement were assumed to imply higher resource demands (i.e., higher cognitive loads) in learners, we assumed that higher levels of cognitive engagement could induce higher levels of learner engagement and knowledge gain.

The designs were investigated in an online experiment, where 164 participants were randomly assigned to one of the four chatbot interaction designs to learn the specific content on phishing email identification. After the interaction, user engagement and knowledge gain were measured. The learning content of phishing emails was assumed to be highly relevant for the study, given the relevance and timeliness of the subject (i.e., the ramification of cybersecurity) and the availability of practices to identify emails at risk as being part of phishing awareness campaigns (Jampen, Gür, Sutter & Tellenbach, 2020).

The study contributes new knowledge of how chatbot designs reflecting different learning strategies vary in terms of user engagement and learning. Specifically, we found that while interaction designs with higher demands on cognitive engagement did require more from users in terms of time spent on the overall learning task, these users also tended to report higher levels of perceived knowledge gain. We did, however, not find differences in knowledge gain in an objective knowledge text – possibly due to a ceiling effect within our learning task setting in the experiment. User engagement was similar across all conditions, and in free text reports on their experience. Users also provided overall positive feedback on the experience, the learning content, and the learning outcome. As such, the study findings provide a first step towards expanding the knowledge base in this area of research while motivating future research.

The remainder of the paper is structured as follows. First, we provide an overview of relevant background within educational chatbots, chatbot interaction design, and educational theory on active learning. We then outline the study research question and hypotheses, before detailing methods and findings. Finally, we discuss the study findings relative to previous research, suggest implications to theory and practice, and point our limitations and promising directions for future research.

2. Background

2.1 Educational chatbots

Motivated by the recent general industry and academic interest in chatbots (Dale, 2016), there has also been a surge of research interest in chatbots for educational purposes. As chatbots allow for presenting content in a conversational and, potentially, engaging mode, there is an assumption that chatbots – either as stand-alone solutions (Fryer, Ainley, Thompson, Gibson, & Sherlock., 2017) or complements to other online educational technology (Winkler, Hobert et al., 2020) – may provide a valuable tool in a toolkit for educators. For example, Chang, Hwang, & Gao (2021) investigated how aspects of an educational program could be strengthened through the use of a chatbot to support reflection in learning. The study suggested that chatbot use could lead to increased self-reported self-efficacy, engagement, and learning.

Educational chatbots have been used for a range of purposes. Language learning through chatbots has been investigated in a series of studies by Fryer and colleagues (Fryer & Carpenter, 2006; Fryer et al., 2017; Huang et al., 2022). Here, students have exploratorily interacted with openly available chatbots for language practice – typically provided as supplement to language learning courses. Chatbots have also been used to support vocational training such as nursing (Chang, Hwang, & Gao, 2021) and programming (Yin, Goh, Yang, & Xiaobin, 2021). Here, the chatbot may offer feedback on training exercises or support the learning process for tasks that require the learner to engage in a reflection process on the study material (Hobert, 2019).

Educational chatbots have been used as stand-alone solutions for specific learning tasks. Chatbots for language learning are good examples of this (Fryer & Carpenter, 2006; Fryer et al., 2017) as well as chatbots for specific content. An early example of the latter is the Freudbot presented by Heller et al., (2005), where psychology students were invited to learn about a famous psychologist through interacting with a chatbot intended as a simple representation of Freud. However, educational chatbots seem to have been studied more in the learning context where they complement other educational technologies. For example, Fidan and Gencel (2022) investigated how chatbots and peer-feedback might augment instructional videos in online education and found increases in learners' intrinsic motivation and learning performance. Similarly, Winkler et al. (2020) investigated chatbots as a means for enriching online learning videos, by providing scaffolding material to break up the video-based lecture session and offer students a means for self-reflection and -assessment. Chatbots have also been used to support more long-term educational processes. Hobert (2019; 2021, 2023) presented a study of a chatbot coding tutor following students throughout full lecture periods of a programming course. Goel and Polepeddi (2018) presented a virtual teaching assistant supporting students throughout a programming course in a semester.

A range of different technologies have been used to implement educational chatbots, reflecting the intended purpose of the chatbots as well as the available technology at the point of implementation. Generally, the technologies used can be divided into what McTear (2021) refers to as rule-based approaches and statistical data-driven approaches. Educational chatbots used for providing support on users' frequently asked questions (FAQs) may employ statistical data-driven approaches, for example, for intent recognition. In this case, the focus typically lies on interpreting arbitrary user input and providing the most appropriate answer based on a knowledge base. The task of the chatbot in this approach is typically to react to the users' input by answering their questions. The advantage of such statistical data-driven approaches is that arbitrary input can be processed. Large language models, which have become available in recent months (see, e.g., Kasneci et al. 2023), can be seen as a further development of these statistical data-driven methods. However, large language models usually do not generate answers from a knowledge base but use generative procedures (see Subsection 6.3 for a more in-depth discussion of the implications of large language models on future research). In contrast to these statistical data-driven approaches, chatbots presenting content through longer structured dialogues (e.g., pre-defined learning paths) may apply rule-based approaches where dialogues follow scripts and users traverse dialogue trees by the use of quick replies and predefined answer alternatives, like in Winkler et al. (2020) or Hobert (2023). Whereas the statistical data-driven approach typically follows a reactive interaction approach, rule-based approaches may also define proactive rules in which the chatbots may start interactions autonomously based on predefined events (like in Winkler et al. (2020) when an interaction is started after a predefined time period). One advantage of predefined learning paths in rule-based educational chatbots is that the chatbot developers (e.g., the lecturers of a class) have full control over the learning processes, as all answer alternatives can be manually defined. Thus, the quality of the dialogues can be assured to a better extent. Overall, it is important to emphasize that both approaches (rule-based and statistical data-driven approaches) are not mutually exclusive. A combination of both methods in a chatbot is also possible, as it was done in Hobert (2023). In this case, the chatbot enables both a guided interaction based on predefined learning paths and a FAQ-like interaction based on a statistical data-driven approach. In doing so, the chatbot is able to react to learners' questions or to actively steer the conversation flow by asking the learners specific questions to decide which learning path to follow. In addition, the chatbot can proactively start new dialogs, for example, if errors occur in the learning task.

A key use-case for educational chatbots is microlearning, that is, online learning content delivered in small units, addressing skill-based knowledge needed in the short term. Taylor & Hung (2022) in their scoping review of research on microlearning noted that this approach to learning may have beneficial outcomes for acquisition of knowledge and skills, confidence, and engagement. Key instructional strategies employed in current microlearning include demonstration, gamification, and questions and answers. The main channels for educational content in microlearning research have been video, text messaging, and traditional course material. However, arguably, chatbots may be particularly suitable for microlearning as learning content broken down in chunks is suitable for a conversational mode of presentation, and a conversational style may foster user engagement. Yin, Goh, Yang, and Xiaobin (2021) studied the use of chatbots for micro-learning as part of a basic computer course. They found that students in groups with chatbot support performed equally well to students in groups with human support, and also attained higher intrinsic motivation.

2.2 Chatbot interaction design

A range of interaction design types may be supported in chatbots. As noted in a typology by Følstad, Brandtzaeg and Skjuve (2018), a fruitful distinction may be made between user-led and chatbot-led conversations. In user-led conversations, the advancement of the dialogue is typically guided by the user, such as in chatbots for customer service and virtual assistants. Here, the topic and progress of the dialogue is driven by a specific task goal defined by the user, typically without proactive initiation in the chatbot. User-led chatbots may be based on statistical data-driven approaches (McTear, 2021), as the conversation typically depends on the chatbot identifying the user's intents and responding accordingly. Chatbot-led conversations, on the other hand, are typically guided by the chatbot. Examples of this are chatbots for mental health (Fitzpatrick, Darcy, & Vierhile, 2017), where users are guided through structured, stepwise processes conveying information and insight on mental health. In chatbot-led conversations, the dialogue typically follows longer predefined dialogue trees, where users may impact the dialogue flow through selecting branches of the trees but where the overall process is provided by the chatbot. For this purpose, users are typically offered response alternatives in the form of buttons or quick replies during the conversation. Chatbot-led conversations may be valuable for educational microlearning in cases where the chatbot is to guide the user through the steps of a predefined learning topic. The chatbot-led approach to chatbot interaction design is also commonly applied in educational chatbots (Kuhail, Alturki, Alramlawi, & Alhejor, 2023), which is beneficial to ensure the pedagogical quality of the interactions.

As noted by Følstad & Brandtzaeg (2017), chatbot interaction design is characterized by a need to consider the conversation as design material. That is, when designing chatbot interactions, developing the conversational content of the interaction is key to the design task. Hence, chatbot interaction design includes both the considerations of the interaction mechanisms for providing content to the user and the content presented through the use of the same interaction mechanisms (Shevat, 2017). In an experiment of chatbot interaction mechanisms and content types (Haugeland Følstad, Taylor, & Bjørkli., 2022), providing users with the opportunity for button interaction was found to strengthen the pragmatic quality of the chatbot interaction as it was seen as efficient and effective compared to interaction in free text messages. In particular, button interaction was seen as beneficial for conversations where the chatbot helped the user explore a particular topic in depth. At the same time, free text interaction is seen as potentially more engaging, provided that it adds value to the conversation. A study by Jain, Kumar, Kota, & Patel (2018) showed that users find free text interaction capabilities in chatbots to reflect conversational intelligence and also to enable needed flexibility. Xiao et al. (2020) investigated how free text interaction could be used as a strategy to strengthen user engagement through mimicking active listening.

Chatbots are seen as potentially promising to strengthen user engagement and involvement, through humanlike interactions (Haugeland et al., 2022) in a format, which is accessible and familiar to users (Følstad & Brandtzaeg, 2017). These beneficial aspects of chatbots have been investigated in a range of application domains, including customer service (Haugeland et al., 2022), mental health (Fitzpatrick et al., 2017), and education (Chang, Hwang, & Gao, 2021). At the same time, important areas of chatbot user experience are still underexplored. Følstad et al. (2021) identified chatbot user experience as an area of key research challenges within chatbot research. In existing research, a range of theoretical frameworks and measurement instruments have been used to investigate chatbot user experience, including perceptions of chatbot anthropomorphism and social presence (Araujo, 2018; Go & Sundar, 2019), assessments of pragmatic and hedonic quality (Følstad & Brandtzaeg, 2020). Also, the relatively recent short form of the User Engagement Scale (O'Brien, Cairns, & Hall, 2018) has been applied in several chatbot studies (e.g., Feine, Morana, & Maedche, 2020; Gabrielli et al., 2021; He, Basar, Wiers, Antheunis, & Krahmer, 2022).

2.3 Active learning and ICAP framework

With the broad range of approaches and use-cases available for educational chatbots, it is particularly important to ground chatbot interaction design in theoretical assumptions regarding how to obtain desired learning outcomes. A key chatbot characteristic of relevance to an educational context is the potential for presenting learning content in an engaging manner through a conversational mode (Chang, Hwang, & Gao, 2021). Furthermore, chatbots hold potential for providing content for microlearning (Yin, et al., 2021), in part due to the ability to break up and adapt content to fit the learning context and the needs of the user.

In response to these characteristics of chatbots, a suitable theoretical basis may be the framework of active learning (Chi & Wylie, 2014). Here, the learning process is seen as strengthened if the educator is able to allow learners to engage cognitively with the material. Specifically, the framework of Chi and Wylie (2014) proposes four modes for cognitive engagement, which increasingly foster active learning:

- *passive mode*, where learners merely receive and store the educational content in memory, e.g., when reading a text without doing anything else.
- *active mode,* where learners act on the educational content through physical manipulation without making their own inferences or expressing the content in their own words, e.g., when underlining or copying content while reading a text.
- *constructive mode,* where learners use the educational content to make inferences in the form of reflections or explanations in their own words, e.g., when taking notes in one's own words or integrating across texts.
- *interactive mode,* where learners use the educational content in co-inferential processes with a learning partner, such as a peer, teacher, or computer agent, e.g., when conversing on comprehension questions.

The four modes of the framework (interactive, constructive, active, passive) lend it the abbreviated name ICAP. Educators using the ICAP framework may motivate learners to engage in different modes by adding didactical methods to the presentation of learning content to increase cognitive engagement. For example, encouraging learners to repeat or rehearse, reflect on presented content, or discuss in dyads or small groups (Chi & Wylie, 2014). In this paper we refer to the four modes of ICAP as *learning modes* or *ICAP modes.*

The ICAP framework has been adopted to inform the design of education in a range of disciplines and at different levels. For example, within lower-level education (Morris & Chi, 2020), higher-level education (Wekerle, Daumiller, & Kollar., 2022), and in disciplines such as health (Lim et al., 2019) and science and technology (Wiggins, Eddy, Grunspan, & Crowe, 2017)

The effects of the different ICAP modes have been the subject of several empirical studies. For example, the following has been investigated: learner engagement (Wekerle et al., 2022) and perceived effort (Lim et al., 2019), subjective learning outcomes and self-reported knowledge acquisition (Wekerle et al., 2022), as well as objective learning outcomes assessed through knowledge tests or assessments (Chi & Wylie, 2014; Lim et al., 2019; Morris & Chi, 2020).

Precious studies have investigated educational chatbots designed in response to the ICAP model (Winkler, Weingart, & Söllner, 2020, Hobert, 2019; 2021). However, to our knowledge, no studies have been conducted to systematically investigate and compare the different learning modes of ICAP in chatbot research. Hence, there is a lack of knowledge both in terms of how the four learning modes may inspire different chatbot interaction designs, and in terms of how such different interaction designs perform in terms of engagement and learning outcome.

3. Research Question and Hypotheses

While chatbots are considered a promising technology for educational purposes, there is still limited research on understanding how different chatbot interaction designs may influence learning processes in terms of engagement and learning outcomes. To address this research gap, we conducted an experimental study with educational chatbots to address our main research question:

How does chatbot interaction design for active learning impact user engagement and learning outcomes?

We grounded our study in the theoretical basis of active learning (Section 2.3).

In line with prior research on chatbot interaction design, it was expected that the interaction design influences the users' engagement with chatbots. By enabling learners to become more engaged with an educational chatbot in an interactive conversation (ranging from passive, active, constructive to interactive according to the ICAP framework, Chi & Wylie, 2014), the human-chatbot interaction becomes more intensive. Thus, we assume that the perceived social presence of the chatbot increases with the progressive level of active learning. Overall, a more engaging interaction design may increase the learners' intention to use educational chatbots. Our corresponding hypotheses are:

- **H1**: Designing chatbot interactions to allow learners to become more engaged increases their perceived user engagement.
- **H2**: Designing chatbot interactions to allow learners to become more engaged increases their perceived social presence of educational chatbots.
- **H3**: Designing chatbot interactions to allow learners to become more engaged increases their intention to use educational chatbots.

Additionally, evaluating the effect of an intervention on learning outcomes is essential for educational research. In line with the ICAP assumption (Section 2.3), our hypotheses in terms of subjective perceptions of learning outcome and objective performance-related learning outcome are:

- **H4**: Designing chatbot interactions to allow learners to become more engaged increases subjective learning outcomes.
- **H5**: Designing chatbot interactions to allow learners to become more engaged increases objective learning outcomes.

4. Method

4.1 Research design

In response to our main research question to investigate the effect of chatbot interaction design for active learning and the related hypotheses, we implemented a stand-alone chatbot software app in which the whole learning process takes place in the text-based interaction and set up an experimental study with four conditions reflecting the four ICAP modes with the text-based interaction: passive, active, constructive and interactive. The chatbot implemented a different interaction design for each condition based on the four ICAP modes. We kept all other components (like the user interface or technical implementation) unchanged.

This experimental setting allowed us to research the effect of the interaction design based on the independent variable *ICAP Mode*. In line with our hypotheses, our dependent variables were *User Engagement*, *Social Presence*, *Learning Outcome* (subjective and objective), and *Intention to Use*. Data collection on the dependent variables were conducted with an after-intervention questionnaire.

We also collected qualitative data by including several open-ended questions in the after-intervention questionnaire to gain further insights into the users' perception of the four different interaction designs with a particular focus on user experience and engagement. In addition, data on the participants' prior experience with chatbots and knowledge of phishing emails was gathered in a prior experience questionnaire administered before the chatbot interaction.

We relied on Prolific (https://www.prolific.co/), a commonly used online crowdsourcing platform, to recruit participants for our study. We confined potential participants to English native speakers and required them to use a desktop computer to have comparable settings across our sample. In total, we included 164 participants in our study. As an incentive to participate, the participants were paid in line with the crowdsourcing platform's proposal of a suitable wage.

4.2 Learning Task and Context for the Experiment

To set up the study, we needed a relevant learning context, which is the basis for the chatbot's interaction with the users. We wanted to ensure that the selected topic is general enough that no specific domain knowledge is required and that it can be implemented in a stand-alone chatbotwhere the whole learning process can be covered in the text-based dialog (without needing further learning materials). To achieve this, we selected the learning topic of identifying phishing attacks in email communication. This topic is suitable for such an experiment, as email communication is relevant for both individuals and organizations, and phishing attacks are considered as one of the most common threats for online communications (Bissell et al. 2019; Jampen et al 2020). Furthermore, the learning topic is relatively small and can be divided into several small micro contents (i.e., several strategies to identify phishing attacks).

From a teaching perspective, the learning objective for deploying the chatbot was to enable learners to identify potential phishing attacks in email communication. To this end, the chatbot's main task was to provide instructional material within the dialog. The instructional material was presented in five steps (the top row in Fig. 1). First, the chatbot introduced the learning objective to enable a transparent learning path. Second, the chatbot provided the instructional material on how to identify email attacks in a natural language conversation. To impart the learning content, the chatbot presented a three-step procedure: (1) checking the sender's email address, (2) checking attachments, and (3) checking links included in the email content. Finally, a brief summary of the learning content was provided.

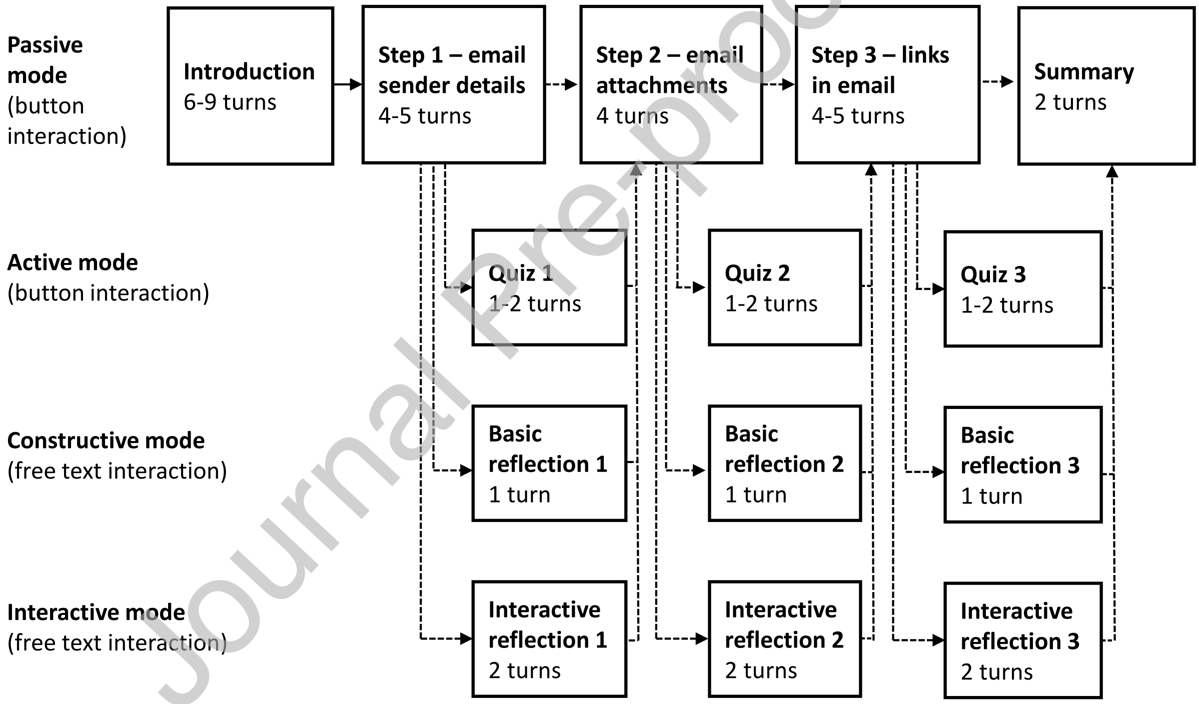
4.3 Chatbot Implementation and Interaction Design

To implement the chatbot used in the experiment, we relied on the Botpress framework (https://botpress.com/), which enabled us to apply different interaction designs without altering the source code. To provide a streamlined user interface for the whole study, we adjusted the chatbot's user interface to match the user interfaces of the study's landing page and questionnaire.

The software implementation and user interface were identical for each version of our chatbot named as *PhiBot*. In all versions, we used button interaction or free text input to allow users to converse with the chatbot depending on the ICAP mode. In passive and active modes, only button interaction with predefined answers was employed. In constructive and interactive engagement, free text input was also requested at several points during the interaction. To ensure the validity of comparisons across the four chatbot designs, we provided the same instructional material for each condition. Variations in the ICAP mode were implemented by including different forms of self-tests during the chatbot interaction.

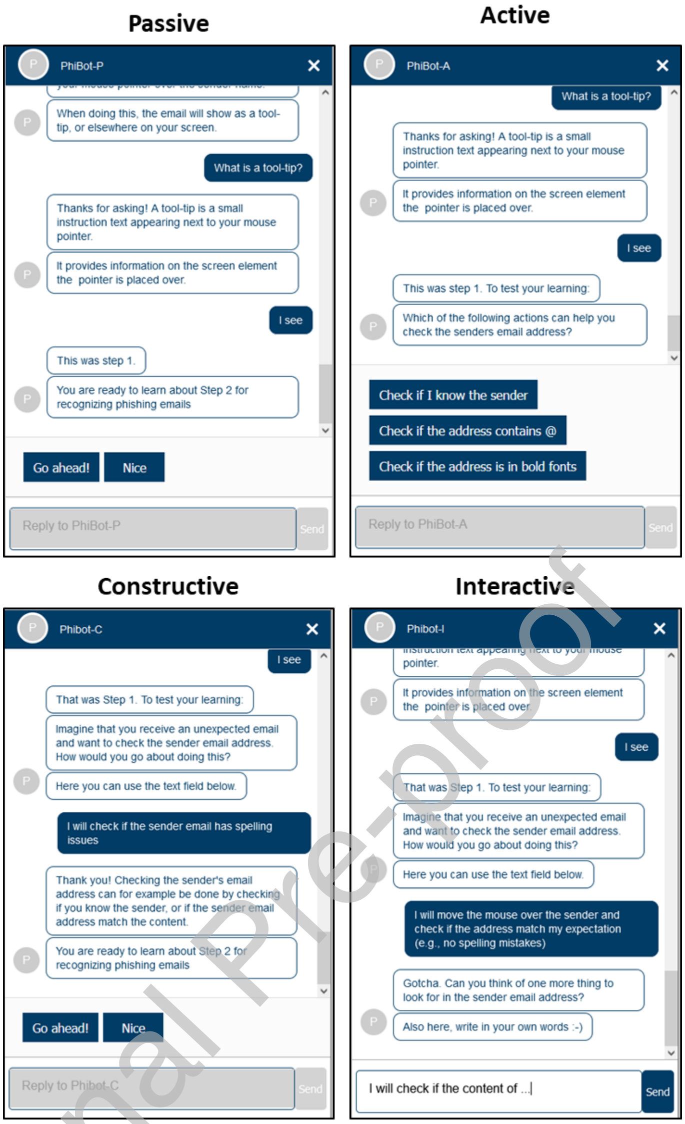
The chatbot instructional material was presented in an introduction, in a sequence of three steps addressing different aspects of identifying phishing emails, and in a brief summary (Fig. 1). In the *passive mode*, the users only received the instructional material. Their main task was reading the messages and using buttons with predefined answers to continue step-by-step through the dialog. In the *active mode*, the instructional material was augmented with a short self-test following each step. This self-test consisted of a one-question quiz to challenge and actively engage the users. The quiz was implemented by providing answer alternatives as predefined buttons, similar to a single-choice quiz often used in teaching settings. To reach the *constructive mode*, the chatbot also challenged the users with a self-test after each step, but this consisted of one reflection question to be answered in the users' own words. By this reflection question, it was expected that the users were stimulated to reflect on the learning content on their own by explaining the previously learned content. Finally, in the *interactive mode,* the chatbot challenged the user with both a reflection question and a subsequent follow-up question as a self-test after each step. For both questions, the users were to respond in their own words. This interactive reflection was intended to motivate the users to engage more deeply in reflection on the instructional material.

The total chatbot interaction included between 20 and 31 turns, depending on the ICAP mode, with exchange of messages back and forth between the chatbot and the user. Fig. 1 summarizes the four different interaction designs by ICAP mode, similar to the original publication of the ICAP framework by Chi and Wylie (2014). The resulting chatbots implementing the four ICAP modes are visualized in Fig. 2.



Passive mode (button interaction)
Introduction 6-9 turns
Step 1 – email sender details 4-5 turns
Step 2 – email attachments 4 turns
Step 3 – links in email 4-5 turns
Summary 2 turns
Active mode (button interaction)
Quiz 1 1-2 turns
Quiz 2 1-2 turns
Quiz 3 1-2 turns
Constructive mode (free text interaction)
Basic reflection 1 1 turn
Basic reflection 2 1 turn
Basic reflection 3 1 turn
Interactive mode (free text interaction)
Interactive reflection 1 2 turns
Interactive reflection 2 2 turns
Interactive reflection 3 2 turns

Fig. 1.Interaction Design by ICAP mode (based on Chi & Wylie, 2014)



Passive

PhiBot-P

When doing this, the email will show as a tool-tip, or elsewhere on your screen.

What is a tool-tip?

Thanks for asking! A tool-tip is a small instruction text appearing next to your mouse pointer.

It provides information on the screen element the pointer is placed over.

I see

This was step 1.

You are ready to learn about Step 2 for recognizing phishing emails

Active

PhiBot-A

What is a tool-tip?

Thanks for asking! A tool-tip is a small instruction text appearing next to your mouse pointer.

It provides information on the screen element the pointer is placed over.

I see

This was step 1. To test your learning:

Which of the following actions can help you check the senders email address?

Check if I know the sender
Check if the address contains @
Check if the address is in bold fonts
Reply to PhiBot-A
Send

Constructive

Phibot-C

I see

That was Step 1. To test your learning:

Imagine that you receive an unexpected email and want to check the sender email address. How would you go about doing this?

Here you can use the text field below.

I will check if the sender email has spelling issues

Thank you! Checking the sender's email address can for example be done by checking if you know the sender, or if the sender email address match the content.

You are ready to learn about Step 2 for recognizing phishing emails

Interactive

Phibot-I

pointer.

It provides information on the screen element the pointer is placed over.

I see

That was Step 1. To test your learning:

Imagine that you receive an unexpected email and want to check the sender email address. How would you go about doing this?

Here you can use the text field below.

I will move the mouse over the sender and check if the address match my expectation (e.g., no spelling mistakes)

Gotcha. Can you think of one more thing to look for in the sender email address?

Also here, write in your own words :-)

I will check if the content of ...|

Fig. 2. Example screenshots visualizing the four different ICAP modes

4.4 Measurement Instruments

4.4.1 Prior Experience with Chatbots and Knowledge of Phishing Emails

Before the chatbot interaction, we included a short prior experience questionnaire focusing on the participants' prior experience with chatbots and knowledge of phishing emails. To this end, we used three 5-point Likert scale items for each aspect (e.g., “I frequently use chatbots.” and “I have the knowledge to reliably identify phishing emails.”).

4.4.2 User Engagement, Social Presence, and Intention to Use

As part of the after-intervention questionnaire, we measured the dependent variables User Engagement, Social Presence, and Intention to Use. User Engagement was measured using the short form of the User Engagement Scale (UES-SF) by O'Brien, Cairns and Hall (2018), covering the four subscales: Focused Attention, Perceived Usability, Aesthetic Appeal, and Reward. Focused Attention concerns the feeling of absorption in the interaction, Perceived Usability concerns negative affect in result of the interaction, Aesthetic Appeal concerns the attractiveness and appeal of the user interface, while Reward concerns the interaction being perceived as worthwhile and interesting.

Social Presence was measured using four items based on Laban and Araujo (2020). Our measure of Intention to Use included two items (e.g., “If I get the chance in the future, I would like to use this type of chatbot to learn new content”). The latter measure was inspired by the Technology Acceptance Model (TAM) (Venkatesh & Davis, 2000) and formed to fit the specific context of our study. For all scales, Cronbach Alphas had acceptable levels (> .7). All measurement instruments are presented in Appendix A.

Additionally, we asked the participants for qualitative feedback on their experience (“In your own words, how was your experience with the chatbot you just used? Please provide a short description of your experience with the chatbot”).

4.4.3 Learning Outcome (Subjective and Objective)

To get insights into the learning outcome after completing the chatbot interaction, we asked the participants to rate their perceived subjective learning outcome using three homegrown items (e.g., “The chatbot strengthened my ability to reliably identify phishing emails”. For a full overview of the items, see Appendix A).

We further developed a short summative assessment consisting of four exercises to get insights into the actual, measurable Objective Learning Outcome. In each exercise, the participants got access to a screenshot of an email and asked to provide their estimation whether the email looks like a possible phishing attack and the reasoning behind. To get a summative score of the assessment, we assigned one point for the correct answer and another point for an understandable explanation. In total, the participant could get up to 8 points if they succeeded in all four exercises.

An example task of the assessment test is included in Appendix B.

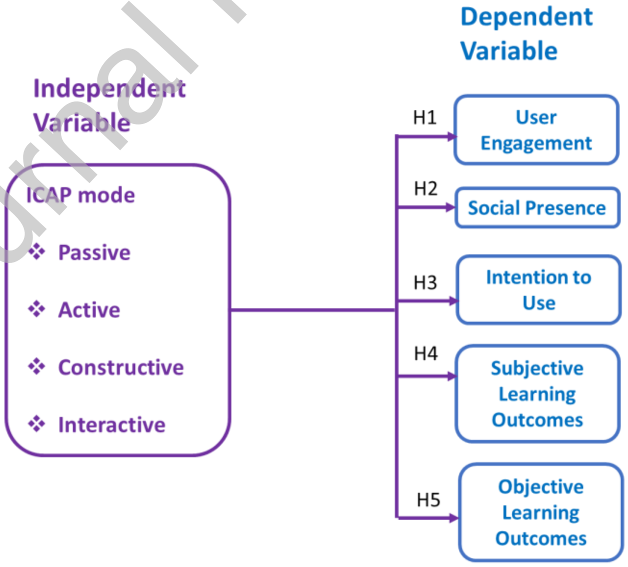


Fig. 3.Overview of the independent and dependent variables of the study

4.5 Data Collection and Experimental Setup

Based on the four implemented variations of the chatbot and the measurement instruments (see Fig. 3), we set up the experiment as visualized in Fig. 4. After the introduction to the study and informed consent, the participants were randomly allocated to one of the four experimental conditions. We then asked the participants to fill out the prior experience questionnaire focusing on their experience with chatbots and phishing emails. Afterward, each participant interacted with the chatbot version that constituted the relevant ICAP mode. After completing the chatbot interaction, the participants were redirected to the after-intervention questionnaire. In addition to the data collected using the questionnaires, we captured the dialogues to analyse and verify the participants' chatbot interactions.

Before running the experiment, we pilot-tested the full setup iteratively by recruiting six participants using the same online crowdsourcing platform. Based on the results, we improved the questionnaires and chatbot implementations before running the experiment and data collection.

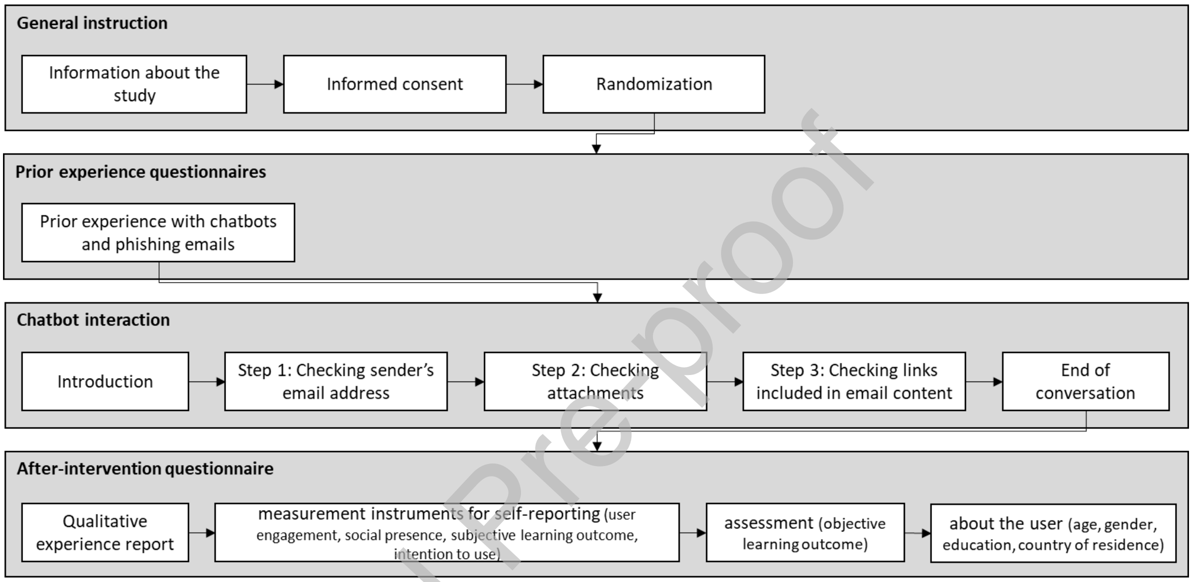

General instruction
Information about the study
Informed consent
Randomization
Prior experience questionnaires
Prior experience with chatbots and phishing emails
Chatbot interaction
Introduction
Step 1: Checking sender's email address
Step 2: Checking attachments
Step 3: Checking links included in email content
End of conversation
After-intervention questionnaire
Qualitative experience report
measurement instruments for self-reporting (user engagement, social presence, subjective learning outcome, intention to use)
assessment (objective learning outcome)
about the user (age, gender, education, country of residence)

Fig. 4. Summary of the experimental setup

4.6 Research ethics

The research study was set up according to the institution's guidelines and did not pose unnecessary demands on the participants. Because the recruitment process was outsourced to a commonly used online crowdsourcing platform, the participants were kept anonymous to the researchers. The whole data collection was set up to be fully anonymous to protect privacy. The participation was voluntary following informed consent. The participants could terminate the study at any time without the need to provide any reason.

5. Results

5.1 Descriptive statistics

The 164 participants recruited for the experiment were between 19 and 74 years old (mean = 38, SD = 14). 70 of them were female, 93 male, one preferred not to say. In line with our prerequisite to recruit native English speakers from the online crowdsourcing platform, the participants' countries of residence were mostly from the UK and the US (63% and 29 %, respectively), followed by other countries with less than 5%. The participants' educational background was mostly higher education (81%). 19 % had high school as their highest level of education, and one finished elementary school.

In terms of chatbot experience prior to the study, it was found to be a medium level of 3.3 on a 5-point Likert scale (SD = 0.9). The participants self-estimated their competence in identifying phishing emails to be high, with a mean value of 4.2 (SD = .8). This indicated that the participants already had good prior knowledge before the study. ANOVA tests showed no significant differences between the four experimental groups (Phishing Competence: F(3,160) = 1.4, p = .24; Chatbot Experience: F(3,160) = 1.0, p = .39.).

The time spent on the chatbot interaction varied markedly between the conditions. We measured the time the participants spent on the instructional material, the self-tests, and the aggregated total time as visualized in Fig. 5.

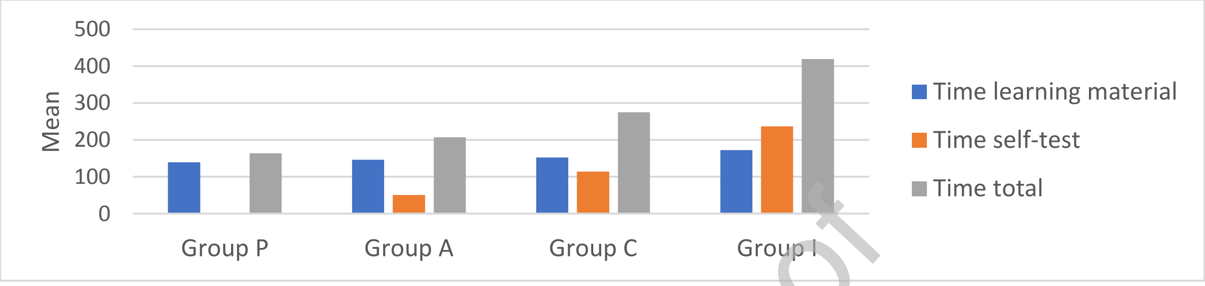


Fig. 5 Visualization of times spent on learning material, self-test and the whole interaction

No study hypotheses were associated with the time spent on the chatbot interaction. However, to verify that different ICAP modes indeed resulted in different forms of engagement, we ran separate one-way ANOVA tests. These tests did not reveal any significant differences between the four ICAP modes on the time spent on the instructional material, varying only slightly between 2.3 minutes and 2.9 minutes. This was expected as the instructional material was implemented identically for all groups. However, the effect of the four ICAP modes on the time the participants spent on the self-tests and the total time in the interaction was shown to be significant (Table 1). Note that in the passive mode, no self-test was implemented, and that for Group C and Group I manual typing was required to solve the tests (see Method, Section 4.3). Tukey post-hoc analyses and corresponding t-tests revealed significant differences (p < .001) in the time total in the following pairs: Group P-I, A-I and C-I (p<.001) and Group P-C (p=.006). Regarding time on self-test, the same pattern of post-hoc pairwise significant group differences was observed.

Table 1: Mean (standard deviation) of the three temporal measures (in seconds)across the four experimental groups (P = Passive; A = Active; C=Constructive; I = Interactive)

	Group P (n = 36)	**Group A (n = 45)**	**Group C (n = 41)**	**Group I (n = 41)**	**F**	**p**	**eta-sq**
Time total	164 (53)	207 (78)	275 (131)	419 (242)	23.30 (3,159)	<.001	.35
Time instructional material	139 (47)	146 (59)	152 (90)	172 (90)	1.46 (3,159)	.23	.03
Time self-test	NA	51 (21)	114 (48)	237 (171)	36.97 (2,124)	<.001	.37

5.2 Impact on User Engagement, Social Presence, and Intention to use

As described in Section 4.4.2 and Fig. 2, different measures (DVs) were taken. Scores were analysed by each experimental group, reflecting the four ICAP modes.

To investigate the impact of the four ICAP modes on User Engagement, Social Presence and Intention to Use (H1-H3), separate one-way ANOVA tests were conducted. No significant differences were observed between the groups. An overview of the descriptive and ANOVA statistics is shown in Table 2.

Table 2: Mean (standard deviation) and ANOVA analysis results of the four factors of User Engagement (UE), Social Presence, and Intention to Use across the four experimental groups.

	Group P (n = 37)	**Group A (n = 45)**	**Group C (n = 41)**	**Group I (n = 41)**	**F (3, 160)**	**p**	**eta-sq**
UE: Focused Attention	2.8 (.7)	3.1 (1.0)	3.0 (0.9)	3.3 (1.1)	1.81	.15	.03
UE: Perceived Usability	4.4 (.6)	4.4 (.6)	4.2 (.7)	4.3 (.8)	.76	.52	.01
UE: Aesthetic Appeal	3.2(.8)	3.2 (.9)	3.2 (.7)	3.4 (.8)	.37	.78	.01
UE: Reward	3.5 (.9)	3.8 (1.1)	3.9 (.7)	3.8 (.9)	1.22	.30	.02
Social Presence	3.4 (1.1)	3.4 (1.3)	3.4 (1.1)	3.7 (1.1)	.60	.61	.01
Intention to Use	3.4 (.9)	3.6 (1.2)	3.8 (.9)	3.7 (1.0)	.95	.42	.02

5.3 Impact on Subjective and Objective Learning Outcomes

Subjective Learning Outcome was measured by a dedicated self-reported instrument. To investigate the impact of the four ICAP modes on subjective learning outcome (H4), a one-way ANOVA test was conducted. Significant differences were observed between the groups in line with the study hypothesis. Specifically, the lowest score was observed for the group with the passive mode, and the highest scores were observed for the group with the interactive mode. This is also reflected in an additional Tukey post-hoc analysis and t-test, which confirmed a significant difference ($p < .015$) between the subjective learning outcome of the groups P and I ($t(76) = 2.744$, $p = .008$).

An overview of the group means and standard deviations, as well as ANOVA statistics, is provided in Table 3. Because non-normality was observed in the distribution of the Subjective Learning Outcome data, this analysis was replicated by the non-parametric Kruskal-Wallis test, which showed similar and significant group differences.

Objective Learning Outcome was measured by a set of assessment exercises where the participants could obtain a maximum of eight points (see Appendix B). To investigate the impact of the four ICAP modes on objective learning outcome (H5), a one-way ANOVA test was conducted. Objective learning outcome scores were fairly high with an average of 6.2 across all four groups. No significant differences were observed between the groups. An overview of the group means and standard deviations, as well as ANOVA statistics, is provided in Table 3.

Table 3: Statistical analysis of the Subjective and Objective Learning Outcome across the four experimental groups

	Group P (n = 37)	Group A (n = 45)	Group C (n = 41)	Group I (n = 41)	F (3, 160)	p	eta-sq
Subjective Learning Outcome	**3.5 (1.3)**	**3.9 (1.1)**	**4.1 (.7)**	**4.2 (1.0)**	**3.40**	**<.05**	**.06**
Objective Learning Outcome	6.2 (1.5)	6.2 (1.5)	6.2 (1.7)	6.2 (1.9)	.01	.99	.00

5.4 Exploring Qualitative Experience Reports on the Chatbot Interaction

As part of the after-intervention questionnaire, the participants were asked to report on their experienced perception of the chatbot interaction by asking them “In your own words, how was your experience with the chatbot you just used?”. After coding the participants’ responses and grouping them into main themes, we looked at the themes in relation to the four ICAP modes. The themes were evenly distributed across the four groups, and we could not identify a specific pattern. Hence, we have no reason to suspect group differences in the qualitative responses. In the following, we report on each theme by citing the relevant quotes.

5.4.1 User Experience

Most of the statements within the reports focused on user experience. 107 mentioned positive aspects of the chatbot interaction with regard to the user experience, 10 stated a negative experience.

More than 50% of the participants commenting on user experience mentioned that the overall experience with the chatbot was positive in general (e.g., “My experience with the chatbot was good” P87, Group A). Others focused on more specific aspects like the responsiveness of the human-chatbot interaction (like “Pleasant chat. No problems. Answered my questions promptly.” P41, Group P) or the chatbot’s tone of voice within the conversation (e.g., “The bot's tone was polite and helpful and it provided useful information.” P87, Group A). Nine participants stated that the conversation with the chatbot was fun or enjoyable (e.g., “I found my experience with the chatbot pleasurable and fun.” P67, Group I).

Additionally, a few people criticized some aspects of the interaction or suggested improvements in the responsiveness and the chatbot’s tone of voice. For example, one person pointed out a dislike for the chatbot’s informal phrasing (“I didn't like when it said, 'lovely' 'goodie' 'phew', I would have preferred the conversation to be more professional.” P157, Group P). Another suggested to reduce the conversation speed (“It was a pretty cool experience actually. I think the bot may have gone too fast when posting information and questions but overall, it was enjoyable.” P27. Group C).

5.4.2 Training Content

The learning content was the second most frequently addressed topic. 85 participants commented on this, 91% of them in a positive manner.

The selected learning content of phishing emails was considered as useful by most participants. The depth of the information and selection of the training content was considered adequate (e.g., "The chatbot gave me a lot of good information on how to detect phishing emails". P14, Group A.). Additionally, the clarity and structure of the instruction materials was positively mentioned (e.g., "Quick and easy to understand. Responses where clear and concise and written in a good format." P105, Group P.)

In only eight statements, negative aspects were addressed. One specific possibility for improvement was pointed out by one person noting that some parts of the materials were too generic ("[…] some of the information it was giving out was not applicable in all situations" P155, Group P.).

5.4.3 Reflections on Learning

Twenty-six statements addressed the learning process or outcome of the chatbot interaction. Even despite the high level of prior self-estimated competence (see subsection "Descriptive overview" above), 20 people stated they learned something new about phishing emails. It was further mentioned that the chatbot was also useful for repeating previously learned topics (e.g., "Useful. It covered most of what I know already about phishing emails, but there were a couple of useful bits of information and always good to refresh the memory." P23, Group P).

In line with the participants' high self-estimated competence in identifying phishing emails, two persons complained about the too low level of the learning materials and a corresponding missing learning success (e.g., "I did already know about the topic so I feel like it would be more suitable to teach someone who is new to computers and technology" P154, Group C).

6. Discussion

In this section, we first discuss the five hypotheses (H1-H5) with reference to the empirical results, then we explore the implications for the theoretical and practical issues pertaining to the design and usage of educational chatbots. Next, we reflect on the limitations of our study to draw insights for future research along this line of inquiry.

6.1 Effect of Chatbot Interaction Design for Active Learning

Participants of all four experimental groups (P=Passive, A=Active, C=Constructive, I=Interactive) interacted with the respective chatbot versions at different levels of engagement (Table 1). As expected, the self-tests time were significantly longer for the more engaged groups (Section 5.1) as the users needed more time to complete the more demanding self-tests. Nonetheless, the significant differences in the interaction times were *not* translated into corresponding differences in the User Engagement (H1), Social Presence (H2), or Intention to Use (H3). Contrary to our hypotheses, no significant differences in these constructs were detected (Section 5.2). In fact, the mean ratings per constructs and constituting factors were comparable across the four groups (Table 3).

The short interaction episode, ranging from about 2.73 minutes (Group P) to 6.98 minutes (Group I) (Table 2), might explain why no significant differences in Focussed Attention could be found. For instance, in their study on changes in sustained attention over age, Fortenbaugh et al. (2015) applied the 4-minute gradual-onset continual performance task (grad-CPT) with a large number of participants aged 10-70 years old. The duration was chosen by those authors as participants were expected to stay focused within it. While the results of that study are irrelevant to our work presented here, it is relevant to point out that the task completion time of our experimental study was not unusually short, but it might be too short for revealing the effect on Focused Attention (Table 2)

Regarding Perceived Usability, the simple interaction mechanism of the chatbot rendered it highly usable, as evident in both the quantitative (4.3 out of 5; Table 3) and qualitative results (e.g. "… experience with the chatbot pleasurable", P67, Group I). The simple look of the text-based interface of the chatbot with the icon "P" standing for PhiBot, which was the same across the four groups, was generally perceived with a medium level of Aesthetic Appeal. Clearly, the length of interaction time may not have played any significant role in the perception of visual attractiveness, as epitomized by the finding of Lindgaard et al., (2006) that the initial impression of a user interface could already be formed in the first 50 milliseconds of interaction. In addition, the highly significant correlation between Perceived Usability and Aesthetic Appeal (r = 0.298, p<.001) resonates with the findings of the existing research on the relations between usability and beauty (e.g., Hassenzahl, 2004; Tractinsky, Katz, & Ikar, 2000).

Concerning Reward, the four groups had a mean rating between 3.5 and 3.9, indicating that in general the participants found the interaction rewarding but not highly so. The medium level of Reward could be attributed to the *form* and *content* of the learning scenario. Specifically, some participants of Group P and Group A gave mixed comments on the button-based interaction, which was perceived as easy, fast and concise on the one hand (e.g., "Quick and easy to understand. Responses where clear and concise and written in a good format." P105, Group P) but as limiting and even meaningless on the other hand (e.g., "… my options were limited and so I could not manipulate the conversation in any meaningful way…" P162, Group P). The former could enhance the perception of Reward whereas the latter could undermine it. Furthermore, curiosity and interest are critical elements determining the extent to which the learning process is perceived as rewarding and worthwhile (e.g., Litman, 2005; Peterson & Hidi, 2019). In fact, some participants in all four groups commented that they were already informed about the topic of phishing. This might dampen their curiosity and thus sense of rewarding. Moreover, as part of their ongoing work on user engagement, O'Brien et al. (2020) recently investigated factors influencing user engagement in information retrieval tasks. While their findings that the task topic and perceived task difficulty played a significant role in user engagement were not surprising, the intriguing result was that effort expended in the tasks had a negative effect on user engagement. This observation was not supported by our result: the extra effort in terms of time spent (Table 2) had no effect on User Engagement across the four groups.

The notion of Social Presence in the context of chatbot interaction implies that a user perceives the chatbot as a partner being able to engage in active dialogue and to demonstrate understanding as well as good will to support the user (Araujo, 2018; Go & Sundar, 2019). Our results showed that the most active dialogue in Group I led to the highest rating of Social Presence (Table 3), but it was not significantly different from the other groups. An illustrative qualitative response from Group I may shed light on the issue: "The experience was informative, although the responses were clearly not the same as if interacting with a real person" (P144, Group I).

As there were no significant differences in any of the four factors of User Engagement Scale or Social Presence, it was logical to find that the trend was the same for Intention to Use. Indeed, a linear regression indicated that the four factors of the User Engagement Scale and Social Presence were significant predictors of Intention to Use (R^2=0.667, F(4,159) = 79.8, p<.001). The rating for Intention to Use was on average moderate, regardless of the groups, suggesting that participants had the intention, albeit not particularly strong, to adopt the chatbot for learning new content.

Overall, H1, H2 and H3 were rejected based on the results of our experimental study with learning tasks constructed for the experiment. To further investigate whether this rejection also holds for other chatbot-based learning settings, additional studies with other and more elaborate or contextually embedded learning tasks might be needed (see Subsection 6.3 for further details).

Regarding H4 and H5 on learning outcomes, some intriguing results were observed. As shown in Table 4, there were differences in the Perceived (self-assessed) Knowledge Gain across the four groups and Group I's was significantly higher than Group P's whereas the mean scores of the four groups were the same in the Objective Learning Outcome. There was no correlation between the subjective and objective assessment (r = -0.076, N = 164, p = .755). This inconsistency corroborates one of the observations identified by Sitzmann et al. (2010) in their meta-analysis on self-assessment of knowledge. They argued against the reliance on self-assessment as a sole indicator of learning performance. Furthermore, Sitzmann et al. (2010) concluded that "self-assessed knowledge is generally more useful as an indicator of how learners feel about a course than as indicator of how much they learned from it" (p.180). Arguing along this line, it implied that Group I felt much more positive about the learning experience with the chatbot than Group P did. Nevertheless, an explanation for no difference in objective performance was the ceiling effect. As indicated by the prior experience questionnaire and after-intervention qualitative feedback, many of the participants already knew the topic of phishing before the study.

6.2 Implications for Theory and Practice

The theoretical model underpinning our empirical study was primarily the ICAP framework postulated by Chi and Wylie (2014) (Section 2.3). Specifically, we have translated the progressively active learning modes into different chatbot interaction designs, from the close-ended button-based to the mixed closed- and open-ended dialogical exchanges between the user and chatbot. Although the objective learning outcomes could not reflect the effects of the varied active learning designs in our experimental learning setting, thanks to the ceiling effect, the significant difference in the subjective self-assessed knowledge gain provided insights into the potential of the ICAP framework.

According to the established pedagogical theories, social constructivism proved relevant for enhancing learning (e.g., Amineh & Asl, 2015). In other words, learners should be enabled to construct knowledge from instructional materials in a socially inspiring environment. The ICAP framework aligns with the social constructivist approaches and instantiates each of the four modes with concrete cases. Furthermore, breaking down instructional materials into small chunks to allow learners to have sufficient space and time for reflection (Brockbank, McGill & Beech, 2017) – one of the critical pedagogical principles of the ICAP framework – is a feature highly compatible with micro-learning (Section 2.4) that chatbots can effectively support (cf. Sarosa, Kusumawadhani, Suyono, & Azis, 2021). In our study, we realised the technique of "self-explaining" (Chi & Wylie, 2014) for the Constructive and Interactive modes where the participants were asked to elaborate their responses. How the other two ICAP-based techniques, namely note-taking and concept mapping, can be implemented in an educational chatbot entails more systematic research effort. Even though our learning setting might be limited in terms of constructive and interactive modes, we present a first operationalization of the four ICAP modes for conversational interaction design in chatbot-supported learning processes. Thus, we present first insights into designing chatbots for active learning. We foresee further research on different conversational interaction designs and pedagogical learning processes in future studies to gain further knowledge of this area.

Regarding the practical implications, the non-significant differences in User Engagement and Social Presence suggest the need for modifying the chatbot interfaces. Specifically, the interface design of

PhiBot was minimalistic with a simple icon as we wanted to focus on the text-based conversational interaction design. To enhance aesthetic appeal, chatbots with humanlike or customisable avatars meeting users' preference (e.g., Schöbel, Janson, & Mishra, 2019) can be deployed. While the conversational style used in PhiBot was anthropomorphic, conveying the sense of social presence, the medium ratings and qualitative responses indicated that the Uncanny Valley effect (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2019; Mori, MacDorman, & Kageki, 2012) might be aroused. Some participants commented on what they saw as inappropriate uses of casual remarks such as "phew" and "lovely". Given the educational (utilitarian) goal of PhiBot, making it sound too fun-loving or friendly could invoke the feeling of incompatibility unless the chatbot is introduced in a way that sets the right expectations at the outset. This compatibility issue needs to be addressed in the future chatbot user interface design.

In order to actually transfer the scientific findings on the chatbot interaction design into learning practice, there are further steps that need to be considered. This appears to be particularly important, as only a few educational chatbots are known that have been used for learning purposes in the education sector on a long-term and continuous basis (e.g., Hobert, 2023). Important steps to consider when introducing chatbots in educational settings include technical barriers, especially in school and university contexts. The use of cloud-based chatbot implementations is not always permitted. This is due, among other things, to particularly high data protection requirements in the education sector (e.g., due to the GDPR in the EU). The EU AI Act that targets the regulation of AI technology might have implications. In addition to that, ethical implications must be considered in order to offer learners suitable learning conditions that do not lead to any disadvantages for individual groups. Here, further investigations are needed to understand if different groups (e.g., age, existing prior knowledge, learning difficulties) impact the effects of educational chatbots.

Overall, weighing the benefits observed and considering the lessons learned from our study, we expect that the operationalization of the ICAP framework to design active text-based conversations has the potential to enhance educational chatbot design. We expect it to be a worthwhile direction for future research.

6.3 Limitations and Future Work

Like most if not all experimental work, ours has also been undesirably affected by limitations, including the selection of learning topic, duration of the interaction episode, and Uncanny Valley effect. We reflect on each of these limitations and infer implications for future work.

Prior to the main study, we conducted pilot tests with six participants to check for issues in the chatbot user interface design and questionnaires. Results from the pilot study suggested no concerns of a ceiling effect. Nevertheless, in the main study the ceiling effect did reach as there were no differences among the four experimental groups. In hindsight, our choice of phishing seemed not ideal, because many of the participants were familiar with it. While micro-learning implies bitesize units of learning activity, the short interaction episode might not be able to fully utilize the benefits of the ICAP framework. Hence, for future work, participatory design activities should be conducted to identify needs and preferences from the target group on topics, and complexity as well as duration of learning activities with chatbots. Some participants, albeit the number was small, regarded some chatbot dialogues of PhiBot as unnatural. As mentioned above, the conversational style of the chatbot should match the expectation set for its usage and goal, eliminating the risk of arousing the unpleasant affective responses in users.

Furthermore, we did not administer any retention test to check the effectiveness of the intervention, something which would have required us to instantiate participant tracking. For future work when complex learning activities are involved, longitudinal studies to track the impact of chatbot-based active learning would be recommendable, aligning with the concepts of the ICAP framework.

In addition, future studies should critically investigate whether the generalization of our hypothesis testing results is possible and whether the rejection of most hypotheses in our experimental study is really due to the fact that active learning in chatbots does not have a sufficient effect. Possibly, future interaction designs with adapted learning tasks and more intense or extensive operationalizations of the four learning modes may lead to stronger effects. Thus, there is a need for further research here, which could be based on our first proposal of an operationalization of ICAP modes for designing conversational interactions in educational chatbots. Finally, it is to be expected that the rapid recent developments in chatbot technology, triggered by the widespread availability of large language models, will also impact the use of chatbots in education (e.g., Kasneci et al. 2023 and Rudolph, Tan, and Tan, 2023). In this regard, it should be noted that currently, many large language models sometimes have problems with ensuring response quality (e.g., Fergus, Botha, and Ostovar, 2023). Nevertheless, it is to be expected that technical improvements will lead to quality improvements here. This appears to be particularly important in the educational context so that learners do not learn from incorrect or biased content. However, further studies will be necessary to keep up with the technical developments of the chatbot technology.In spite of the limitations, our study supports the potential of chatbots as a viable option for presenting learning materials in accessible micro-units, and how active learning may be enabled through chatbot interaction design.

7 Conclusion

Educational chatbots introduce a promising approach to provide instructional content and facilitate active learning processes. Knowledge on suitable chatbot interaction designs is, however, limited. In our study, we grounded our research in the ICAP framework of Chi and Wylie (2014) and derived four chatbot interaction designs for active learning. With an experimental study, we investigated how these chatbot interaction designs impacted user engagement and learning outcomes. Our results suggested that chatbot interaction designs implementing an interactive ICAP mode might be superior regarding the time users spent within their learning processes and subjective learning outcome. We could, however, not support our other hypotheses, especially regarding objective learning outcome. Nevertheless, we assume that the effects of interactive chatbot interaction designs might be positive when providing instructional content. To get further insights on how to create suitable interaction designs for educational chatbots and address the limitations of our study (particularly the potential ceiling effect), future research studies are needed.

Author contributions

[removed for review]

Declaration of Competing Interest

The authors have no competing interests to declare for this study.

Funding

[removed for review]

Appendix A: Questionnaires

Table 4 Overview over the prior experience questionnaire

Topic	Questions
Prior experience with chatbots	I frequently use chatbots. I use chatbots when this is provided as a service alternative. I have used chatbots for a long time.
Prior experience with phishing emails	I know what phishing emails are. I have the knowledge to reliably identify phishing emails. I understand how phishing emails work.

Note: Response alternatives from 'strongly disagree' (1) to 'strongly agree' (5)

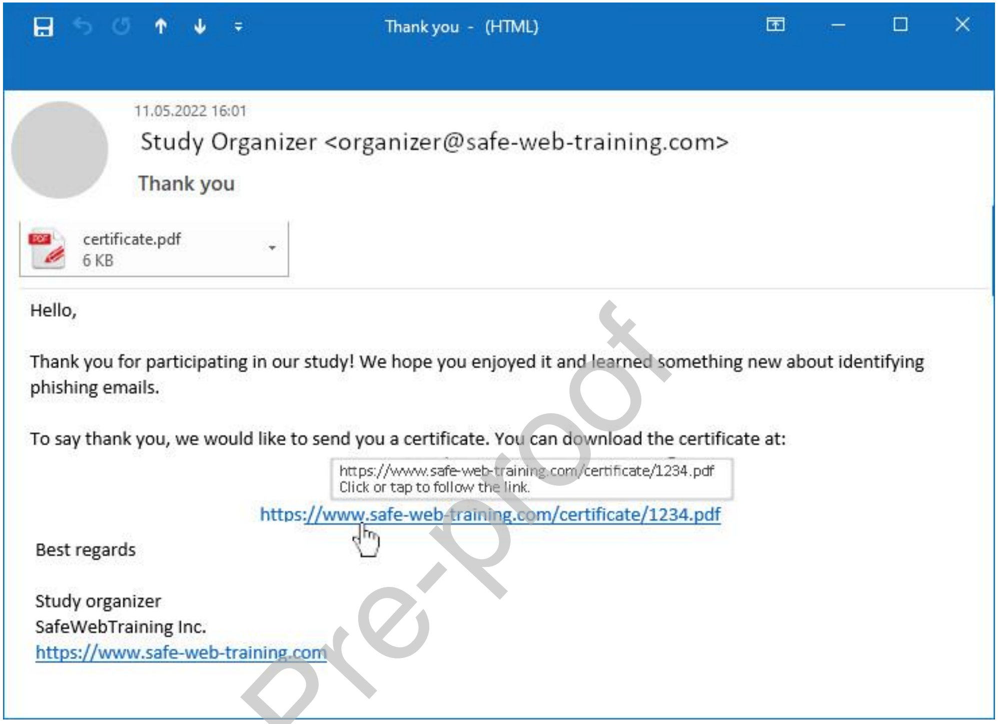
Table 5 Overview over the after-intervention questionnaire

Topic	Questions
Qualitative experience report	In your own words, how was your experience with the chatbot you just used?
User Engagement – Focused Attention based on O'Brien et al. (2018).	I lost myself in the interaction with the chatbot. The time I spent using the chatbot just slipped away. I was absorbed in the interaction with the chatbot.
User Engagement – Perceived Usability based on O'Brien et al. (2018).	I felt frustrated while using this chatbot. I found this chatbot confusing to use. Using this chatbot was taxing.
User Engagement – Aesthetic Appeal based on O'Brien et al. (2018).	This chatbot was attractive. This chatbot was aesthetically appealing. This chatbot appealed to my senses.
User Engagement – Reward based on O'Brien et al. (2018).	Using the chatbot was worthwhile. My experience with the chatbot was rewarding. I felt interested in the experience with the chatbot.
Social Presence based on Laban Araujo, 2019.	I felt like I was engaged in an active dialogue with the chatbot. My interaction with the chatbot felt like a back and forth conversation. I felt as if the chatbot and I were involved in a mutual task. The chatbot followed up on my activity in a good way.
Subjective Learning Outcome	The chatbot dialogue improved my knowledge of what phishing emails are The chatbot strengthened my ability to reliably identify phishing emails. After using the chatbot, I understand better how phishing emails work.
Intention to Use inspired by the Technology Acceptance Model (Venkatesh & Davis, 2000)	If I get the chance in the future, I would like to use this type of chatbot to learn new content. I would prefer to use chatbots as part of learning processes.

Note: Qualitative experience report with free text response. For all other questions, response alternatives from 'strongly disagree' (1) to 'strongly agree' (5)

Appendix B: Example Task of Objective Learning Outcome Assessment

Table 6 One of the tasks used to assess Objective Learning Outcome, presented as an example.

Topic	**Task**
Objective Learning Outcome (example assessment task)	Please consider this image of an email.  11.05.2022 16:01 Study Organizer <organizer@safe-web-training.com> Thank you certificate.pdf 6 KB Hello, Thank you for participating in our study! We hope you enjoyed it and learned something new about identifying phishing emails. To say thank you, we would like to send you a certificate. You can download the certificate at: https://www.safe-web-training.com/certificate/1234.pdf Click or tap to follow the link. https://www.safe-web-training.com/certificate/1234.pdf Best regards Study organizer SafeWebTraining Inc. https://www.safe-web-training.com From what you have learnt from the chatbot, does this email look like a possible phishing email? • Yes, it looks like a possible phishing email • No, it does not look like a possible phishing email Why did you give this answer? (please give your main reason)

References

Amineh, R. J., & Asl, H. D. (2015). Review of constructivism and social constructivism. *Journal of Social Sciences, Literature and Languages, 1*(1), 9-16.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183-189.

Bissell K., LaSalle R. M., Cin P. D. (2019). Accenture's ninth annual cost of cybercrime study: Unlocking the value of improved cybersecurity protection. Technical Report, Accenture. https://www.accenture.com/us-en/insights/security/cost-cybercrime-study

Brockbank, A., McGill, I., & Beech, N. (2017). Reflective learning in practice. In *Reflective Learning in Practice* (pp. 18-28). Routledge.

Chang, C. Y., Hwang, G. J., & Gau, M. L. (2021). Promoting students' learning achievement and self-efficacy: A mobile chatbot approach for nursing training. *British Journal of Educational Technology, 53*(1), 171-188.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219-243.

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems, 92*, 539-548.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811-817.

Feine, J., Morana, S., & Maedche, A. (2020). Designing interactive chatbot development systems. In *Proceedings of ICIS 2020* (paper no. 1870).

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education*, 100(4), 1672–1675.

Fidan, M., & Gencel, N. (2022). Supporting the instructional videos with chatbot and peer feedback mechanisms in online learning: The effects on learning performance and intrinsic motivation. *Journal of Educational Computing Research*. https://doi.org/10.1177/07356331221077901.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health, 4*(2), e7785.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *ACM interactions, 24*(4), 38-42.

Følstad, A., & Brandtzaeg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience, 5*(1), 1-14.

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2018). Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *Proceedings of the International Conference on Internet Science – INSCI 2018* (pp. 145-156). Cham, Switzerland: Springer.

Følstad, A., Araujo, T., Law, E. L. C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., ... & Luger, E. (2021). Future directions for chatbot research: an interdisciplinary research agenda. *Computing, 103*(12), 2915-2942.

Fortenbaugh F. C., DeGutis J., Germine L., Wilmer J. B., Grosso M., Russo K., Esterman M. (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. *Psychological Science. 26*(9), 1497–1510.

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology, 10*(3), 8-14.

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.

Fryer, L. K., Nakao, K., & Thompson, A. (2019). Chatbot learning partners: Connecting learning experiences, interest and competence. *Computers in Human Behavior, 93*, 279-289.

Gabrielli, S., Rizzi, S., Bassi, G., Carbone, S., Maimone, R., Marchesoni, M., & Forti, S. (2021). Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. *JMIR mHealth and uHealth, 9*(5), e27965.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304-316.

Goel, A. K., & Polepeddi, L. (2018). Jill Watson: A virtual teaching assistant for online education. In *Learning Engineering for Online Education* (pp. 120-143). Routledge.

Grand View Research (2022). Education technology market size, share & trends analysis report. Technical report, Grand View Research. https://www.grandviewresearch.com/industry-analysis/education-technology-market

Hansch, A., Hillers, L., McConachie, K., Newman, C., Schildhauer, T. & Schmidt, J. P. (2015). Video and Online Learning: Critical Reflections and Findings from the Field. HIIG Discussion Paper Series No. 2015-02.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human–Computer Interaction, 19*(4), 319-349.

Haugeland, I. K. F., Følstad, A., Taylor, C., & Bjørkli, C. A. (2022). Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies, 161*, 102788.

He, L., Basar, E., Wiers, R. W., Antheunis, M. L., & Krahmer, E. (2022). Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health, 22*(1), 1-14.

Heller, B., Proctor, M., Mah, D., Jewell, L., & Cheung, B. (2005). Freudbot: An investigation of chatbot technology in distance education. In *EdMedia+ Innovate Learning* (pp. 3913-3918). Association for the Advancement of Computing in Education (AACE).

Hobert, S. (2019). Say hello to 'Coding Tutor'! Design and evaluation of a chatbot-based learning system supporting students to learn to program. In *Proceedings of the International Conference on Information Systems – ICIS 2019* (paper no. 2661).

Hobert, S. (2021). Individualized learning patterns require individualized conversations–Data-driven insights from the field on how chatbots instruct students in solving exercises. In *Proceedings of CONVERSATIONS 2021 - International Workshop on Chatbot Research and Design* (pp. 55-69). Cham, Switzerland: Springer.

Hobert, S. (2023). Fostering skills with chatbot-based digital tutors – training programming skills in a field study. *i-com* (Ahead-of-Print).

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning, 38*(1), 237-257.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference – DIS 2018* (pp. 895-906). New York, NY: ACM.

Jampen, D., Gür, G., Sutter, T., & Tellenbach, B. (2020). Don’t click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences, 10*(1), 1-41.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. Education and Information Technologies, 28(1), 973-1018.

Laban, G., Araujo, T. (2020). Working together with conversational agents: The relationship of perceived cooperation with service performance evaluations. In *Proceedings of Conversations 2019 - International Workshop on Chatbot Research and Design (pp.* 215–228*)*. Cham, Switzerland: Springer.

Lim, J., Ko, H., Yang, J. W., Kim, S., Lee, S., Chun, M. S., ... & Park, J. (2019). Active learning through discussion: ICAP framework for education in health professions. *BMC Medical Education, 19*(1), 1-8.

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: You have 50 milliseconds to make a good first impression!. *Behaviour & Information Technology, 25*(2), 115-126.

LinkedIn Learning (2020). 2020 workplace learning report. Technical report, LinkedIn Learning. https://learning.linkedin.com/content/dam/me/learning/resources/pdfs/LinkedIn-Learning-2020-Workplace-Learning-Report.pdf

LinkedIn Learning (2022). 2022 workplace learning report: The transformation of L&D. Technical report, LinkedIn Learning. https://learning.linkedin.com/content/dam/me/learning/en-us/pdfs/workplace-learning-report/LinkedIn-Learning_Workplace-Learning-Report-2022-EN.pdf

Litman, J. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition & Emotion, 19*(6), 793-814.

McTear, M. (2021) *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots.* Morgan & Claypool

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98-100.

Morris, J., & Chi, M. T. (2020). Improving teacher questioning in science using ICAP theory. *The Journal of Educational Research, 113*(1), 1-12.

O'Brien, H. L., Arguello, J., & Capra, R. (2020). An empirical study of interest, task complexity, and search behaviour on user engagement. *Information Processing & Management, 57*(3), 102226.

O’Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies, 112*, 28-39.

Peterson, E. G., & Hidi, S. (2019). Curiosity and interest: current perspectives. *Educational Psychology Review, 31*(4), 781-788.

Rudolph, J., Tan, S., Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 6(1), 1-22.Shevat, A. (2017). Designing Bots: Creating Conversational Experiences. Boston, MA: O'Reilly Media.

Sarosa, M., Kusumawadhani, M., Suyono, A., & Azis, Y. M. (2021). The effectiveness of chatbot as an online learning method on active and reflective learning styles. *Multicultural Education, 7*(9), 92-100.

Schöbel, S., Janson, A., & Mishra, A. (2019) A configurational view on avatar design–the role of emotional attachment, satisfaction, and cognitive load in digital learning. In Fortieth International Conference on Information Systems – ICIS 2019 (paper no. 1909).

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure?. Academy of Management Learning & Education, 9(2), 169-191.

Taylor, A. D., & Hung, W. (2022). The effects of microlearning: a scoping review. *Educational Technology Research and Development*, 1-33.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers, 13*(2), 127-145.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186-204.

Wekerle, C., Daumiller, M., & Kollar, I. (2022). Using digital technology to promote higher education learning: The importance of different learning activities and their relations to learning outcomes. *Journal of Research on Technology in Education, 54*(1), 1-17.

Wiggins, B. L., Eddy, S. L., Grunspan, D. Z., & Crowe, A. J. (2017). The ICAP active learning framework predicts the learning gains observed in intensely active classroom experiences. *AERA Open, 3*(2), https://doi.org/10.1177/2332858417708567.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (paper no. 652). New York, NY, ACM.

Winkler, R., Weingart, P., & Söllner, M. (2020). Using smart personal assistants for online learning activities: What benefits can we expect? In Proceedings der 15. Internationalen Tagung Wirtschaftsinformatik 2020 (pp. 465-479). Berlin, Germany: GITO Verlag, https://doi.org/10.30844/wi_2020_d7-winkler

Xiao, Z., Zhou, M. X., Chen, W., Yang, H., & Chi, C. (2020). If I hear you correctly: Building and evaluating interview chatbots with active listening skills. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (paper no. 4). New York, NY: ACM

Yin, J., Goh, T. T., Yang, B., & Xiaobin, Y. (2021). Conversation technology with micro-learning: The impact of chatbot-based learning on students' learning motivation and performance. *Journal of Educational Computing Research, 59*(1), 154-177.

CRediT author statement

Sebastian Hobert: Conceptualization, Methodology, Data curation, Formal analysis, Resources, Writing - original draft, Funding acquisition. **Asbjørn Følstad:** Conceptualization, Methodology, Data curation, Formal analysis, Resources, Writing - original draft, Funding acquisition. **Effie L.-C. Law:** Conceptualization, Methodology, Formal analysis, Writing - original draft.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: