Computers and Education Open 9 (2025) 100315

Contents lists available at ScienceDirect

Computers and Education Open

journal homepage: www.sciencedirect.com/journal/computers-and-education-open

Enhancing enthusiasm for STEM education with AI: Domain-specific chatbot as personalized learning assistant

Cilia Ricarda Rücker *, Sebastian Becker-Genschow

Department of Physics Education, University of Cologne, Gronewaldstr. 2, 50931 Cologne, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Chatbot
Personalization
Education
STEM
Mathematics

ABSTRACT

Generative AI has been increasingly integrated into educational settings for its potential to facilitate personalized learning. However, evidence-based implementation concepts remain essential to realize its pedagogical value. Therefore, this study investigates the educational efficacy of ADA, a domain-specific AI chatbot customized for secondary mathematics education, in enhancing learning outcomes and student engagement. A cluster-randomized controlled study was conducted with 195 ninth-grade students from German secondary schools in authentic classroom settings. The objective of the study was to compare ADA's personalized support against conventional differentiation materials. The intervention targeted the Heron method for estimating square roots in a single-lesson intervention with a pre-post assessment. As one of the few cluster-randomized controlled studies on custom AI chatbots in secondary mathematics education, this investigation assessed both affective and cognitive aspects of learning simultaneously to provide a comprehensive research picture. The findings indicated a high degree of student acceptance across the dimensions of the Technology Acceptance Model. Additionally, situational interest significantly improved in the chatbot condition. The study found that trends in performance and emotional responses were favorable, while cognitive load increased slightly. The findings emphasize the evident potential of customized AI chatbots to supplement differentiated instruction in the domain of mathematics education. They provide a critical foundation for future research investigating the potential of AI chatbots to enhance student learning through careful integration and customization.

1. Introduction

Conventional teaching methods in current STEM education frequently fall short in providing sufficient individualized learning support [1]. The rapid advancement of artificial intelligence (AI) has introduced innovative tools with the potential to address these systemic challenges in STEM education on a large scale [2–4]. Among these technological innovations, AI chatbots have emerged as particularly promising educational technologies, offering the capacity to provide personalized, interactive support tailored to individual learner needs [5–7].

AI chatbots enable students to interact through natural language, facilitating immediate feedback, comprehension support, and access to supplementary learning materials [2]. By responding to individual needs in real time, these systems have the capacity to engender student-centered, interactive classroom environments and, concomitantly, to reduce the instructional burden associated with differentiated teaching [8].

Recent meta-analytic evidence demonstrates that custom AI chatbots in particular have a statistically significant positive effect on student learning performance, with text-based interactions and STEM disciplines showing particularly substantial effectiveness [9]. Furthermore, the educational benefits of custom AI chatbots include enhanced learning motivation and the promotion of self-directed learning and problem-solving skills [10,11].

The present study investigates ADA, a domain-specific AI chatbot specifically developed for secondary STEM education. In contradiction to generic large language models, ADA incorporates domain-specific pedagogical frameworks and curriculum-aligned content that have been shown to enhance chatbot effectiveness in STEM contexts [9]. The objective of this study is to evaluate the effectiveness of ADA in comparison to conventional differentiation strategies with regard to students' learning performance, emotional responses, situational interest, and cognitive load. By examining the implementation of ADA in authentic classroom settings, the current study provides empirical insights into the pedagogical potential of customized AI chatbots in

* Corresponding author.
E-mail addresses: cilia.ruecker@uni-koeln.de (C.R. Rücker), sebastian.becker-genschow@uni-koeln.de (S. Becker-Genschow).

https://doi.org/10.1016/j.caeo.2025.100315
Received 24 June 2025; Received in revised form 31 October 2025; Accepted 23 November 2025
Available online 24 November 2025

secondary STEM education.

2. Current state of research

2.1. AI chatbots in education

2.1.1. Current applications and capabilities

Unlike conventional conversational agents, AI chatbots have the potential to function as learning assistants by providing personalized, interactive support that is tailored to the specific needs of individual learners [12–14]. Within educational environments, AI chatbots have been shown to possess a high degree of versatility with regard to their potential applications. Primarily, the provision of various types of immediate formative feedback and personalized correction on student work in a low-risk environment where learners need not fear making mistakes [15–19]. Additionally, they address individual learning difficulties by providing answers to questions and offering accurate explanations [15,20]. These explanations range from the transfer of solution strategies to complex scientific concepts and are available at various levels of understanding [21,22]. AI chatbots further offer supplementary learning resources, including additional exercises and explanatory texts comparable in quality to those found in textbooks [15,23,24]. Moreover, they systematically guide students step-by-step through the development of explanations for scientific phenomena [25] and the intricacies of problem-solving processes, thereby facilitating a more profound comprehension of underlying concepts [23].

2.1.2. Impacts on learning performance and experience

A growing body of research indicates that chatbot-assisted learning can surpass traditional learning in certain contexts, exhibiting favorable effects on both cognitive and affective dimensions of learning (e.g. [26–28]).

AI chatbots have demonstrated improvements in learning performance and reductions in cognitive load across various educational contexts [26,27,29]. Subsequent studies have demonstrated in particular the potential of ChatGPT to enhance students' productivity, engagement, knowledge construction, and higher-order thinking [2, 30–34].

Beyond cognitive dimensions, AI chatbots have demonstrated positive effects on student motivation, interest, and emotions [2,26,27,33]. AI chatbots have demonstrated efficacy in mitigating students' anxieties, stress, and the occurrence of a sense of learned helplessness [29, 34]. Furthermore, they have the capacity to augment the enjoyment of learning and the sense of well-being [34,35]. These findings are consistent with research that has documented an increase in positive attitudes toward learning and an enhancement in perceived value of learning [2,33].

These favorable outcomes may be linked to several characteristics observed in interactions with AI chatbots in educational settings [36]. A salient feature is their capacity to address particular learning challenges [37]. Furthermore, the promotion of low-pressure learning environments that encourage learners to seek clarification with greater ease were evidenced, particularly for ChatGPT [38].

However, critical considerations regarding the implementation of chatbots in educational settings have also emerged. While AI chatbots can reduce mental effort in research tasks compared to conventional methods, they may simultaneously result in a decline in the quality of reasoning and depth of understanding, presenting a significant trade-off for educators [39]. Furthermore, the educational effectiveness of AI chatbots tends to diminish over extended experiment durations, suggesting the need for strategic implementation and the development of novel pedagogical approaches to maintain student motivation and engagement over time [40].

In summary, AI chatbots demonstrate substantial potential to enhance both the cognitive and affective dimensions of learning. However, there remains a critical need for further research to identify successful didactic concepts and meaningful implementation strategies, particularly in subject-specific contexts [41]. Addressing these gaps is imperative to enable educators to maximize the pedagogical advantages of AI chatbots while mitigating potential drawbacks.

2.1.3. Applications and advantages in STEM education

The efficacy of chatbot-assisted learning is substantiated by domain-specific evidence across STEM (Science, Technology, Engineering, and Mathematics) disciplines. In the domain of biology education, significant improvements in students' academic achievements have been documented through the implementation of educational AI chatbots [42]. Computer science education research demonstrated that AI chatbots facilitate self-paced learning, enhance time efficiency as well as active participation, and yield superior academic outcomes compared to traditional methodologies [43,44]. Additionally, AI chatbots in computer science contexts have been shown to reduce stress, increase motivation, and positively influence situational interest [43,44].

Physics education research has similarly demonstrated benefits of AI chatbots integration. Studies document significant improvements in student achievement in both experiential learning dimensions and measurable performance metrics as well as effective rectification of preexisting misconceptions, contributing to more comprehensive understanding of fundamental physical concepts such as electromagnetism [45–47].

Mathematics education shows analogous benefits from AI chatbot integration. Studies demonstrate that AI chatbots facilitate robust conceptual understanding and enhance critical thinking processes in mathematical problem-solving contexts [48,49]. Additionally, chatbot utilization in mathematics instruction has been shown to enhance student self-efficacy [48].

Interdisciplinary research has further illuminated the potential of AI chatbots at the intersection of mathematics and physics education. While chatbot-generated explanations focusing on proportional relationships demonstrated indefinite impact on learning outcomes, they significantly reduced both intrinsic and extrinsic cognitive load [50]. Regarding affective dimensions, AI-generated materials have been demonstrated to elicit positive-activating emotions, situational interest, and self-efficacy among students [50].

These findings indicate that AI chatbots can serve as valuable resources for more personalized STEM education by catering to students' unique learning needs with substantial impact documented across various disciplines [51]. Notwithstanding these promising results, a significant research gap persists concerning the development and validation of specific didactic concepts for the meaningful integration of AI chatbots in subject-specific STEM contexts. Further investigations are required to identify successful didactic approaches and to develop meaningful implementation strategies that can harness the full pedagogical potential of these technologies [41]. Absent such research-based didactic frameworks, there is a risk that the integration of AI chatbots may result in disadvantages such as diminished reasoning quality or diminishing effects over extended usage periods [39,40].

2.1.4. Custom chatbots

While general-purpose AI chatbots have demonstrated considerable potential in education, recent research highlights the superior efficacy of custom AI chatbots that are tailored to specific pedagogical contexts. These systems address the critical limitations of general-purpose chatbots in education by integrating curriculum-aligned content, pedagogically informed prompt engineering, and design features that promote active learning rather than passive answer retrieval.

In the context of mathematics education, the implementation of custom chatbots has been shown to enhance practice performance in comparison to conventional ChatGPT interfaces [52]. This enhancement was achieved while circumventing the deleterious learning effects identified in standard ChatGPT usage [52]. In a similar vein, engineering education research demonstrated that customized chatbots exhibited superior efficacy in enhancing learning outcomes and self-efficacy in comparison to both standard ChatGPT and conventional instruction, with students reporting greater confidence and perceived effectiveness [53].

The mechanisms underlying these enhanced outcomes remain an active area of investigation. Custom AI chatbots appear to empower learners by catering to individual learning needs, enabling them to independently overcome challenges, study according to their preferred schedules, and progress at a pace that suits them [34,44]. Initial studies indicate that designs that facilitate interactive or constructive communication are more effective than passive approaches [11,54]. Further preliminary studies suggest a positive correlation between students' perceptions of social presence and human-likeness in AI chatbots and their motivation levels [55]. However, the number of empirical studies examining custom AI chatbots in educational contexts remains limited, highlighting substantial gaps in our understanding of how these systems function across diverse STEM disciplines and educational settings. Consequently, future research should endeavor to address this gap by developing theoretically grounded and empirically validated didactic approaches for the effective integration of custom AI chatbots across various STEM disciplines.

3. Research questions

The current state of research demonstrates the promising potential of AI chatbots in educational settings, particularly within STEM disciplines. The utilization of AI chatbots has been shown to enhance both affective dimensions (e.g., situational interest, emotions) and cognitive dimensions (e.g., learning performance, cognitive load). This enhancement is primarily attributable to their capacity to facilitate personalized learning experiences.

Nevertheless, despite these promising findings, the current state of research reveals notable limitations. Firstly, extant studies are predominantly situated in higher education, with limited empirical insights into domain-specific applications within secondary STEM education that align with curricula standards [7,56]. Secondly, there persists a substantial research gap regarding the development and validation of custom AI chatbots. As argued by Küchemann et al. [41], the absence of theoretically grounded and empirically supported integration concepts poses a risk of undermining their potential benefits or even counterproductive effects. Individual studies have cautioned that substandard implementation of AI chatbots has the potential to inadvertently compromise learning quality [39,40]. Moreover, few investigations have systematically examined how these technologies simultaneously affect both cognitive and affective learning dimensions in authentic classroom environments.

The present study aims to address the research gaps identified in the field of STEM education by investigating the efficacy of an AI chatbot named ADA, specifically designed for mathematics education. Utilizing the Heron method for estimating square roots as the learning context, the study compares student outcomes achieved through AI chatbot assistance to those obtained via traditional analog learning support. This study addresses the following five research questions:

RQ1. *How high is the technological acceptance of ADA among students?*

The successful integration of AI-based educational tools such as ADA is contingent not only on pedagogical efficacy but also on user acceptance [57]. Research has demonstrated that perceived usefulness and ease of use have a significant impact on students' propensity to engage with new learning technologies [58].

RQ2-5. *Does the utilization of the learning assistant ADA, in comparison to conventional differentiation measures, lead to*

RQ2. *… an increase in positive-activating emotions and a reduction in negative-deactivating emotions?*

Guided by Pekrun's [59] control-value theory of achievement emotions, this study explores how AI chatbot ADA influences learners' emotional states during the intervention. The study focuses on emotions that have been shown to impact engagement and learning: positive-activating emotions such as satisfaction, and negative-deactivating emotions such as frustration or boredom. We measure these emotions directly in response to the chatbot interaction rather than long-term affective outcomes.

RQ3. *… an improvement in situational interest?*

Building on AI chatbots' potential to stimulate situational interest, this study assesses whether ADA can spark greater situational interest than traditional learning materials. We prioritized dynamic, context-dependent constructs, such as emotions and situational interest, over more stable constructs, including self-efficacy, due to their heightened responsiveness to immediate, personalized AI chatbot interactions [59–61].

RQ4. *… a reduction in cognitive load?*

This study also investigates whether AI chatbot ADA supports more efficient cognitive processing by reducing intrinsic and extraneous cognitive load relative to traditional differentiation materials.

RQ5. *… an enhancement in learning performance?*

Given the growing evidence of AI chatbots' potential to enhance academic performance, this study explores whether AI chatbot ADA fosters measurable gains in mathematical achievement.

4. Methodology

4.1. Sample

A total of $N = 195$ ninth-grade students (91 female, 104 male) from secondary schools in Germany participated voluntarily in the randomized controlled study. The mean age of the students was 14.33 years (SD=0.67). The sample was divided into two groups: an experimental group (EG) and a control group (CG). The EG comprised 102 students, with 41 female and 61 male participants. The CG comprised 93 students, with 50 female and 43 male participants.

The sample size was determined using G*Power with parameters of an effect size (f) of 0.2, an alpha level of 0.05, and a power level (1-beta) of 0.8. The calculated sample size was 200, thus enabling the detection of medium effect sizes between groups.

4.2. Data collection procedure

A quasi-experimental, controlled intervention design with cluster randomization was employed. Due to technical constraints, existing class groups were assigned to either the CG or the EG. This methodological approach ensured the uniformity of treatment across classes, thereby preventing contamination effects and enhancing external validity. The fundamental distinction between the groups pertained to the nature of supplementary support provided. The students in the EG were able to access the AI chatbot ADA via a web browser on school-provided laptops. Conversely, the CG students employed static reference materials, namely help cards attached to their materials.

The data collection was executed during three scheduled mathematics lessons, with the objective of preserving ecological validity. The initial single lesson comprised a pre-test, the purpose of which was to collect demographic and prior mathematical knowledge data. This served to assess comparability between groups and to provide baseline measures for evaluating the intervention's impact on learning gains. The subsequent double lesson comprised the intervention followed immediately by a post-test measuring all dependent variables (see Section 4.4). In order to ensure procedural consistency, teachers were required to adhere to detailed lesson plans, which included standardized timing and instructions. Prior to the commencement of the study, all instructors received a structured online briefing.

The study received ethical approval from the University of Cologne's Faculty of Humanities Ethics Committee. Further, written informed consent was obtained from school administrators, parents, and students. The participants were thoroughly informed about data handling, the voluntary nature of participation, and withdrawal rights.

4.3. Chatbot design

The intervention employed ADA, a Custom Generative Pre-Trained Transformer (C-GPT) that had been adapted specifically for the purpose of STEM education. The system accessed the GPT-4 language model through a privacy-compliant Application Programming Interface (API) provided by fobizz, a widely used AI platform in German schools. This enabled interaction through an intuitive chat interface requiring no prior technical knowledge while ensuring General Data Protection Regulation (GDPR) compliance through anonymized data transmission [62,63].

ADA was conceptualized as a learning assistant for STEM education. It employs Socratic dialogues to enhance conceptual understanding by encouraging students to articulate their reasoning and examine assumptions [64,65]. Rather than merely providing answers, ADA utilized questioning techniques to assist students in discovering mathematical relationships independently, thereby providing scaffolding tailored to individual learning needs. The configuration process combined prompt engineering and retrieval-augmented generation (RAG; [66,67]). System prompts underwent a process of refinement through iterative testing to ensure mathematical accuracy, age-appropriate clarity, and pedagogical effectiveness. The integration of motivational design elements was informed by the Self-Determination Theory [68]. The language was carefully adapted for the student demographic, with particular emphasis placed on gender-inclusive phrasing [69,70]. In order to enhance accuracy and minimize hallucinations, ADA accessed a curated knowledge base containing curriculum content, the intervention materials, and in-depth background information on Ada Lovelace and the Heron method. The integration of the Wolfram Alpha plugin facilitated the accurate verification of student calculations and the immediate provision of mathematical steps when required.

4.4. Material

Both groups received identical paper-based materials explaining the Heron method, an iterative procedure for calculating square roots (see the Supplementary Material *I. Intervention Material*). The materials followed a scaffolded progression based on evidence-based practice for mathematical procedures (see [71]), moving systematically from prerequisite knowledge through guided exploration to independent application. The material commenced with prerequisite knowledge of square roots (Worksheet 1), followed by guided exploration through geometric visualizations (Worksheet 2), independent application (Worksheet 3), and critical optimization (Worksheet 4). The tasks were designed to span a range of competence levels, encompassing reproduction, application, and transfer, and are grounded in established textbooks by Griesel et al. [72] and Neubert and Benedix [73].

4.5. Measured variables

A variety of constructs were measured using established instruments, all of which were adapted to the specific learning context and translated into German if necessary. The assessment of all items was conducted using four-point Likert scales with anchors: 1 = strongly disagree, 2 = somewhat disagree, 3 = somewhat agree, and 4 = strongly agree. The variables of technological acceptance, emotional responses, situational interest, and cognitive load were measured exclusively post-intervention. The learning performance was collected at both the pre-test and post-test in order to facilitate comparison of changes over time.

4.5.1. Technology acceptance model

To assess students' acceptance of the AI chatbot in the experimental group, this study employed the Technology Acceptance Model (TAM; [74]). The TAM provides the basis for the iterative refinement of AI tools with the aim of sustainable technology adoption [75]. Four established TAM constructs were used to evaluate students' perceptions of the AI chatbot [76–78]. The constructs encompassed perceived ease of use (three items), perceived usefulness (four items), attitudes toward using (five items), and perceived competence (four items).

4.5.2. Emotions

The emotional states of the students during the learning phase were measured using five items adapted from the Achievement Emotions Questionnaire [79]. The items assessed positive-activating emotions (two items: pleasure and satisfaction) and negative-deactivating emotions (three items: boredom, frustration, and uncertainty).

4.5.3. Situational interest

The situational interest of the students was assessed using a set of seven items drawn from the International Situational Interest Questionnaire by Potvin et al. [60]. The assessment draws on Hidi and Renninger's [80] conceptualization of situational interest as a precursor to sustained STEM participation.

4.5.4. Cognitive load

The cognitive load experienced by the students was measured using a set of five items drawn from Klepsch et al. [81]. The items were designed to assess intrinsic cognitive load (two items) and extrinsic cognitive load (three items).

4.5.5. Learning performance

The learning outcomes were assessed using a set of four items derived from mathematics textbooks [73,82]. The test items systematically progressed from basic conceptual recall to advanced problem-solving, mirroring the scaffolded structure of the learning materials. The students were able to demonstrate their acquired knowledge at the reproduction level by completing clozes, at the application level by applying the Heron method, and at the transfer level by critically optimizing the method.

5. Data analysis

All analyses were conducted using R statistical software (Version 4.4.2; [83]).

5.1. Validation of the measurement models

The structural configuration's appropriateness for all models was confirmed through exploratory factor analysis (EFA) (R package "psych") and confirmatory factor analysis (CFA) (R package "lavaan").

5.1.1. Technology acceptance model

The parallel analysis supported a three-factor solution. Both three-factor ($\chi^2(101)=91.95$, $p = 0.729$; CFI=1.00; RMSEA=0.00; SRMR=0.06) and four-factor models ($\chi^2(98)=87.28$, $p = 0.773$; CFI=1.00; RMSEA=0.00; SRMR=0.06) demonstrate exemplary model fit, with no significant difference between the two ($\Delta\chi^2=4.68$, $\Delta df=3$, $p = 0.130$). The limited empirical discriminability within the current dataset can be ascribed to the fact that TAM dimensionality has been shown to be context-sensitive, particularly when user experience is limited [58,84]. In order to ensure theoretical consistency with the TAM [76,77], the four-factor model was retained.

5.1.2. Emotions

The parallel analysis supported a two-factor solution. The two-factor model ($\chi^2(4)$=12.67, p = 0.013; CFI=1.00; RMSEA=0.11; SRMR=0.04) demonstrated a significantly superior fit in comparison to the single-factor model ($\Delta\chi^2$=16.54, Δdf=1, $p < 0.0001$), confirming the separability of positive-activating and negative-deactivating emotions in our data, thereby aligning with the dimensional structure proposed by Pekrun et al. [79].

5.1.3. Situational interest

The results of both exploratory and confirmatory factor analyses indicated a two-factor structure ($\chi^2(13)$=11.86, p = 0.539; CFI=1.00; RMSEA=0.00; SRMR=0.05). As demonstrated by the model comparison tests, a significant improvement was observed in relation to the unidimensional structure ($\Delta\chi^2$=21.82, Δdf=1, $p < 0.0001$). However, it is likely that this statistical structure is an artifact of methodological issues rather than an indication of substantive conceptual distinctions [85]. Given that the simpler unidimensional model exhibited an acceptable fit and demonstrated a stronger theoretical alignment with prevailing frameworks in educational intervention research [60,86], it was adopted. This approach circumvents arbitrary distinctions that lack substantial theoretical foundation while enhancing explanatory power for intervention effects [87].

5.1.4. Cognitive load

The parallel analysis supported a two-factor solution. The two-factor model ($\chi^2(4)$=3.94, p = 0.414; CFI=1.00; RMSEA=0.00; SRMR=0.03) demonstrated a significantly superior fit compared to the single-factor model ($\Delta\chi^2$=11.18, Δdf=1, $p < 0.0001$), validating the distinction between intrinsic and extrinsic cognitive load in our data, thereby aligning with the cognitive load theoretical framework proposed by Klepsch et al. [81].

5.2. Test for normal distribution

The Shapiro-Wilk test indicated that the variables of emotions, cognitive load, situational interest, and learning performance are not normally distributed (R package "stats"). While non-parametric tests (Mann-Whitney U test, R package "stats") were employed for the purpose of group comparisons, a Repeated Measures (RM-) ANOVA (Type III sum of squares, R package "ez") was used for pre-post analysis. Given the adequate sample size (N = 195) and approximately balanced group sizes, RM-ANOVA remains robust to normality violations, with Monte Carlo studies demonstrating acceptable Type I error control for moderate departures from normality [88].

5.3. Test for initial group differences

Despite the implementation of cluster randomization, Mann-Whitney U tests indicated significant disparities between groups with regard to prior knowledge pertaining to the Heron method on all evaluable pre-test performance variables ($p < 0.05$). To account for these initial differences, a RM-ANOVA was employed to analyze learning outcomes.

5.4. Test for interrater reliability

The majority of performance test items were closed-ended, with tasks 4.2(a) and 4.2(b) requiring open-ended responses (see the Supplementary Material *II. Post-test*). The responses were evaluated by two independent raters according to a three-dimensional framework that encompassed depth of understanding, factual accuracy, and content accuracy (see Appendix Table A.1). Cohen's weighted kappa with quadratic weighting (R package "irr") demonstrated excellent interrater agreement for the composite scores (κ=0.921–0.974) and the individual dimensions (κ=0.816–1.000).

6. Results

6.1. Descriptive statistics

Table 1 presents a comprehensive overview of the descriptive statistics for all outcome variables across EG and CG.

6.1. Technological acceptance

Descriptive analyses within the EG revealed positive perceptions toward the AI chatbot (see Appendix Fig. A.1). Students reported favorable assessments of ADA's perceived ease of use (M = 3.44, SD=0.83), perceived usefulness (M 3.03, SD=0.81), attitudes toward use (M = 3.01, SD 0.85), and perceived competence (M 3.09, SD=0.83).

6.2. Emotions

Descriptive analyses revealed subtle differences in emotional responses across the EG and the CG. The EG demonstrated slightly elevated mean scores for positive-activating emotions (M 2.37, SD=1.03) relative to the CG (M = 2.10, SD 0.96), alongside slightly reduced negative-deactivating emotion scores (M 2.45, SD=1.15 vs. M 2.61, SD=1.15) (Fig. 1).

The Mann-Whitney U test revealed no significant differences between the EG and the CG for positive-activating (p = 0.053) or negative-deactivating emotions (p = 0.265).

6.3. Situational interest

Descriptive analyses revealed substantial differences in situational interest across the EG and the CG. The EG demonstrated notably higher mean scores for situational interest (M 2.63, SD=0.51) compared to the CG (M = 2.43, SD=0.62), indicating a pronounced increase in student interest during the learning activity (Fig. 2).

The Mann-Whitney U test revealed a highly significant difference between the EG and the CG for situational interest (p = 0.00005). The effect size analysis using Cohen's d revealed a medium positive effect (d = 0.63), suggesting a substantial enhancement of situational interest in the EG.

Table 1
Descriptive statistics.

Variable	Group ($n_♀$, $n_♂$)	M	SD	95 % CI
Technological Acceptance				
Perceived ease of use	EG (41, 61)	3.44	0.83	[3.28,3.60]
Perceived usefulness	EG (41, 61)	3.03	0.81	[2.87,3.19]
Attitudes toward use	EG (41, 61)	3.01	0.85	[2.84,3.18]
Perceived competence	EG (41, 61)	3.09	0.83	[2.93,3.25]
Emotions				
Positive-activating	EG (41, 61)	2.37	1.03	[2.17,2.57]
	CG (50, 43)	2.10	0.96	[1.90,2.30]
Negative-deactivating	EG (41, 61)	2.45	1.15	[2.23,2.67]
	CG (50, 43)	2.61	1.15	[2.38,2.84]
Situational Interest				
	EG (41, 61)	2.63	0.51	[2.53,2.73]
	CG (50, 43)	2.43	0.62	[2.30,2.56]
Cognitive Load				
Intrinsic	EG (41, 61)	2.75	1.01	[2.55,2.95]
	CG (50, 43)	2.63	0.97	[2.43,2.83]
Extrinsic	EG (41, 61)	2.62	0.98	[2.43,2.81]
	CG (50, 43)	2.38	1.02	[2.17,2.59]
Learning Performance				
Pre-test	EG (41, 61)	3.56	4.18	[2.75,4.37]
	CG (50, 43)	8.61	8.28	[6.93,10.29]
Post-test	EG (41, 61)	12.42	12.72	[9.95,14.89]
	CG (50, 43)	9.38	12.51	[6.84,11.92]

Notes: EG = Experimental Group, CG = Control Group, M = Mean, SD = Standard Deviation, CI = Confidence Interval (95 %).

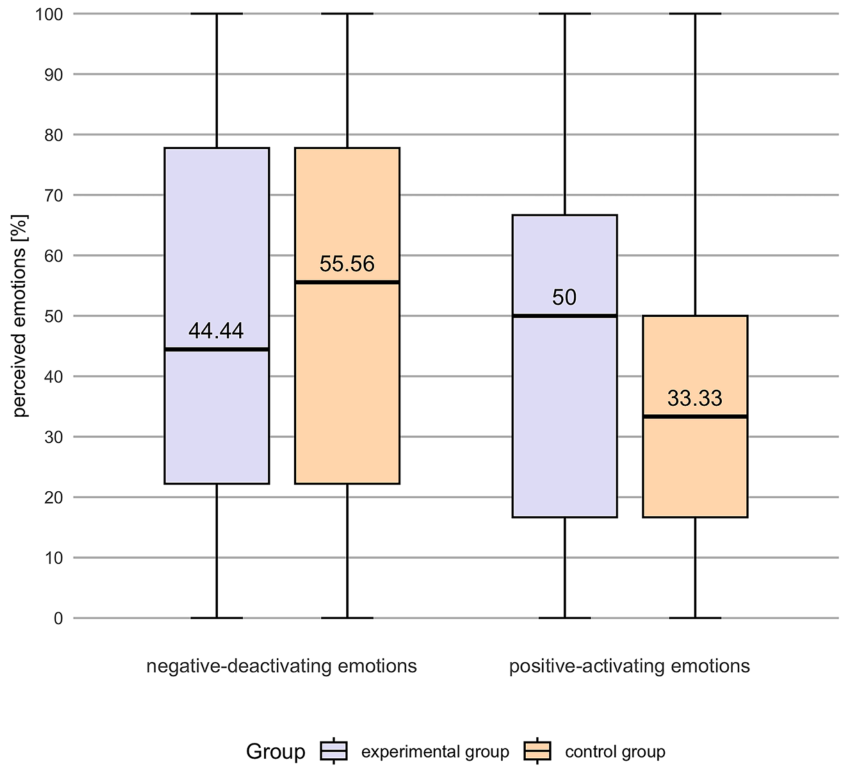

Fig. 1. Boxplots of perceived emotions for EG and CG.

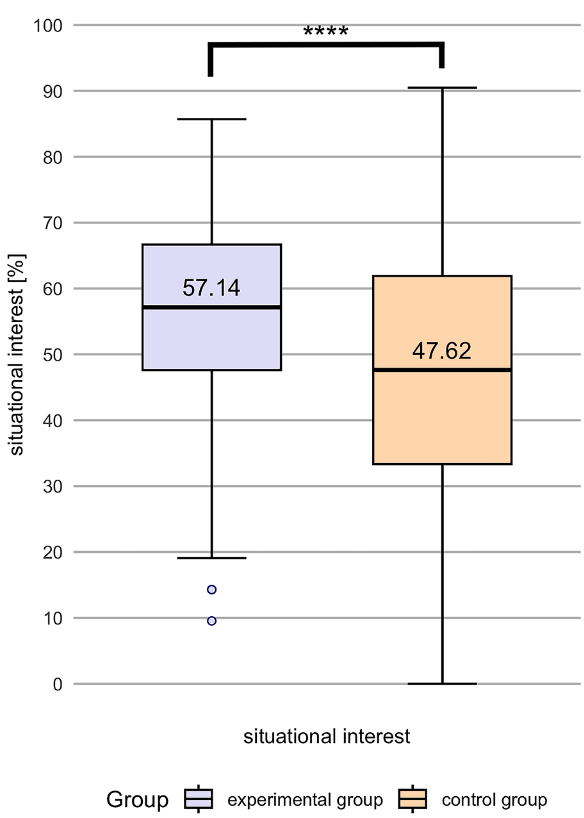

Fig. 2. Boxplots of situational interest for EG and CG.

6.4. *Cognitive load*

Descriptive analyses revealed nuanced differences in cognitive load across the EG and the CG. The EG demonstrated marginally elevated mean scores for intrinsic cognitive load (M 2.75, SD 1.01) relative to the CG ($M = 2.63$, SD=0.97), alongside slightly increased extrinsic cognitive load scores ($M = 2.62$, SD=0.98 vs. $M = 2.38$, SD 1.02) (Fig. 3).

The Mann-Whitney U test indicated no significant differences between the EG and the CG for intrinsic (p 0.299) or extrinsic cognitive load (p 0.060).

6.5. *Learning performance*

Descriptive analyses revealed notable differences in the learning progression between the EG and CG. Despite the CG's higher initial performance in the pre-test across all tasks (EG: M 3.56, SD=4.18; CG: M 8.61, SD 8.28), the EG showed greater overall improvement from pre- to post-test (EG: M 12.42, SD 12.72; CG: M 9.38, SD 12.51) (Fig. 4).

An RM-ANOVA confirmed a significant main effect of time, F (1194)=145.76, $p < 0.001$, ges=0.22, and group, F(1194)=9.26, $p =$ 0.003, ges=0.03, but no significant group by time interaction, F(1194)= 2.84, $p = 0.094$ (Fig. 5).

At the subtask level, Task A3 (application task; EG: $M = 9.44$, SD=10.37; CG: M 6.68, SD 11.31) and Task A4.2(a) (transfer task; EG: $M = 0.49$, SD=1.42; CG: M 0.29, SD 1.03) exhibited marked differences. In addition to the descriptive trends, each subtask demonstrated a significant time effect (all ps<0.001), along with significant group differences for A1 (p 0.001), A3 ($p = 0.022$), and A41 (p 0.005). Although no interaction reached significance at the 0.05 level, Task A3 came closest, F(1194)=3.18, $p = 0.076$ (Fig. 6).

7. Discussion

The purpose of this investigation was to examine the efficacy of a domain-specific AI chatbot ADA in ninth-grade mathematics classrooms.

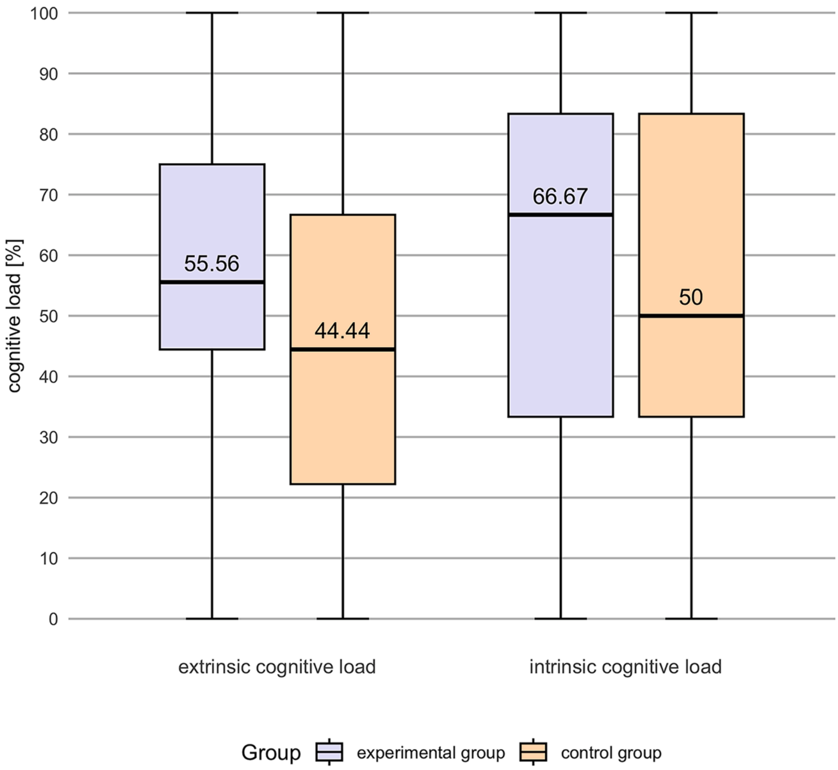

Fig. 3. Boxplots of cognitive load for EG and CG.

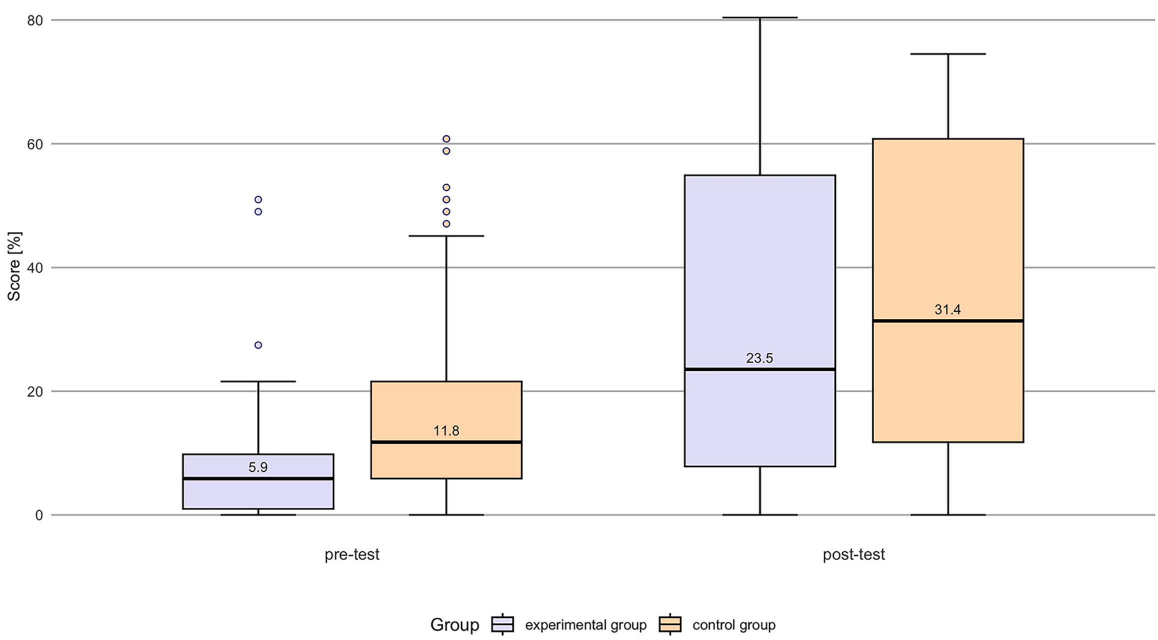

Fig. 4. Boxplots of learning performance for EG and CG.

As one of the few studies to employ a cluster-randomized design in authentic secondary education settings whilst simultaneously assessing both affective and cognitive outcomes, our findings regarding situational interest, emotional responses, cognitive load, and learning performance provide particularly valuable empirical evidence for the efficacy of individualized learning process support through custom bots in STEM education.

7.1. *Technological acceptance*

Concerning RQ 1 ("How high is the technological acceptance of ADA among students?"), the findings demonstrate a high level of acceptance of the AI chatbot ADA across all TAM dimensions. The mean scores persistently exceed 3.0 on the 4-point scale. The high usability scores of ADA can be attributed to architectural innovations that address common Natural Language Processing limitations, including RAG techniques for improved response accuracy and prompt engineering for enhanced

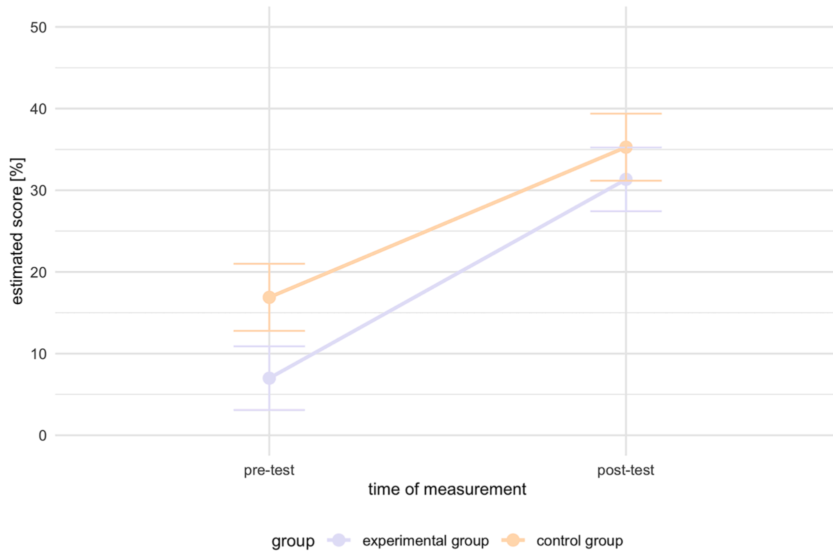

Fig. 5. Interaction plot for RM-ANOVA for overall performance.

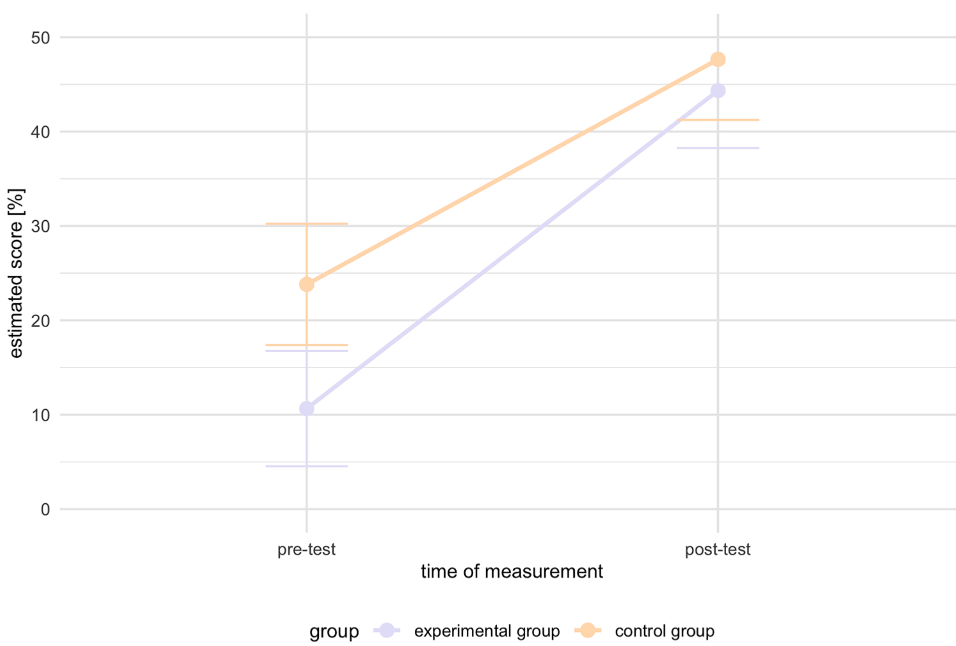

Fig. 6. Interaction plot for RM-ANOVA for task A3.

response patterns [66]. The observed high perceived usefulness of ADA aligns with prior research demonstrating high perceived usefulness of educational chatbots across diverse learning contexts [89–91]. Furthermore, the positive ratings of competence indicate the successful conveyance of domain expertise, a crucial factor in the acceptance and adoption of technology [92]. Finally, the students demonstrated a positive attitude towards using ADA, which is consistent with TAM theory given the strong usability and usefulness ratings [93–95]. This finding aligns with the broader body of evidence suggesting that well-designed educational chatbots tend to engender positive attitudes when incorporating empathic design elements [54] and maintaining consistent response quality [96].

7.2. Emotions

Turning to RQ 2 ("Does the utilization of the learning assistant ADA […] lead to an increase in positive-activating emotions and a reduction in negative-deactivating emotions?"), the findings reveal a directional but non-significant trend. The utilization of AI chatbot ADA in the context of math education has been observed to elicit elevated positive-activating emotions and slightly diminished negative-deactivating emotions in comparison with learning employing conventional differentiation measures. Despite the absence of statistical significance, the observed trends align directionally with prior research, even though previous studies have demonstrated more pronounced positive emotional responses to AI chatbot interventions, particularly in relation to negative-deactivating emotions [29,35]. The absence of significant effects could be attributed to the relatively brief nature of the intervention, as prior research demonstrates that measurable emotional shifts in educational contexts typically require sustained exposure to manifest robustly [97].

7.3. Situational interest

In relation to RQ 3 ("Does the utilization of the learning assistant ADA […] lead to an improvement in situational interest?"), the results provide robust evidence for a significant positive effect. The employment of ADA within the domain of mathematics education significantly enhanced situational interest compared to learning employing conventional differentiation measures. This finding is consistent with prior research indicating that AI chatbots can enhance learners' interest across various educational settings, including nursing, teacher, language, computer science, physics, and chemistry education (e.g., [31,32, 98]). However, Huang et al. [11] emphasize that the motivational impact of chatbots is context-dependent. This assertion is corroborated by Kumar [99], who reported no significant benefits in relation to affective-motivational outcomes in their study of educational chatbots in design education.

7.4. Cognitive load

The results pertaining to RQ 4 ("Does the utilization of the learning assistant ADA […] lead to a reduction in cognitive load?") indicate a modest yet non-significant tendency towards increased intrinsic and extrinsic cognitive load. This finding is in contrast to the results of previous research, which has reported reductions in cognitive load as a key benefit of AI-assisted learning environments [29,39,50,]. Two principal factors may account for the observed patterns of cognitive load. Firstly, in contradistinction to conventional learning materials with predetermined scaffolding, the AI chatbot imposes dual-task demands on students [100]. Secondly, the brevity of the intervention may have been insufficient to allow students to achieve cognitive efficiency in their interactions with ADA [75,84].

7.5. Learning performance

Addressing the influence of the AI chatbot ADA on learning performance, the results concerning RQ 5 ("Does the utilization of the learning assistant ADA […] lead to an enhancement in learning performance?") indicate non-significant yet promising trends. In comparison to conventional differentiation methods, the implementation of ADA was associated with a more favorable development in learning performance. While the effects did not reach statistical significance, the observed tendencies point toward a potential benefit of the chatbot in fostering conceptual understanding and higher-order thinking. This interpretation is supported by recent research (e.g., [27,30,101]). Analogous to previous arguments, the lack of significant effects on learning performance may be primarily attributed to the brevity of the intervention, the dual-task demands imposed by the AI chatbot, and the limited scope of the participant sample.

8. Conclusion & future directions

8.1. Major findings

The present study demonstrates that the domain-specific AI chatbot ADA was robustly accepted by students, who perceived it as a usable, useful, and competent tool for mathematics learning and also reported a positive attitude towards its use. The positive attitude scores observed in this study suggest that ADA successfully fulfills the foundational conditions for sustained technology adoption in educational settings. While its immediate affective impact on emotional states was limited, the intervention with ADA significantly enhanced situational interest, with a medium effect size. This finding underscores the potential of ADA to foster motivational aspects of the learning process in STEM education. Notably, cognitive load patterns deviated from prior research, demonstrating modest increases rather than reductions. The necessity for scaffolding in order to optimize interaction design is suggested, particularly regarding the dual-task nature of inexperienced chatbot use. Regarding learning outcomes, results revealed an asymmetry between affective and cognitive effects. While the intervention significantly enhanced situational interest, these gains did not translate into significant improvements in learning performance. Although promising trends emerged, particularly in higher-order application tasks, the findings indicate that in this intervention ADA was more effective at enhancing motivational engagement than at producing immediate performance gains. Taken together, these findings establish AI chatbot ADA as an educationally viable and future-oriented approach, particularly effective in enhancing motivational engagement rather than producing immediate performance gains.

8.2. Limitations

Several limitations must be acknowledged, many of which reflect deliberate design choices aimed at balancing internal validity with ecological realism. Firstly, the relatively brief nature of the intervention may have resulted in a limitation of the potential for significant effects to emerge. Secondly, as a consequence of this brief exposure, only immediate short-term effects measurements were collected, precluding any assessment of long-term effects. Thirdly, the study was conducted with a limited sample size within a specific learning context and with a particular learning group. This limits the extent to which the results can be generalized. Fourthly, while analyzing validated educational constructs rather than usage metrics, we acknowledge that detailed interaction patterns could provide complementary insights. Finally, novelty effects cannot be entirely ruled out, despite the observed situational interest suggesting effects that extend beyond mere technological curiosity.

8.3. Implications for future research

The study's findings suggest that domain-specific AI chatbots have the potential to provide targeted support for STEM learning processes while simultaneously highlighting several crucial avenues for future investigation. The primary objective is to replicate the observed trends with substantially larger samples. This is imperative to evaluate the statistical robustness and generalizability of the effects identified in this study. Beyond validation, it is imperative to extend the scope of application. Future research should explore the transferability of chatbot-based interventions to a broader range of mathematical and scientific learning contexts. It is further recommended that comparative studies be conducted across different educational stages in order to examine the manner in which age-related differences shape the interaction with chatbots and the learning outcomes thereof. Furthermore, the short-term nature of the observed effects necessitates the implementation of longitudinal designs. Particularly, the observed asymmetry between enhanced situational interest and learning performance reveals a critical need to examine whether increased motivational engagement translates into improved learning outcomes over extended intervention periods. Longitudinal studies could further elucidate whether immediate performance gains are attenuated by dual-task demands of simultaneously managing paper-based materials and digital devices, whether the short-term effects persist over time, and whether initial, non-significant trends evolve into statistically significant effects when students engage with chatbots across multiple lessons. Another critical avenue that merits exploration is the dual-task challenge. In order to address this issue, future work should focus on a more seamless integration of chatbot dialogue with instructional content.

8.4. Didactic considerations and practical implications

Despite the implementation of technical safeguards implemented in ADA (see Section 4.3.), hallucinations persist as an inherent limitation of language models, arising from the fundamental statistical properties of their training, even with perfect training data [102]. For educational contexts, this means that despite RAG substantially reducing hallucination rates compared to parametric-only models [66], the risk of plausible but incorrect outputs cannot be eliminated. Further, there is a risk of inconsistent terminology or Socratic questions that are misleading. These inherent limitations underscore that ADA was designed as a learning assistant to support and relieve, not replace, teacher expertise. Ensuring sustainable implementation therefore requires teacher oversight, contemporary adaptation of the curated database to the learning group, and fostering AI literacy skills among both teachers and students (e.g., [103,104]). Crucially, maximizing the didactic utility and individualization potential of domain-specific custom AI chatbots hinges on empowering educators with the competence to configure and customize such bots. This capability allows teachers to precisely align them with their distinct pedagogical objectives, personal teaching preferences, and the unique needs of their students.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude (Anthropic) and ChatGPT (OpenAI) in order to assist with proofreading. After using this tools/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article

CRediT authorship contribution statement

Cilia Ricarda Rücker: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sebastian Becker-Genschow:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors have declared no conflict of interest

Acknowledgements

The authors wish to express their sincere gratitude to the participating secondary schools in Germany, their administrators, and teaching staff who facilitated this research. We would like to express our sincere gratitude to the ninth-grade students who voluntarily participated in this study. We would also like to express our gratitude to Jannik Henze for his contribution as an independent second rater for the open-ended mathematics tasks, thereby ensuring the reliability of our scoring procedures.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.caeo.2025.100315.

Appendix

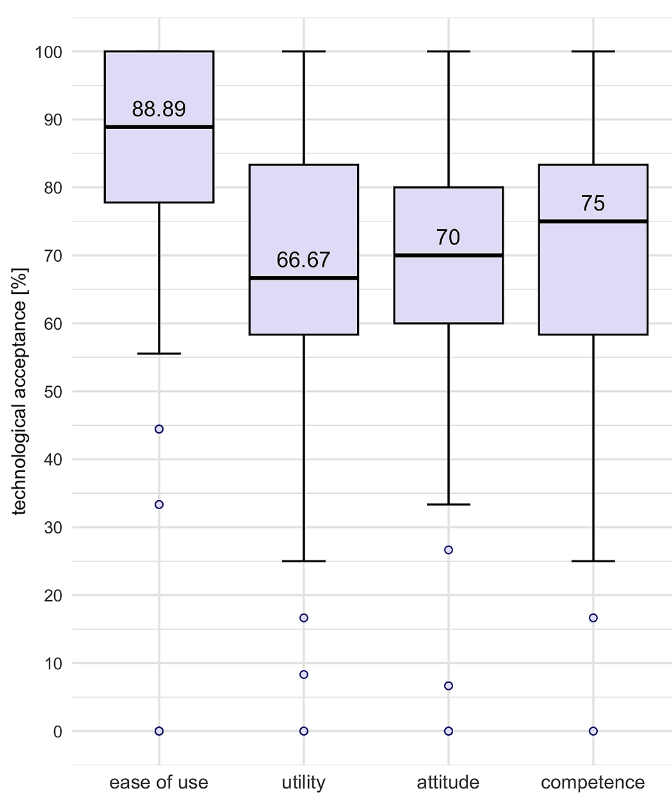

Fig. A.1. Boxplots of TAM for the EG 1.

Table A.1
Coding rubric.

Dimension	0	1	2
Depth of understanding	No understanding	Minimal to adequate understanding	Deep understanding
Factual accuracy	Full of errors	Partial errors	Error-free
Content accuracy (precision)	Largely inaccurate or irrelevant	Partially accurate and relevant	Accurate and relevant

References

[1] Lange-Hegermann M, Schmohl T, Watanabe A, Heiss S, Rubart J. AI-based STEM education: Generating individualized exercises in mathematics. Proceedings og the 10thInternational Conference on New Perspectives in Science Education. 2021. STEM4992, https://doi.org/10.25656/01:27944.

[2] Chang C-Y, Hwang G-J, Gau M-L. Promoting students' learning achievement and self-efficacy: a mobile chatbot approach for nursing training. Br J Educ Technol 2022;53(1):171–88. https://doi.org/10.1111/bjet.13158.

[3] Chen L, Chen P, Lin Z. Artificial intelligence in education: a review. IEEE Access 2020;8:75264–78. https://doi.org/10.1109/ACCESS.2020.2988510.

[4] Mahligawati F, Allanas E, Butarbutar MH, Nordin NAN. Artificial intelligence in Physics Education: a comprehensive literature review. J Phys 2023;2596:012080. https://doi.org/10.1088/1742-6596/2596/1/012080.

[5] Adiguzel T, Kaya MH, Cansu FK. Revolutionizing education with AI: exploring the transformative potential of ChatGPT. Contemp Educ Technol 2023;15(3):ep429. https://doi.org/10.30935/cedtech/13152.

[6] García-Martínez I, Fernández-Batanero JM, Fernández-Cerero J, León SP. Analysing the impact of artificial intelligence and computational sciences on student performance: systematic review and meta-analysis. J New Approach Educ Res 2023;12(1):171–97. https://doi.org/10.7821/naer.2023.1.1240.

[7] Sari HE, Tumanggor B, Efron D. Improving educational outcomes through adaptive learning systems using AI. Int Trans Artif Intell 2024;3(1):21–31. https://doi.org/10.33050/italic.v3i1.647.

[8] Baidoo-Anu D, Owusu Ansah L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. J AI 2023;7(1):52–62. https://doi.org/10.2139/ssrn.4337484.

[9] Laun M, Wolff F. Chatbots in education: hype or help? A meta-analysis. Learn Individ Differ 2025;119:102646. https://doi.org/10.1016/j.lindif.2025.102646.

[10] Xu W, Ouyang F. The application of AI technologies in STEM education: a systematic review from 2011 to 2021. Intern J STEM Edu 2022;9(59). https://doi.org/10.1186/s40594-022-00377-5.

[11] Huang AYQ, Lu OH, Yang SJ. Effects of artificial Intelligence–Enabled personalized recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. Comput Educ 2023;194:104684. https://doi.org/10.1016/j.compedu.2022.104684.

[12] Kohnke L. A pedagogical chatbot: a supplemental language learning tool. RELC J 2023;54(3):828–38. https://doi.org/10.1177/00336882211067054.

[13] Limo FAF, Tiza DRH, Roque MM, Herrera EE, Murillo JPM, Huallpa JJ, Flores VAA, Castillo AGR, Peña PFP, Carranza CPM, Gonzáles JLA. Personalized tutoring: chatGPT as a virtual tutor for personalized learning experiences. Przestrzeń Społeczna (Social Space) 2023;23(1):293–312.

[14] Tossell CC, Tenhundfeld NL, Momen A, Cooley K, Visser EJd. Student perceptions of ChatGPT use in a college essay assignment: implications for learning, grading, and trust in artificial intelligence. IEEE Trans Learn Technol 2024;17:1069–81. https://doi.org/10.1109/TLT.2024.3355015.

[15] Ait Baha T, El Hajji M, Es-Saady Y, Fadili H. The impact of educational chatbot on student learning experience. Educ Inf Technol 2024;29:10153–76. https://doi.org/10.1007/s10639-023-12166-w.

[16] Escalante J, Pack A, Barrett A. AI-generated feedback on writing: insights into efficacy and ENL student preference. Int J Educ Technol High Educ 2023;20(57). https://doi.org/10.1186/s41239-023-00425-2.

[17] Latif E, Zhai X. Fine-tuning ChatGPT for automatic scoring. Compu Educ 2024;6:100210. https://doi.org/10.1016/j.caeai.2024.100210.

[18] Steinert S, Avila KE, Ruzika S, Kuhn J, Küchemann S. Harnessing large language models to develop research-based learning assistants for formative feedback. Smart Learn Environ 2024;11(62). https://doi.org/10.1186/s40561-024-00354-1.

[19] Yuan Y. An empirical study of the efficacy of AI chatbots for English as a foreign language learning in primary education. Interactl Learnl Environ 2024;32(10):6774–89. https://doi.org/10.1080/10494820.2023.2282112.

[20] Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. Int J Educ Technol High Educ 2023;20(56). https://doi.org/10.1186/s41239-023-00426-1.

[21] Liang Y, Zou D, Xie H, Wang FL. Exploring the potential of using ChatGPT in physics education. Smart Learn Environ 2023;10(52). https://doi.org/10.1186/s40561-023-00273-7.

[22] Orrù G, Piarulli A, Conversano C, Gemignani A. Human-like problem-solving abilities in large language models using ChatGPT. Front Artif Intell 2023;6:1199350. https://doi.org/10.3389/frai.2023.1199350.

[23] 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, Stadler M, Kasneci G. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ 2023;103:102274. https://doi.org/10.1016/j.lindif.2023.102274.

[24] Küchemann S, Steinert S, Revenga N, Schweinberger M, Ding Y, Avila KE, Kuhn J. Can ChatGPT support prospective teachers in physics task development? Phys Rev Phys Educ Res 2023;19:020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128.

[25] Adair A. AI-driven assessment and scaffolding for mathematical modeling and explanations during science investigations. 2024. Retrieved from, *https://www.proquest.com/dissertations-theses/ai-driven-assessment-scaffolding-mathematical/docview/3126862548/se-2* [Accessed 19 May 2025].

[26] Gökçearslan Ş, Tosun C, Erdemir ZG. Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: a systematic literature review. Int J Technol Educ 2024;7(1):19–39. https://doi.org/10.46328/ijte.600.

[27] Wu R, Yu Z. Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. Br J Educ Technol 2023;55(1):10–33. https://doi.org/10.1111/bjet.13334.

[28] Yuan C-C, Li C-H, Peng C-C. Development of mobile interactive courses based on an artificial intelligence chatbot on the communication software LINE. Interactl Learnl Environ 2023;31(6):3562–76. https://doi.org/10.1080/10494820.2021.1937230.

[29] Stathakarou N, Nifakos S, Karlgren K, Konstantinidis ST, Bamidis PD, Pattichis CS, Davoody N. Students' perceptions on chatbots' potential and design characteristics in healthcare education. Stud Health Technol Inform 2020;272:209–12. https://doi.org/10.3233/SHTI200531.

[30] Fauzi F, Tuhuteru L, Sampe F, Ausat AMA, Hatta HR. Analysing the role of ChatGPT in improving student productivity in higher education. J Educ 2023;5(4):14886–91. https://doi.org/10.31004/joe.v5i4.2563.

[31] Lee H-Y, Chen P-H, Wang W-S, Huang Y-M, Wu T-T. Empowering ChatGPT with guidance mechanism in blended learning: effect of self-regulated learning, higher-order thinking skills, and knowledge construction. Int J Educ Technol High Educ 2024;21(16). https://doi.org/10.1186/s41239-024-00447-4.

[32] Li H. Effects of a ChatGPT-based flipped learning guiding approach on learners' courseware project performances and perceptions. Austral J Educ Technol 2023;39(5):40–58. https://doi.org/10.14742/ajet.8923.

[33] Sandu N, Gide E. Adoption of AI-chatbots to enhance student learning experience in higher education in India. In: 2019 18th International conference on information technology based higher education and training; 2019. p. 1–5. https://doi.org/10.1109/ITHET46829.2019.8937382.

[34] Wu T-T, Lee H-Y, Li P-H, Huang C-N, Huang Y-M. Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning Aid. J Educ Comput Res 2024;61(8):1539–67. https://doi.org/10.1177/07356331231191125.

[35] Al-Hafdi FS, AlNajdi SM. The effectiveness of using chatbot-based environment on learning process, students' performances and perceptions: a mixed exploratory study. Educ Inf Technol 2024;29:20633–64. https://doi.org/10.1007/s10639-024-12671-6.

[36] Song C, Song Y. Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. Front Psychol 2023;14:1260843. https://doi.org/10.3389/fpsyg.2023.1260843.

[37] Huang W, Hew K, Fryer L. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. J Comput Assist Learn 2021;38(1):237–57. https://doi.org/10.1111/jcal.12610.

[38] Lo CK, Hew KF, Jong MS-Y. The influence of ChatGPT on student engagement: a systematic review and future research agenda. Comput Educ 2024;219:105100. https://doi.org/10.1016/j.compedu.2024.105100.

[39] Stadler M, Bannert M, Sailer M. Cognitive ease at a cost: lLMs reduce mental effort but compromise depth in student scientific inquiry. Comput Human Behav 2024;160:108386. https://doi.org/10.1016/j.chb.2024.108386.

[40] Alemdag E. The effect of chatbots on learning: a meta-analysis of empirical research. J Res Technol Educ 2023;57(2):459–81. https://doi.org/10.1080/15391523.2023.2255698.

[41] Küchemann S, Steinert S, Kuhn J, Avila K, Ruzika S. Large language models–Valuable tools that require a sensitive integration into teaching and learning physics. Phys Teach 2024;62(5):400–2. https://doi.org/10.1119/5.0212374.

[42] Lin Y-T, Ye J-H. Development of an educational chatbot system for enhancing students' biology learning performance. J Internet Technol 2023;24(2):275–81. https://doi.org/10.53106/160792642023032402006.

[43] Essel HB, Vlachopoulos D, Tachie-Menson A, Johnson EE, Baah PK. The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher

education. Int J Educ Technol High Educ 2022;19(57). https://doi.org/10.1186/s41239-022-00362-6.
[44] Ait Baha T, El Hajji M, Es-Saady Y, Fadili H. The impact of educational chatbot on student learning experience. Edu Inform Technol 2023;29:10153–76. https://doi.org/10.1007/s10639-023-12166-w.
[45] Alarbi K, Halaweh M, Tairab H, Alsalhi NR, Aldarmaki F. Making a revolution in physics learning in high schools with ChatGPT: a case study in UAE. Eur J Math Sci Technol Educ 2024;20(9):em2499. https://doi.org/10.29333/ejmste/14983.
[46] Alneyadi S, Wardat Y. ChatGPT: revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. Contemp Educ Technol 2023;15(4):ep448. https://doi.org/10.30935/cedtech/13417.
[47] Kotsis KT. Correcting students' misconceptions in physics using experiments designed by ChatGPT. Eur J Contemp Educ E-Learn 2024;2(2):83–100. https://doi.org/10.59324/ejceel.2024.2(2).07.
[48] Canonigo AM. Levering AI to enhance students' conceptual understanding and confidence in mathematics. J Comput Assist Learn 2024;40(6):3215–29. https://doi.org/10.1111/jcal.13065.
[49] Martínez-Téllez R, Camacho-Zuñiga C. Enhancing mathematics education through AI chatbots in a flipped learning environment. In: 2023 World engineering education forum - global engineering deans council; 2023. p. 1–8. https://doi.org/10.1109/WEEF-GEDC59520.2023.10343838.
[50] Lademann J, Henze J, Becker-Genschow S. Augmenting learning environments using AI custom chatbots: effects on learning performance, cognitive load, and affective variables. Phys Rev Phys Educ Res 2025;21:010147. https://doi.org/10.1103/PhysRevPhysEducRes.21.010147.
[51] Luzano JFP. An integrative review of AI-powered STEM education. Int *J Acad Peadgog Res* 2024;8(4):113–8.
[52] Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative ai can harm learning. University of Pennsylvania; 2024. https://doi.org/10.2139/ssrn.4895486.
[53] Al Hakim VG, Paiman NA, Rahman MHS. Genie-on-demand: a custom AI chatbot for enhancing learning performance, self-efficacy, and technology acceptance in occupational health and safety for engineering education. Comput Appl Eng Educ 2024:e22800. https://doi.org/10.1002/cae.22800.
[54] Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychol Behav Soc Netw 2018;21(10):625–36. https://doi.org/10.1089/cyber.2018.0110.
[55] Ebadi S, Amini A. Examining the roles of social presence and human-likeness on Iranian EFL learners' motivation using artificial intelligence technology: a case of CSIEC chatbot. Interact Learn Environ 2024;32(2):655–73. https://doi.org/10.1080/10494820.2022.2096638.
[56] Albadarin Y, Saqr M, Pope N, Tukiainen M. A systematic literature review of empirical research on ChatGPT in education. Discover Edu 2024;3(60). https://doi.org/10.1007/s44217-024-00138-2.
[57] Brandhofer G, Tengler K. Zur Akzeptanz von KI-Applikationen bei Lehrenden und Lehramtsstudierenden. R&E-Source 2024;11(3):7–25. https://doi.org/10.53349/resource.2024.i3.a1277.
[58] Scherer R, Siddiq F, Tondeur J. The technology acceptance model (TAM): a meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. Comput Educ 2019;128:13–35. https://doi.org/10.1016/j.compedu.2018.09.009.
[59] Pekrun R. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educ Psychol Rev 2006;18:315–41. https://doi.org/10.1007/s10648-006-9029-9.
[60] Potvin P, Ayotte-Beaudet JP, Hasni A, Smith J, Giamellaro M, Lin T-J, Tsai C-C. Development and validation of a questionnaire to assess situational interest in a science period: a study in three cultural/linguistic contexts. Res Sci Educ 2023;53:99–120. https://doi.org/10.1007/s11165-022-10050-0.
[61] Schlegel RJ, Chu SL, Chen K, Deuermeyer E, Christy AG, Quek F. Making in the classroom: longitudinal evidence of increases in self-efficacy and STEM possible selves over time. Comput Educ 2019;142:103637. https://doi.org/10.1016/j.compedu.2019.103637.
[62] European Parliament and of the Council of the European Union. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing. Directive 95/46/EC (General Data Protection Regulation). Offic J Eur Union 2016;L 119: 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[63] fobizz. Wie gewährleistet fobizz, dass keine personenbezogenen Daten mit KI-Anbietern oder Betreibern von LLMs geteilt werden?. 2025. Retrieved from, https://app.fobizz.com/de/help/articles/376-datenschutz-ki-personenbezogene-daten [Accessed 19 May 2025].
[64] Rasiman Z, Fitriawan D. Development of socrates-based mathematical statistics teaching materials. J Didaktik Math 2023;10(1):152–69. https://doi.org/10.24815/jdm.v10i1.28881.
[65] Weber K, Inglis M, Mejia-Ramos JP. How mathematicians obtain conviction: implications for mathematics instruction and research on epistemic cognition. Educ Psychol 2014;49(1):36–58. https://doi.org/10.1080/00461520.2013.865527.
[66] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th international conference on neural information processing systems. 793; 2020. p. 9459–74. https://dl.acm.org/doi/abs/10.5555/3495724.3496517.
[67] OpenAI. OpenAI Platform. Core Concepts 2023. Retrieved from, https://platform.openai.com/docs/guides/text [Accessed 19 May 2025].
[68] Ryan RM, Deci EL. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. Am Psychol 2000;55(1):68–78. https://doi.org/10.1037//0003-066X.55.1.68.
[69] Kollmayer M, Pfaffel A, Schober B, Brandt L. Breaking away from the male stereotype of a specialist: gendered language affects performance in a thinking task. Front Psychol 2018;9:985. https://doi.org/10.3389/fpsyg.2018.00985.
[70] Sczesny S, Formanowicz M, Moser F. Can gender-fair language reduce gender stereotyping and discrimination? Front Psychol 2016;7:25. https://doi.org/10.3389/fpsyg.2016.00025.
[71] Belland BR, Walker AE, Kim NJ. A Bayesian Network Meta-Analysis to Synthesize the Influence of Contexts of Scaffolding Use on Cognitive Outcomes in STEM Education. Rev Educ Res 2017;87(6):1042–81. https://doi.org/10.3102/0034654317723009.
[72] Griesel H, Postel H, Suhr F. Elemente der Mathematik. Hannover: Schroedel; 2007.
[73] Neubert K, Benedix H. Westermann-mathematik, schuljahr 9. Braunschweig: Westermann; 1974.
[74] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 1989;13(3):319–40. https://doi.org/10.2307/249008.
[75] Ifenthaler D, Schweinbenz V. Students' acceptance of Tablet PCs in the classroom. J Res Technol Educ 2016;48(4):306–21. https://doi.org/10.1080/15391523.2016.1215172.
[76] Bailey DR, Almusharraf N, Almusharraf A. Video conferencing in the e-learning context: explaining learning outcome with the technology acceptance model. Educ Inf Technol 2022;27:7679–98. https://doi.org/10.1007/s10639-022-10949-1.
[77] Ghani MTA, Hamzah M, Ramli S, Daud WAAW, Romli TRM, Mokhtar NNM. A questionnaire-based approach on technology acceptance model for mobile digital game- based learning. J Glob Bus Soc Entrepr 2019;5(14):11–21.
[78] Schlimbach R, Robra-Bissantz S. Exploring the impact of inclusive PCA design on perceived competence, trust and diversity. In: SIGHCI 2022 Proceedings. 9; 2022. https://aisel.aisnet.org/sighci2022/9.
[79] Pekrun R, Goetz T, Frenzel AC, Barchfeld P, Perry RP. Measruing emotions in students' learning and performance: the Achievement Emotions Questionnaire (AEQ). Contemp Educ Psychol 2011;36(1):36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002.
[80] Hidi S, Renninger KA. The Four-Phase Model of Interest Development. Edu Psychol 2006;41(2):111–27. https://doi.org/10.1207/s15326985ep4102_4.
[81] Klepsch M, Schmitz F, Seufert T. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front Psychol 2017; 8:1664. https://doi.org/10.3389/fpsyg.2017.01997. –1078.
[82] Griesel H, Postel H, Suhr F. Elemente der mathematik. Hannover: Schroedel; 2007.
[83] R Core Team. R: a language and environment for statistical computing (Version 4.4.2) [Computer software]. The R Foundation; 2024. https://www.r-project.org/.
[84] Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Quarterly 2003;27(3):425–78. https://doi.org/10.2307/30036540.
[85] Marsh HW, Morin AJS, Parker PD, Kaur G. Exploratory structural equation modeling: an integration of the best features of exploratory and confirmatory factor analysis. Annu Rev Clin Psychol 2014;10:85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700.
[86] Linnenbrink-Garcia L, Durik AM, Conley AM, Barron KE, Tauer JM, Karabenick SA, Harackiewicz JM. Measuring situational interest in academic domains. Educ Psychol Meas 2010;70(4):647–71. https://doi.org/10.1177/0013164409355699.
[87] Potvin P, Hasni A, Sy O, Riopel M. Two crucial years of science and technology schooling: a longitudinal study of the major influences on and interactions between self-concept, interest, and the intention to pursue S&T. Res Sci Educ 2020;50:1739–61. https://doi.org/10.1007/s11165-018-9751-6.
[88] Blanca MJ, Arnau J, Garcìa-Castro FJ, Alarcón R, Bono R. Non-normal data in repeated measures ANOVA: impact on Type I error and power. Psicothema 2023; 35(1):21–9. https://doi.org/10.7334/psicothema2022.292.
[89] Benedetto L, Cremonesi P. Rexy, a configurable application for building virtual teaching assistants. In: Lamas D, Loizides F, Nacke L, Petrie H, Winckler M, Zaphiris P, editors. Human-computer interaction – interact 2019. Cham: Springer; 2019. p. 233–41.
[90] Daud SHM, Teo NHI, Zain NHM. e-JAVA chatbot for learning programming language: a post-pandemic alternative virtual tutor. Int J Emerg TRans Eng Res 2020;8(7):3290–8. https://doi.org/10.30534/ijeter/2020/67872020.
[91] Lee L-K, Fung Y-C, Pun Y-W, Wong K-K, Yu MT-Y, Wu N-I. Using a Multiplatform Chatbot as an Online Tutor in a University Course. In: 2020 International symposium on educational technology (ISET); 2020. p. 53–6. https://doi.org/10.1109/ISET49818.2020.00021.
[92] De Cicco R, Lacobucci S, Aquino A, Romana Alparone F, Palumbo R. Understanding users' acceptance of chatbots: an extended TAM approach. In: Følstad A, Araujo T, Papadopoulos S, Law EL-C, Luger E, Goodwin M, Brandtzaeg PB, editors. Chatbot research and design. 5th international workshop, conversations 2021. Cham: Springer; 2022. p. 3–18.
[93] Al-Gahtani SS. Empirical investigation of e-learning acceptance and assimilation: a structural equation model. Appl Comput Inform 2016;12(1):27–50. https://doi.org/10.1016/j.aci.2014.09.001.

[94] Lee YH, Hsiao C, Purnomo SH. An empirical examination of individual and system characteristics on enhancing e-learning acceptance. Austral J Educ Technol 2014;30(5):561–79. https://doi.org/10.14742/ajet.381.

[95] Tarhini A, Hone K, Liu X. Measuring the moderating effect of gender and age on e-learning acceptance in England: a structural equation modeling approach for an extended technology acceptance model. Educ Comput Res 2014;51(2):163–84. https://doi.org/10.2190/EC.51.2.b.

[96] Wollny S, Schneider J, Di Mitri D, Weidlich J, Rittberger M, Drachsler H. Are we there yet? A systematic literature review on chatbots in education. Front Artif Intell 2021;4:654924. https://doi.org/10.3389/frai.2021.654924.

[97] Loderer K, Pekrun R, Lester JC. Beyond cold technology: a systematic review and meta-analysis on emotions in technology-based learning environments. Learn Instr 2020;70:101162. https://doi.org/10.1016/j.learninstruc.2018.08.002.

[98] Kulkarni A, Chu SL, Sharma N, Sathe S. Interest-based learning through a contextualizing Chatbot for video-based online learning platforms. In: International conference on advanced learning technologies, ICALT; 2022. https://doi.org/10.1109/ICALT55010.2022.00027.

[99] Kumar JA. Educational chatbots for project-based learning: investigating learning outcomes for a teambased design course. Int J Educ Technol High Educ 2021. https://doi.org/10.1186/s41239-021-00302-w.

[100] Sweller J, van Merriënboer JJG, Paas F. Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev 2019;31:261–92. https://doi.org/10.1007/s10648-019-09465-5.

[101] Deng R, Jiang M, Yu X, Lu Y, Liu S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Comput Educ 2025;227:105224. https://doi.org/10.1016/j.compedu.2024.105224.

[102] Kalai, A.T., Nachum, O., Vempala, S.S., & Zhang, E. (2025). Why language models hallucinate. *arXiv* [Preprint]. https://doi.org/10.48550/arXiv.2509.04664.

[103] Huwer J, Becker-Genschow S, Thyssen C, Jochen Thoms L, v. Kotzebue L, Finger A, Kremser E, Meier M, Bruckermann T. Kompetenzen für den Unterricht mit und über Künstliche Intelligenz. Perspektiven, Orientierungshilfen und Praxisbeispiele für die Lehramtsausbildung in den Naturwissenschaften. Waxmann; 2024.

[104] UNESCO. AI Competency Framework for Teachers. UNESCO. https://doi.org/10.54675/ZJTE2084.