Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables

Julia Lademann,[1,*] Jannik Henze,[2] and Sebastian Becker-Genschow[1]

[1]*Faculty of Mathematics and Natural Sciences, Digital Education Research, University of Cologne, 50931 Cologne, Germany*

[2]*Faculty of Mathematics and Natural Sciences, Institute for Physics Education, University of Cologne, 50931 Cologne, Germany*

(Received 18 December 2024; accepted 25 March 2025; published 7 May 2025)

This work explores the integration of artificial intelligence (AI) custom chatbots in educational settings, with a particular focus on their applicability in the context of mathematics and physics. In view of the increasing deployment of AI tools such as ChatGPT in educational contexts, the present study explores their potential in generating topic-related learning material. The study assesses the impact of learning with AI-generated explanations as Supplemental Material on the learning experiences and performance of sixth-grade students, with a particular focus on proportional relationships in mathematical and physical contexts. The randomized controlled study with $N = 214$ students compared supplementary learning material in the form of traditional textbook material with explanations previously generated by an AI custom chatbot. The results demonstrated that while the AI-generated materials had an indefinite impact on learning outcomes, they significantly enhanced positive-activating emotions, situational interest, and self-efficacy while reducing intrinsic and extrinsic cognitive load. These findings underscore the potential of AI to transform educational practices by fostering a superior learning experience. However, further research is required to clarify its impact on learning performance and long-term learning outcomes. The study highlights the importance of careful integration and customization of AI tools to maximize their benefits in physics education.

DOI: 10.1103/PhysRevPhysEducRes.21.010147

I. INTRODUCTION

Since the release of ChatGPT in 2022, there has been growing attention to how AI chatbots can be used profitably in an educational context, and the integration of artificial intelligence (AI) is emerging as a key factor at both the school and university levels [1]. For instance, AI tools can be utilized by educators to develop lesson plans, create differentiated learning materials, and provide feedback. Additionally, they can be employed by students as intelligent tutoring systems, research, or writing tools [2–5]. Especially, large language models (LLMs) have the potential to transform educational experiences, particularly in specialized fields such as physics [1]. LLMs are a category of artificial intelligence based on neural networks, trained on vast datasets. Their objective is to comprehend and generate text in a manner that imitates human communication, thereby opening up a multitude of opportunities for educational applications [3]. By enabling more personalized and interactive learning approaches, these models have the potential to transform the way students learn. One of the most widely recognized applications of LLMs is in the field of chatbots, which are software tools that engage with the user in written dialogue [1]. In STEM education, chatbots can be utilized as learning assistants to provide feedback on exercises and individual topic-related questions. They can generate a variety of additional exercises, which also can be adapted to the individual student's learning status. In the natural sciences, chatbots can assist students during experiments [2,6–9]. However, despite this promise, the question of how exactly learning with AI chatbots affects students' learning performance and experience remains unanswered [5]. This indicates the necessity for further research in this field. The present study thus seeks to address this question by initially focusing on the question of whether AI chatbots are capable of designing engaging materials for self-learning. The effects on emotional aspects, cognitive load, self-efficacy expectations, and situational interest, in particular, are to be measured while learning with an AI-generated explanation of a selected topic.

*Contact author: julia.lademann@uni-koeln.de

Regarding the use of AI chatbots in schools, challenges, including a lack of in-depth contextual understanding, the difficulty in assessing the quality of responses, and a deficiency in higher-order cognitive abilities are mentioned [10]. For instance, learners frequently utilize chatbots without reflection when engaged in physics tasks, without attempting to solve the tasks independently [11]. These findings highlight the necessity for an in-depth analysis of responses generated by ChatGPT, particularly with regard to their scientific veracity, by both instructors and students [12] and underscore the need for cautious incorporation of AI into the education system by conducting a thorough and comprehensive evaluation of this potentially profitable tool.

The present study explores the impact of AI-generated explanations on students' learning experiences and performance in physics. The AI-generated explanations were previously generated by an AI custom chatbot. The aim of this study is to conduct a comparative analysis of traditional and AI-generated supplementary learning materials.

The primary findings of the study indicate that the use of AI-generated material has a notable impact on the learning experience of students. In the experimental group, higher levels of positive-activating emotions, self-efficacy expectations, and situational interest were measured. Conversely, the levels of intrinsic and extrinsic cognitive load were significantly lower.

II. STATE OF RESEARCH

A. Context of physics teaching

Students' comprehension problems in physics lessons are often related to a lack of mathematical skills [13–15]. However, these difficulties are also frequently attributed to the challenge of transferring the skills acquired in mathematics to the physical context [16,17]. The causes of these difficulties are not uniform across all students—they may arise from misconceptions about the physical meaning of, for example, ratios, products, functions, and neutral elements [18].

Moreover, in the fields of physics and mathematics, comprehension problems potentially emerge from difficulties in interpreting and linking different forms of representation, which can be defined as multiple representations [19]. These forms of representation, including formulas, graphs, diagrams, and tables, are a means of encoding important information. It is therefore beneficial for learners to possess the capacity to interpret and transition between the various forms, thereby facilitating a full comprehension of scientific concepts [19–21]. By providing targeted assistance in establishing connections between the structures of depicted forms of representation, students are more likely to develop a better understanding of the concepts and are less likely to hold misconceptions [20].

As an example, the topic "Proportional Relationships" was chosen for this study. It is taught in seventh-grade mathematics in Germany. A proportional relationship can be represented in various forms (graphs, tables, formulas), and students must develop the ability to switch between these representations with ease. To do so, students have to interpret the mathematical models from a physics point of view or translate physical behavior into a mathematical context [22,23]. AI chatbots with vision capabilities like ChatGPT may offer a potential solution to this issue [2,3,5], as they can read and analyze images of representations. In this manner, learners can be aided in both the interpretation of discrete forms of representation and in the recognition of encoded information and the relationships between different forms of representation.

In the context of educational materials and environments, a multitude of factors influence the efficacy of learning in classroom settings. Primarily, achievement emotions perceived during the learning process can exert a positive or negative influence. Emotions such as enjoyment, hope, pride, anger, anxiety, shame, hopelessness, and boredom play a critical role in students' motivation and learning performance [24].

Moreover, situational interest is regarded as an important element in the learning process. Contrary to individual interest, which is defined as a permanent motivational tendency, situational interest is initially temporary and is associated with a specific learning situation [25]. The research literature on the subject demonstrates a divergence of trends with respect to the relationship between situational interest and students' learning performance. Studies indicate that situational interest exerts a positive effect on learning performance [26], while others demonstrate an absence of a measurable effect [27].

The model for cognitive load on which this work is based is a two-factorial model for an intervention-induced cognitive load [28]. Cognitive load, caused by the limitations of working memory, is measured as intrinsic cognitive load (ICL) and extrinsic cognitive load (ECL). ICL refers to the complexity of the information to be processed and is consequently influenced by the task and the prior knowledge of the learner. ECL can be influenced by the design of the learning environment and material and refers to cognitive processes that do not lead to a significant learning outcome [29]. Cognitive load theory (CLT) posits the notion that the capacity of working memory is constrained in comparison to that of long-term memory [30,31]. The underlying objective of CLT in instructional design is to design learning environments in a manner that optimizes the utilization of working memory resources and minimizes learning-irrelevant cognitive load [32–34]. In the case of a successful implementation, learners find themselves with augmented cognitive resources in their working memory, a condition that fosters the active construction of new knowledge and enhances the learning outcome.

The concept of self-efficacy expectation describes a person's subjective conviction that they can perform specific actions successfully. Students with higher self-efficacy expectations are characterized by a higher level of ambition and set themselves more challenging goals. They are less likely to give up when faced with difficulties and invest more time and effort in solving complex problems. This enables more in-depth learning, which ultimately results in affecting learning performance [35]. Positive self-efficacy expectations have been demonstrated to encourage motivation to engage with novel and challenging tasks and to exert effort in doing so [36].

This study explores the extent to which learning with AI-generated material as Supplemental Material can promote the aforementioned concepts and thereby contribute to enhanced learning performance in physics lessons.

B. Textual explanations

In this study, an AI-generated textual explanation on the topic of "Proportional Relationships" is compared with nontextual textbook material in its function as Supplemental Material. A text-based explanation should fulfill certain characteristics. It should be as precise and coherent as possible, explaining the core concepts and principles while omitting irrelevant details [37]. The inclusion of excessive detail has been demonstrated to increase cognitive load, thereby diminishing the effectiveness of provided explanations. Consequently, it is imperative to prioritize the relevance of the information contained within these explanations [38]. In addition, overly dense text structures have the potential to impede comprehension, especially for students with limited prior knowledge [39]. In order to ensure efficacy and maximize cognitive engagement, it is essential that explanations be adapted to align with the learners' current knowledge level and interests [37,40]. Research has demonstrated that an absence of adaptation to existing knowledge and misconceptions can impede the efficacy of explanations. In this regard, the utilization of AI custom chatbots can prove beneficial, as they possess the capability to adapt their explanations to the given context. Explanations, tailored to the learners' existing knowledge level, have been shown to be advantageous for students [38].

C. AI chatbots in education

1. Learning experience and learning performance

The integration of AI into the field of education is rapidly transforming teaching and learning experiences. AI is gaining recognition for its ability to improve educational outcomes, thereby fostering a growing interest in research within the domain of AI in education in recent years [41,42]. Generative AI, in particular, opens up new opportunities for the design of teaching and learning environments through the personalized assistance of chatbots. AI chatbots can provide learners with individualized feedback and address specific learning difficulties and questions [2,3,5]. They have the potential to enhance student productivity and foster motivation [4,43]. Moreover, a meta-analysis demonstrates that AI chatbots, in general, can significantly impact student learning outcomes [44] and result in advantages such as better access to information and a simplification of personalized and complex learning [10]. ChatGPT for instance, is not only capable of solving tasks but can also explain solutions and approaches and create tasks itself [7,8]. This additionally gives learners the opportunity to generate a variety of exercises and to comprehend solution paths. Moreover, personal chatbots can be configured to operate as so-called custom chatbots, thereby providing users with assistance and feedback.

However, particularly within the school context, the potential of AI chatbots, especially custom chatbots, as intelligent tutoring assistants is still largely underexplored. One remaining question is how learners perceive AI chatbot-generated feedback and their interaction with AI chatbots, and whether it influences the learning process [5]. An increasing number of studies are investigating the influence of the use of AI chatbots, in particular ChatGPT, on the learning experience and learning performance.

It was found that the utilization of ChatGPT fosters a low-pressure environment, thereby encouraging learners to seek further clarification and assistance with greater ease [45]. Learners exhibited greater satisfaction due to the assurance of privacy. Additionally, learning with chatbots may result in the cultivation of positive emotions and an increased sense of well being [46]. In the field of mathematics, the utilization of ChatGPT has been demonstrated to enhance self-efficacy and to improve the development of conceptual understanding [47]. Similarly, in the field of physics, ChatGPT has been shown to positively influence both learning experiences and learning performance [48,49]. Moreover, it was found that employing ChatGPT for experimentation in school helped to correct misconceptions in physics and facilitated a more profound understanding of physical concepts [9]. Concerning research methods, it was shown by a comparative analysis between the utilization of ChatGPT and conventional search engines that LLMs can reduce the mental effort demanded by learners in research tasks, yet simultaneously result in a decline in the quality of reasoning and conclusions [50].

At this point, it is unclear to what extent learners can identify errors in AI-generated content and whether learning with ChatGPT fosters critical thinking or, conversely, reinforces misconceptions [5]. The persuasive manner in which ChatGPT presents information can obscure inaccuracies. Therefore, learners should not rely on it as their sole source of information [51]. Nevertheless, nearly half of the students enrolled in an undergraduate-level introductory physics course expressed confidence in the ability of

ChatGPT to provide accurate responses, regardless of their accuracy [52]. Conversely, the utilization of an AI assistant in the context of experimentation presents an effective method for the rectification and resolution of potential misconceptions within the field of physics education, as well as facilitating a more profound comprehension of fundamental physics concepts [9].

In conclusion, AI chatbots such as ChatGPT are valuable tools, but their integration into physics teaching must be handled with care. They offer the advantage of enabling personalized learning and reducing the workload of teachers. However, despite these advantages, further research is needed to identify successful didactic concepts and meaningful implementations in physics teaching [1].

2. The capabilities of ChatGPT in physics

When discussing the potential use of AI chatbots in an educational setting, the question of the accuracy of AI-generated responses cannot be overlooked. These should provide accurate answers to technical queries with a high degree of probability. This is particularly crucial in mathematical and physical contexts, as errors can be easily overlooked and are challenging for students to verify. Various studies are currently examining the capabilities of AI chatbots in a physical context.

For instance, ChatGPT-4 demonstrated high accuracy in solving the FCI as well as concept tasks in the field of mechanics. Its performance exceeded that of engineering students [53]. Even ChatGPT-3, which was able to pass an introductory physics course, only made mistakes that resemble those of beginners [54]. Subsequently, it was demonstrated that ChatGPT-4 has achieved a significantly higher score in the aforementioned context. Indeed, the responses exhibit a level of competence that is nearly indistinguishable from that of an expert, with a few notable exceptions and limitations [55]. The studies demonstrate that ChatGPT-4 has proficient fundamental physical capabilities. However, they also indicate that there is an upward trend in performance across the models. Nevertheless, a more profound comprehension is occasionally absent, particularly in the capacity to prove theorems or derive physical laws [51].

As a vision-capable chatbot, ChatGPT-4 can also analyze images. This can be particularly interesting for mathematics and physics, since graphs, tables, and formulas are often difficult to describe or transfer into the chatbot's text field. However, a study examining ChatGPT-4's capacity to solve the TUK test revealed that while ChatGPT-4 typically accurately describes the approaches to solving graph-related tasks, it exhibits deficiencies in graph analysis. For instance, it fails to discern intersections with the axes correctly [56]. However, a trend toward improvement could be seen here as well, since ChatGPT-4o was able to achieve significantly better results than its predecessor ChatGPT-4 and also outperforms all other tested models [57]. This observation also shows a successive improvement in LLMs for the physical learning context.

3. Custom chatbots

Despite the existence of a multitude of chatbots, it is often necessary to adapt them to achieve the desired performance in a given application [53]. A study with electrical engineering students showed that learning with an AI custom chatbot (based on ChatGPT) significantly improved learning performance and self-efficacy, both in comparison to traditional learning methods and to learning with ChatGPT. Students reported heightened confidence and effectiveness in utilizing the custom chatbot as a learning assistant [58]. This motivates a systematic configuration and evaluation of chatbots for implementation in educational settings.

For the educational context, a good option to tailor a chatbot is called augmentation. This method does not require detailed knowledge in machine learning and neural networks [59]. The chatbots are based on already existing LLMs and adapted to a specific purpose. A way of augmentation is called retrieval augmented generation (RAG), which is a method used to enhance and specify the capabilities of LLMs by integrating external data sources [59,60]. Documents that serve as the external knowledge base are divided into smaller segments and converted into an embedding. When a user submits a prompt, it is compared with the document embeddings. These selected segments are then combined with the user's prompt and sent to the LLM as a new prompt reading: "Reply to [user prompt] using the following background materials: [relevant text segments]." [59]. The LLM then generates a response based on both the user prompt and the external knowledge. RAG provides a practical method for customization for specific use cases by curating the database with relevant materials. The reliance on external documents reduces the risk of hallucinations [59]. Therefore, in educational contexts, RAG can be employed to create customized chatbots that provide precise answers based on course-specific materials, such as lecture notes, syllabi, or problem sets [59]. A variety of platforms, for example, operated by OpenAI[1] and Anthropic,[2] offer the possibility of augmentation based on their respective LLM, thereby enabling the creation of custom chatbots aligned with specific requirements and domains of use. This is a relatively straightforward process for teachers or students to configure their own custom chatbots. The custom chatbots at OpenAI are designated as GPTs. The users may instruct their GPT to behave and react in a desired manner, such as using situation-specific language or a certain length of

[1]https://openai.com/index/introducing-gpts/accessed: 2024-12-14.

[2]https://docs.anthropic.com/en/docs/build-with-claude/tool-use/ (accessed: December 14, 2024).

answers, by means of a configuration prompt that they define themselves. Moreover, subject- or situation-specific knowledge can be provided to the GPT by uploading corresponding files.

For this study, we decided to use the method of augmentation and OpenAI to customize an own GPT (see Sec. IV A 3.).

III. RESEARCH QUESTIONS

The current state of research implies that further studies in this still new and under-researched area are both useful and imperative. AI chatbots have the potential to address learning deficiencies among students and facilitate personalized learning. However, given the current limitations of research in this area, further inquiry is necessary to gain a more comprehensive understanding of the impact of generative AI on students' learning experiences and learning outcomes. As previously stated in Sec. II A, the constructs and concepts, such as self-efficacy expectations and cognitive load, have been demonstrated to have a significant influence on learning performance in school. However, at this point it is open to question whether chatbots can be used to design suitable learning materials and environments and generate text-based explanations that positively influence the aforementioned constructs.

In this study, we used a customized chatbot to generate explanations for graphical representations (see Sec. IV A 3). The study is designed as a field study, with students as the target group and the objective of addressing the following key research questions:

How does learning with AI-generated explanations using a custom chatbot affect

RQ1: Emotional aspects?
RQ2: Situational interest?
RQ3: Cognitive load?
RQ4: Self-efficacy expectations?
RQ5: Learning performance in a mathematical and physical learning context?

IV. METHODS

A. Study design

1. Sample

A total of $N = 214$ sixth-grade students (146 female, 66 male, 2 n.a.) at secondary schools in Germany participated voluntarily in the randomized controlled study. The average age of the students was 11.7 years (SD = 0.51).

2. Procedure

The students were randomly assigned to either the experimental (EG) or the control group (CG). Both groups were provided with learning materials on the topic of "Proportional Relationships" in a mathematical and physical context. The topic had not been covered in class before the study was conducted. The CG was provided with conventional textbook material, comprising a topic overview and explanatory examples. The EG was also provided with the topic overview from the textbook. Instead of the aforementioned examples, the participants were presented with an explanation of the topic, which was generated previously by an AI chatbot (see Sec. IV A 4) using a previously defined prompt. To eliminate potential bias caused by the use of digital media, both groups engaged in their studies exclusively with paper-based materials. To ensure a fair comparison, the learning time was identical for both groups and amounted to 15 min. Moreover, the learners were not informed in advance about the specific type of learning material they would be using.

Prior to the learning phase, demographic data and individual, subject-specific interest in mathematics were collected. This was done in order to ascertain whether there was a discrepancy in mathematical interest between the two groups. Immediately following the learning phase, the dependent variables, as outlined in Sec. IV A 5, were collected sequentially by category. Participants' emotions and situational interest were recorded. In addition, the intrinsic and extrinsic cognitive load when learning with the materials and the self-efficacy expectation with regard to solving a topic-related task were assessed. Subsequently, in order to measure learning performance, the learners completed a performance test (see Sec. IV A 6) within 30 min.

3. Chatbot design

The custom GPT was configured in OpenAI and adapted to the specific learning context and target group. Furthermore, the chatbot specializes in the analyzing and interpreting various forms of representation for proportional relationships, including graphs, tables, and formulas. It uses language that is both engaging and age appropriate, making it a suitable tool for students in the sixth grade of secondary school. The chatbot is capable of providing targeted and individualized support in both learning mathematical content and applying mathematics in physics lessons.

The chatbot was configured in multiple cycles based on the OpenAI GPT-4.0 model. Following each configuration phase, a test phase was conducted to systematically assess whether the chatbot's responses met the predefined criteria. Based on the results of this evaluation, the configuration prompt was successively adjusted after each test phase. In addition, the chatbot was equipped with subject knowledge that corresponds to the core curriculum for the respective grade. In evaluating the quality of the chatbot, particular attention was paid to ensuring that the answers were factually correct and corresponded to the notations used in common textbooks. Additionally, care was taken to ensure that the language used contained motivating

vocabulary and was age appropriate, assuming a learning group between the ages of 10 and 13.

4. Material

Schoolbook material. Both the control and the experimental group received an overview of the topic "Proportional Relationships," taken from a schoolbook [61] for the learning phase. The CG was also provided with additional examples from the same schoolbook that illustrate and explain the topic, thus compensating for the AI-generated explanation of the topic that was provided to the EG.

The textbook material (see Supplemental Material, Chapters A.a. and A.c. [62]) was selected by a group of experts according to the following criteria, among others:

- A structured overview of the given topic and the forms of representation (graphs, tables, and formulas).
- An overview that does not contain many additional explanations.
- The selected examples are consistent, follow on from the content of the overview (chosen from the same book), and provide a sufficiently in-depth presentation of the topic.

AI-generated explanation. In selecting the AI-generated explanation, particular consideration was given to the following criteria:

- The explanation is accurate and complete.
- All three forms of presentation are adequately explained.
- No terms are used that are unfamiliar to the students.
- No discriminatory language is used.

To achieve a satisfactory output, a prompt was initiated, reiterated, and refined until it resulted in the final selected explanation (see Supplemental Material, Chapter A.b. [62]). In addition to the prompt, the custom chatbot was presented with an image of the textbook overview.

The textbook material and the AI-generated explanation can be found in the Supplemental Material (Chapter A [62]).

5. Data collection

Subject-specific interest. Interest in mathematics as a school subject was measured using five items and, like all subsequent variables, evaluated on a four-point Likert scale. Subject-specific interest was measured before the learning phase and served to examine the comparability of the CG and EG. A scale was utilized, based on the one developed by Rakoczy *et al.* [63]. The scale was shortened from eight to five items for reasons of test economy, and the remaining items were adapted to the specific learning context.

Emotions. To assess the emotional state of the learners during the learning phase, positive-activating (two items: pleasure, satisfaction) and negative-deactivating emotions (three items: boredom, frustration, and uncertainty) were measured retrospectively, directly after the learning phase. For this purpose, five items were selected and translated into German based on the achievement emotions questionnaire (AEQ) [29,64,65].

Situational interest. It was measured with four items following the students' engagement with the learning materials. In order to maintain a low number of items for the age group, four items were chosen based on a reliable and validated scale by Linnenbrink-Garcia *et al.* [66], translated into German, and adapted to the learning context.

Cognitive load. To measure the cognitive load during the learning process, a two-factorial model of cognitive load was used in this work, measuring intrinsic cognitive load (ICL) and extrinsic cognitive load (ECL). Although based on a three-factorial model, the 10-item questionnaire developed and validated by Leppink *et al.* [28] was utilized (cognitive load scale; CLS) to ensure the validity of the measurement. Three items each were used to assess intrinsic cognitive load (ICL) and extrinsic cognitive load (ECL). These six items were translated into German and adapted to the specific context of the learning environment. Furthermore, the original response scale, which ranged from 0 to 10, was reduced to levels 1 to 4 for reasons of consistency with the rest of the questionnaire.

Self-efficacy expectation. It was evaluated in regard to the independent solving of a topic-related task on the basis of a scale comprising five items. The concept of self-efficacy expectation was first developed by Bandura [35] and the German adaptation was conducted by Schwarzer and Jerusalem [67]. The items utilized in this study are derived from the scale by Jerusalem and Satow [68].

6. Performance test

Following the examination of motivational aspects, the students completed a performance test (see Supplemental Material, Chapter B. [62]) that addressed the physical concept of proportional relationships using the "spring scale" as an example. The tasks were developed internally based on standardized textbook tasks. The objective of the performance test was to ascertain whether the students had acquired an understanding of the various ways in which proportional relationships can be represented and whether they were able to switch between these different forms of representation. The test is composed of four subtasks A, B, C, and D, after which students were requested to indicate the degree of certainty associated with their response. They may select one of four options: "completely sure," "rather sure," "unsure," or "guess." In the interest of ensuring the integrity of the evaluation process, students who had not completed a task are instructed to select "guess" to exclude their responses from being considered during the subsequent evaluation.

B. Data analysis

All analyses were conducted using R version 4.4.0 [69].

1. Internal consistency

Cronbach's alpha was calculated for the purpose of evaluating the internal consistency of the individual scales (R package "PSYCH," version 2.4.6.26, function "alpha") [70]. In this context, values with $\alpha > 0.7$ are considered acceptable, values with $\alpha > 0.8$ are considered good, and values with $0.9 < \alpha < 1.0$ excellent [71,72].

As the values for Cronbach's alpha are acceptable to excellent (Table I) and no significant improvements were identified when individual items were removed from the scales, the scales will be retained for further data analysis. In order to proceed with the evaluation, the mean of the recorded values of the items on each scale was calculated for all participants. These resulting scores formed the basis for the subsequent data evaluation.

2. Test for normal distribution

Prior to testing the data of the experimental and control groups for significant deviations, the entire dataset was initially evaluated for normal distribution. This is done to enable the drawing of conclusions regarding the applicability of specific statistical tests. Both the data of the EG and those of the CG were tested for normal distribution using the Shapiro-Wilk test with a significance level $\alpha = 0.05$ (R package "STATS," function "shapiro.test") [69]. In accordance with the selected significance level, a normal distribution of the data can be assumed if the p value of the Shapiro-Wilk test is $p > 0.05$. However, this is only the case for the self-efficacy score in the experimental group for the given dataset ($p = 0.101$) (see Supplemental Material, Chapter F.a. [62]). No variable demonstrated a normal distribution in both groups, therefore, nonparametric tests were employed.

3. Test for significant deviations

To test for significant differences between the experimental and control group scores, the Mann-Whitney U test with significance level $\alpha = 0.05$ was used (R package "STATS," function "wilcox.test") [69]. This test is standardized for two independent non-normally distributed samples. Cohen's d was calculated to estimate the effect size for significant differences (R package "PSYCH," version 2.4.6.26, function "cohen.d") [70].

TABLE I. Cronbach's alpha indicating scale consistency.

Scale	$\alpha =$	
Interest in mathematics	0.92	Excellent
Positive-activating emotions	0.71	Acceptable
Negative-deactivating emotions	0.76	Acceptable
Intrinsic cognitive load	0.73	Acceptable
Extrinsic cognitive load	0.72	Acceptable
Situational interest	0.84	Good
Self-efficacy	0.90	Excellent

4. Evaluation of the performance test

Interrater reliability. It was not possible to determine a clear allocation of points for one of the subtasks (task D). The task required the creation of a graph within a coordinate system. A rating system of six categories was devised for the evaluation of the task (see Supplemental Material, Chapter C. [62]) and a second, independent rater was consulted. To ascertain the interrater reliability, Cohen's weighted kappa was computed for the two raters and for each category of the rating system (R package "IRR," version 0.84.1, function "kappa2") [73]. Quadratic weighting was selected to account for the discrepancy in point allocations across the rating scale (the difference between 0 and 2 carries greater weight than that between 0 and 1). Computing the mean value of the interrater reliability from all categories, results in a mean value of $\kappa = 0.8469$, which indicates a very good degree of agreement between the two raters for task D [74–76].[3] To further evaluate the data, the mean of the two ratings was calculated and taken as the score for task D.

Filtering data and calculating scores. As previously outlined in Sec. IV A 6, the students were asked to evaluate their confidence in their responses following the completion of each subtask. Consequently, the points awarded for solutions marked as "guessed" were set to zero. Furthermore, an item designated as the "overall score" was introduced for the performance test, wherein all subtasks were given equal weighting.

Test for normal distribution. Prior to the group comparison, the data for the performance test were also evaluated for normal distribution using the Shapiro-Wilk test with $\alpha = 0.05$ (R package "STATS," function "shapiro.test") [69]. The results of this analysis indicated that the data did not meet the criteria for a normal distribution.

Test for significant deviations. Given that the data from the performance test were not normally distributed, the Mann-Whitney U test for independent, non-normally distributed samples was also applied here for the comparative analysis of the groups (R package "STATS," function "wilcox.test") [69]. Cohen's d was calculated to estimate the effect size for significant differences (R package "PSYCH," version 2.4.6.26, function "cohen.d") [70].

[3]A value of 0.60 or higher is considered "good" and a value of 0.8 or higher is considered "very good" or "almost perfect."

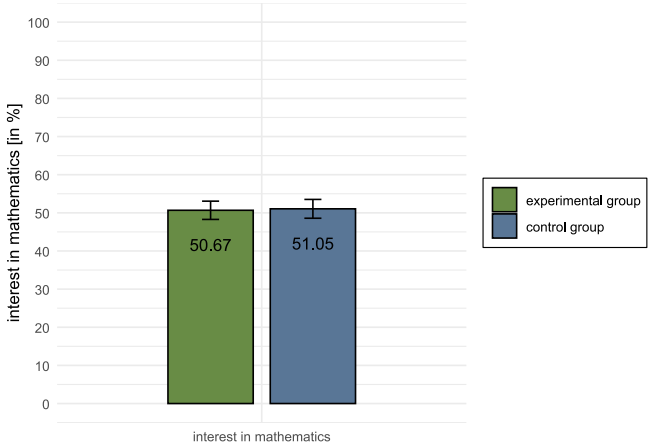

FIG. 1. Interest in mathematics with standard error.

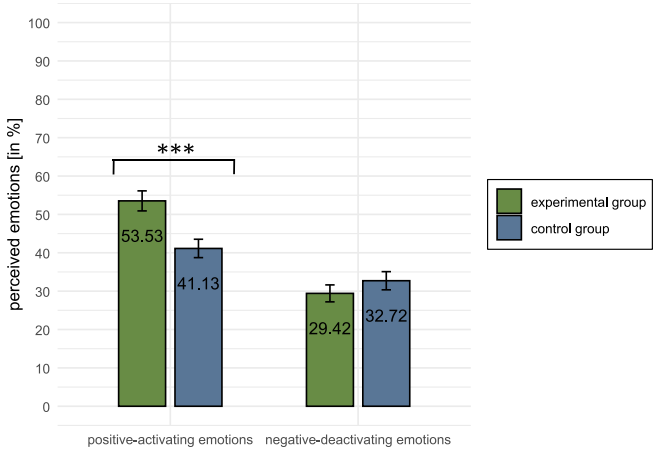

FIG. 2. Perceived emotions with standard error and level of significance.

V. RESULTS

A. Mathematics—grade

To ensure a variable to assess the comparability of the two groups with regard to performance in math classes, the math grade from the last term's report card was requested.[4] The CG exhibits slightly superior mathematical performance ($M = 2.24$; $\mathrm{SD} = 0.88$) than the EG ($M = 2.38$; $\mathrm{SD} = 0.95$), but the Mann-Whitney U test revealed no statistically significant discrepancy between the two groups with respect to their mathematical performance ($p = 0.252 > 0.05$).

B. Measured variables

1. Subject-specific interest

Subject-specific interest was evaluated prior to the introduction of the instructional materials. Therefore, the assignment to a group could not have any effect on the measured results at this point (Fig. 1). The collected data show no significant differences between the two groups in terms of interest in mathematics. The Mann-Whitney U test calculated $p = 0.961$, indicating that there was no statistically significant difference between the two groups.[5]

2. Emotions

The elicited emotions were classified into two categories: positive activating and negative deactivating. For the former, a statistically significant difference ($p = 0.00093$, $d = 0.48$, small effect size) was observed in favor of the EG.

No significant deviation could be detected for negative-deactivating emotions ($p = 0.391$). Only a trend can be identified, which shows that the CG tended to perceive more negative emotions (Fig. 2).

3. Situational interest

The measured situational interest was significantly higher in the EG ($p = 0.00223$, $d = 0.45$, small effect size) than in the CG (Fig. 3).

4. Cognitive load

Cognitive load was measured as intrinsic cognitive load (ICL) and extrinsic cognitive load (ECL). In both cases, the CG experienced significantly greater cognitive load than the EG (Fig. 4):

The analysis showed that a significant deviation ($p = 0.00060$, $d = 0.47$, small effect size) was calculated for ICL, and that a significant deviation ($p = 0.0001$, $d = 0.59$, medium effect size) was calculated for ECL.

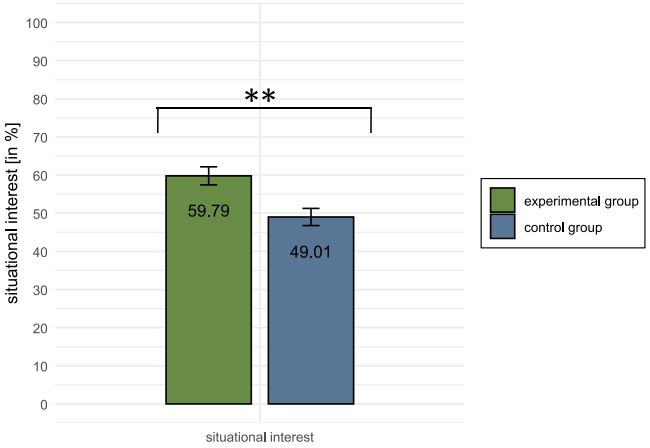

FIG. 3. Situational interest with standard error and level of significance.

[4]In Germany, grades range from 1 to 6, with 1 representing the highest level of achievement and 6 indicating the lowest.

[5]The complete results of the Mann-Whitney U test, including those relating to individual items, can be found in the Supplemental Material, Chapter F.b. [62].

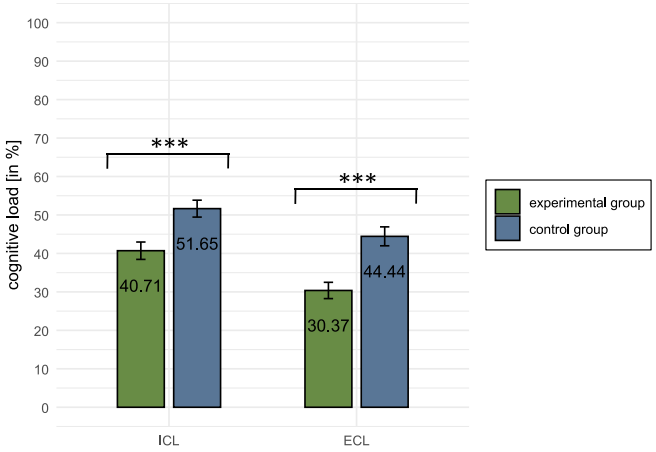

FIG. 4. Intrinsic cognitive load (ICL) left and extrinsic cognitive load (ECL) right, both with standard error and level of significance.

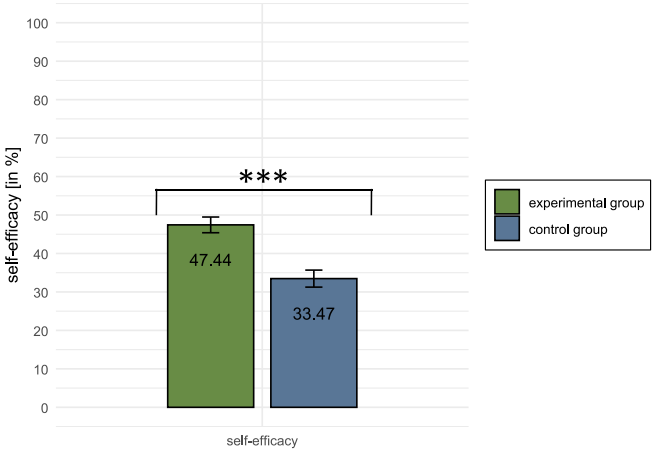

FIG. 5. Self-efficacy with standard error and level of significance.

5. Self-efficacy expectation

Self-efficacy was measured in terms of solving a topic-related task. The values of the EG were found to be significantly higher than those of the CG ($p = 0.00001$, $d = 0.63$, medium effect size).

The results demonstrate that students who engaged in learning activities with AI-generated explanations exhibited higher levels of self-efficacy than those who utilized textbook materials exclusively (Fig. 5).

C. Performance test

For the overall test score, no significant difference between the two groups could be determined ($p = 0.750$) (Fig. 6).

VI. DISCUSSION

As no significant differences were observed between the EG and CG with respect to mathematics grade or subject-specific interest, these two groups were found to be comparable in this regard.

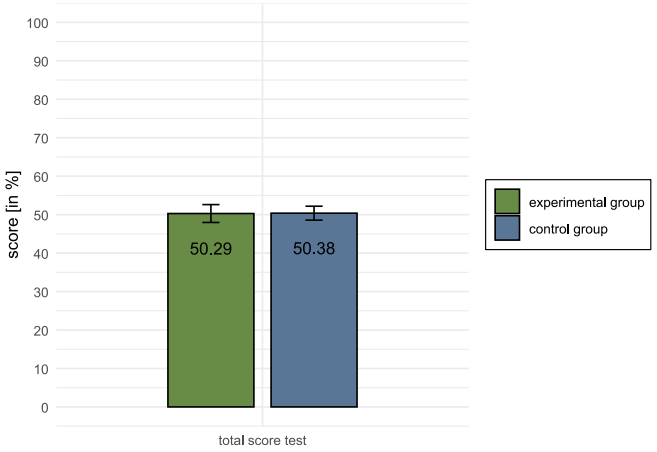

FIG. 6. Overall score of the test with standard error.

Concerning the influence of emotional aspects (RQ1), the EG exhibited a significantly higher incidence of positive-activating emotions (pleasure and satisfaction) than the CG. This finding is consistent with the results of previous research [46]. With regard to negative-deactivating emotions (boredom, frustration, and uncertainty), only a trend could be identified that suggests that learning with AI-generated explanations may potentially contribute to a reduction of negative emotions. In sum, the results of the evaluation indicate that learning with AI-generated explanations as Supplemental Material to textbooks can have a positive impact on the emotional states of the learners.

Regarding the influence on situational interest (RQ2), it was found to be significantly higher in the EG than in the CG. When learning with the explanation generated by the custom chatbot as Supplemental Material, a higher situational interest could be triggered than when learning with the textbook materials only. One potential explanation for this outcome is the nature of the custom chatbot's tone, which is designed to be encouraging and motivating. For instance, the chatbot uses encouraging questions such as "Has that become a bit clearer?" to ensure comprehension of the content or motivating phrases such as "Let's look at that …" (see Supplemental Material, Chapter A.b. [62]).

The results concerning the effects on cognitive load (RQ3) show significant differences for both intrinsic cognitive load (ICL) and extrinsic cognitive load (ECL)—in both cases, the values were higher in the CG. These findings suggest that learning with textbooks and supplementary AI-generated materials reduces cognitive load compared to learning with textbook-only materials. One potential explanation for this outcome is the use of accessible language and a conversational and humanlike tone in the chatbot's responses (Fig. 7).

As mentioned in Sec. II B, the quality of textual explanations has been demonstrated to have a significant

> So, a proportional relationship means that two quantities are in a constant ratio to each other. This means that if one quantity increases, then the other quantity also increases in such a way that the ratio (the result of the division) of both quantities always remains the same.
>
> Let's look at that using the materials you have here:
> […]

FIG. 7. Introductory text of the AI-generated explanation of the learning material using a conversational and human-like tone (see Supplemental Material, Chapter A.b. [62]).

influence on cognitive load. It is imperative that explanations not only be precise and avoid excessive details but also be aligned with the learners' current knowledge level and interests. The findings suggest that the significantly lower cognitive load, particularly the ECL, may be indicative of the custom chatbot's effective use in generating an explanation adapted to the knowledge level and personal needs of sixth-grade students.

Regarding research question RQ4, which concerned self-efficacy expectations, the EG showed significantly higher values than the CG. This indicates that the explanation of the AI chatbot can foster students' confidence in their ability to complete tasks and corresponds to the findings of Canonigo in mathematics [47]. A reason for this may be the encouraging tone of the AI-generated text, the reduced cognitive load, or a combination of both.

The fifth research question (RQ5) related to the influence of the learning material on learning performance in the mathematical and physical learning context. No clear trend could be identified in the overall result of the performance test. The data collected in this study were insufficient to permit a definitive conclusion and answer to RQ5. A larger-scale study in which students could utilize the chatbot independently and individually would be beneficial to enable more in-depth observations and allow for a higher degree of individualization of the feedback. Prior research has demonstrated that the utilization of chatbots can enhance learning outcomes [9,44,47–49]. Whether there is a connection between the learning performance and the previously measured variables cannot be definitively determined on the basis of the present study. Despite the fact that the experimental group exhibited a notable reduction in cognitive load, this did not translate into a statistically significant improvement in performance on the achievement test. It is possible that the lack of independent use of the chatbot in this study may have contributed to the observed outcomes, as it prevented the identification and addressing of individual learning difficulties. Furthermore, the brief duration of the intervention may have been a contributing factor to the observed outcome. For instance, the heightened self-efficacy expectation may only manifest, resulting in positive effects on learning strategies, ambition, and the willingness to tackle more complex problems in longer-term studies (see Sec. II A). Consequently, an impact on learning performance would only become evident at that juncture.

VII. CONCLUSION AND OUTLOOK

The findings of the present study provide valuable insights into the effects of learning with explanations generated by an AI custom chatbot on students' learning experience and learning performance. It was shown that learning with these explanations as Supplemental Material has a significant impact on learners' positive-activating emotions, situational interest, and self-efficacy, while also reducing cognitive load in comparison to traditional learning with textbook materials only. Nevertheless, the influence on learning outcomes remains unclear. No notable differences in the overall result of the performance test could be identified and no definitive conclusions can be drawn.

The generalizability of the results is constrained by different factors: the size of the sample and the brief implementation period (snapshot) as well as the necessity of selecting a particular topic for a specific target group. Additionally, the learning materials were provided in a centralized manner, which resulted in the lack of consideration for individual differences in utilization and the potential for interaction with the chatbot.

Further research could serve to confirm the statements and assumptions posited here through the implementation of larger and longer-term studies with a variety of topics and a wider target group. Moreover, subsequent to the present study, learners should be afforded the opportunity to utilize the AI chatbot independently and individually in order to measure the impact of individualized feedback.

The findings of the study demonstrate that AI custom chatbots have the capacity to generate learning materials or explanations on topics that significantly enhance the learning experience of students. A key benefit of utilizing a chatbot for this purpose is its capacity for immediate, customized responses, which can be generated with a high degree of flexibility and adaptability. Teachers are empowered to create explanations and materials for diverse purposes, including different subjects, class levels, and integrated classrooms. Additionally, students have the option of utilizing AI custom chatbots to obtain feedback on topics and tasks they do not comprehend. To gain a more nuanced understanding of the usability of custom chatbots in mathematics and physics, it would be prudent to implement a design with a more pronounced emphasis on learning outcomes.

Overall, the present work shows that the use of AI in the learning process offers promising possibilities, although further research is needed to fully realize its potential and increase effectiveness in different learning scenarios.

ACKNOWLEDGMENTS

We would like to express our sincere thanks to all participating schools, teachers, and students for contributing to our study.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

[1] S. Küchemann, S. Steinert, J. Kuhn, K. Avila, and S. Ruzika, Large language models—Valuable tools that require a sensitive integration into teaching and learning physics, Phys. Teach. **62**, 400 (2024).

[2] T. Adiguzel, M. H. Kaya, and F. K. Cansu, Revolutionizing education with AI: Exploring the transformative potential of ChatGPT, Contemp. Educ. Technol. **15**, ep429 (2023).

[3] E. Kasneci *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Diff. **103**, 102274 (2023).

[4] S. Küchemann *et al.*, Are large multimodal foundation models all we need? On opportunities and challenges of these models in education (2024), 10.35542/osf.io/n7dvf.

[5] K. Neumann, J. Kuhn, and H. Drachsler, Generative Künstliche Intelligenz in Unterricht und Unterrichtsforschung—Chancen und Herausforderungen [Generative artificial intelligence in teaching and education research—opportunities and challenges], Unterrichtswissenschaft **52**, 227 (2024).

[6] K. E. Avila, S. Steinert, S. Ruzika, J. Kuhn, and S. Küchemann, Using ChatGPT for teaching physics, Phys. Teach. **62**, 536 (2024).

[7] Y. Liang, Di Zou, H. Xie, and F. L. Wang, Exploring the potential of using ChatGPT in physics education, Smart Learn. Environ. **10**, 52 (2023).

[8] G. Orrù, A. Piarulli, C. Conversano, and A. Gemignani, Human-like problem-solving abilities in large language models using ChatGPT, Front. Artif. Intell. **6**, 1199350 (2023).

[9] K. T. Kotsis, Correcting students' misconceptions in physics using experiments designed by ChatGPT, Eur. J. Contemp. Educ. E-Learn. **2**, 83 (2024).

[10] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, A SWOT analysis of ChatGPT: Implications for educational practice and research, Innov. Educ. Teach. Int. **61**, 460 (2024).

[11] L. Krupp *et al.*, Unreflected acceptance—investigating the negative consequences of ChatGPT-assisted problem solving in physics education, in *HHAI 2024: Hybrid Human AI Systems for the Social Good* (2024), pp. 199–212, 10.3233/FAIA240195.

[12] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT, Phys. Rev. Phys. Educ. Res. **19**, 010142 (2023).

[13] W. M. Christensen and J. R. Thompson, Investigating graphical representations of slope and derivative without a physics context, Phys. Rev. ST Phys. Educ. Res. **8**, 023101 (2012).

[14] G. Leinhardt, M. K. Stein, and O. Zaslavsky, Functions, graphs, and graphing: Tasks, learning, and teaching, Rev. Educ. Res. **60**, 1 (1990).

[15] E. B. Pollock, J. R. Thompson, and D. B. Mountcastle, Student understanding of the physics and mathematics of process variables in P-V diagrams, AIP Conf. Proc. **951**, 168 (2007).

[16] S. Becker, L. Knippertz, S. Ruzika, and J. Kuhn, Persistence, context, and visual strategy of graph understanding: Gaze patterns reveal student difficulties in interpreting graphs, Phys. Rev. Phys. Educ. Res. **19**, 020142 (2023).

[17] E. F. Redish and E. Kuo, Language of physics, language of math: Disciplinary culture and dynamic epistemology, Sci. Educ. **24**, 561 (2015).

[18] O. Uhden, Verständnisprobleme von Schülerinnen und Schülern beim Verbinden von Physik und Mathematik [Students' comprehension problems when combining physics and mathematics], Z. Didakt. Naturwiss. **22**, 13 (2016).

[19] D. F. Treagust, The importance of multiple representations for teaching and learning science, in *Education Research Highlights in Mathematics, Science and Technology* (ISRES, 2018), pp. 215–223.

[20] J. Hansen, L. E. Richland, and I. Davidesco, Teaching and learning science through multiple representations: Intuitions and executive functions, CBE Life Sci. Educ. **19**, ar61 (2020).

[21] M. Opfermann, A. Schmeck, and H. E. Fischer, *Multiple Representations in Physics Education*, edited by D. F. Treagust, R. Duit, and H. E. Fischer (Springer, Cham, 2017), p. 1, 10.1007/978-3-319-58914-5_1.

[22] T. J. Bing and E. F. Redish, Analyzing problem solving using math in physics: Epistemological framing via warrants, Phys. Rev. ST Phys. Educ. Res. **5**, 020108 (2009).

[23] T. Nilsen, C. Angell, and L. S. Grønmo, Mathematical competencies and the role of mathematics in physics education: A trend analysis of TIMSS Advanced 1995 and 2008, Acta Didact. Nor. **7** (2013).

[24] *Emotion in Education*, edited by P. A. Schutz and R. Pekrun (Elsevier Academic Press, New York, NY, 2007).

[25] A. S. Willems, *Bedingungen des situationalen Interesses im Mathematikunterricht. Eine mehrebenenanalytische Perspektive [Conditions of situational interest in mathematics teaching. A multi-level analytical perspective]* (Waxmann, Münster, 2011).

[26] K. A. Renninger, S. Hidi, A. Krapp, and A. Renninger, *The Role of Interest in Learning and Development* (Psychology Press, New York, 2014).

[27] J. I. Rotgans and H. G. Schmidt, Interest development: Arousing situational interest affects the growth trajectory of individual interest, Contemp. Educ. Psychol. **49**, 175 (2017).
[28] J. Leppink, F. Paas, C. P. M. Van der Vleuten, T. Van Gog, and J. J. G. Van Merriënboer, Development of an instrument for measuring different types of cognitive load, Behav. Res. Meth. Instrum. Comput. **45**, 1058 (2013).
[29] S. Becker, P. Klein, A. Gößling, and J. Kuhn, Investigating dynamic visualizations of multiple representations using mobile video analysis in physics lessons, Z. Didakt. Naturwiss. **26**, 123 (2020).
[30] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988).
[31] J. J. G. van Merrienboer and J. Sweller, Cognitive load theory and complex learning: Recent developments and future directions, Educ. Psychol. Rev. **17**, 147 (2005).
[32] J. Leppink and A. van den Heuvel, The evolution of cognitive load theory and its application to medical education, Perspect. Med. Educ. **4**, 119 (2015).
[33] J. Leppink, Cognitive load theory: Practical implications and an important challenge, J. Taibah Univ. Med. Sci. **12**, 385 (2017).
[34] J. Sweller, J. J. G. van Merriënboer, and F. Paas, Cognitive architecture and instructional design: 20 years later, Educ. Psychol. Rev. **31**, 261 (2019).
[35] A. Bandura, *Self-Efficacy: The Exercise of Control* (W.H. Freeman, New York, 1997).
[36] *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen [Self-efficacy and motivation processes in educational institutions]*, edited by M. Jerusalem and D. Hopf, Zeitschrift für Pädagogik, Beiheft 44 (Beltz, Weinheim, 2010).
[37] C. Kulgemeyer, Erklären im Physikunterricht [Explanations in physics teaching], in *Physikdidaktik Grundlagen*, edited by E. Kircher, R. Girwidz, and H. E. Fischer (Springer Spektrum, Berlin, Heidelberg, 2020), pp. 403–426, 10.1007/978-3-662-59490-2_11.
[38] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, Cognitive tutors: Lessons learned, J. Learn. Sci. **4**, 167 (1995).
[39] D. Meyer and V. Pietzner, Reading textual and non-textual explanations in chemistry texts and textbooks—a review, Chem. Educ. Res. Pract. **23**, 768 (2022).
[40] J. Wittwer and A. Renkl, Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations, Educ. Psychol. **43**, 49 (2008).
[41] L. Chen, P. Chen, and Z. Lin, Artificial Intelligence in education: A review, IEEE Access **8**, 75264 (2020).
[42] F. Mahligawati, E. Allanas, M. H. Butarbutar, and N. A. N. Nordin, Artificial intelligence in Physics Education: A comprehensive literature review, J. Phys. Conf. Ser. **2596**, 012080 (2023).
[43] F. Fauzi *et al.*, Analysing the role of ChatGPT in improving student productivity in higher education, J. Educ. **5**, 14886 (2023).
[44] R. Wu and Z. Yu, Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis, Br. J. Educ. Technol. **55**, 10 (2024).
[45] C. K. Lo, K. F. Hew, and M. S. Jong, The influence of ChatGPT on student engagement: A systematic review and future research agenda, Comput. Educ. **219**, 105100 (2024).
[46] F. S. Al-Hafdi and S. M. AlNajdi, The effectiveness of using chatbot-based environment on learning process, students' performances and perceptions: A mixed exploratory study, Educ. Inf. Technol. **29**, 20633 (2024).
[47] A. M. Canonigo, Levering AI to enhance students' conceptual understanding and confidence in mathematics, J. Comput. Assist. Learn. **40**, 3215 (2024).
[48] K. Alarbi, M. Halaweh, H. Tairab, N. R. Alsalhi, N. Annamalai, and F. Aldarmaki, Making a revolution in physics learning in high schools with ChatGPT: A case study in UAE, Eurasia J. Math. Sci. Technol. Educ. **20**, em2499 (2024).
[49] S. Alneyadi and Y. Wardat, ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools, Contemp. Educ. Technol. **15**, 448 (2023).
[50] M. Stadler, M. Bannert, and M. Sailer, Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry, Comput. Human Behav. **160**, 108386 (2024).
[51] A. Marinosyan, LLMs in physics and mathematics education and problem solving: Assessment of ChatGPT-4 level and suggestions for improvement (2024), 10.13140/RG.2.2.20196.18568.
[52] L. Ding, T. Li, S. Jiang, and A. Gapud, Students' perceptions of using ChatGPT in a physics class as a virtual tutor, Int. J. Educ. Technol. High. Educ. **20**, 63 (2023).
[53] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, Educational data augmentation in physics education research using ChatGPT, Phys. Rev. Phys. Educ. Res. **19**, 020150 (2023).
[54] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).
[55] C. West, Advances in apparent conceptual physics reasoning in ChatGPT-4, arXiv:2303.17012.
[56] G. Polverini and B. Gregorcic, Performance of ChatGPT on the test of understanding graphs in kinematics, Phys. Rev. Phys. Educ. Res. **20**, 010109 (2024).
[57] G. Polverini and B. Gregorcic, Evaluating vision-capable chatbots in interpreting kinematics graphs: A comparative study of free and subscription-based models, Front. Educ. **9** (2024).
[58] V. G. A. Hakim, N. A. Paiman, and M. H. S. Rahman, 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., **32**, e22800 (2024).
[59] G. Kortemeyer, Tailoring chatbots for higher education: Some insights and experiences, arXiv:2409.06717.
[60] P. Lewis *et al.*, Retrieval-augmented generation for knowledge-intensive NLP tasks, Adv. Neural Inf. Process. Syst. **33**, 9459 (2020).
[61] A. Lergenmüller and A. Baeger, *Mathematik neue Wege [Mathematics New approaches] (Dr. A 5)* (Schroedel, Hannover, 2012).

[62] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.21.010147 for learning material, performance test, scales/items, and results of data analysis.

[63] K. Rakoczy, A. Buff, and F. Lipowsky, *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie. "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". 1. Befragungsinstrumente [Documentation of the survey and evaluation instruments for the Swiss-German video study. "Teaching quality, learning behavior and mathematical understanding". 1. survey instruments]* (2005), https://www.pedocs.de/frontdoor.php?source_opus=3106 [accessed November 13, 2024].

[64] R. Pekrun *et al.*, Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. Educ. Psychol. **37**, 91 (2002).

[65] R. Pekrun *et al.*, Measuring emotions in students' learning performance: The achievement emotions questionnaire (AEQ), Contemp. Educ. Psychol. **36**, 36 (2011).

[66] L. Linnenbrink-Garcia, A. M. Durik, A. M. Conley, K. E. Barron, J. M. Tauer, S. A. Karabenick, and J. M. Harackiewicz, Measuring situational interest in academic domains, Educ. Psychol. Meas. **70**, 647 (2010).

[67] *Skalen zur Erfassung von Lehrer- und Schülermerkmalen: Dokumentation der psychometrischen Verfahren im Rahmen der wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen [Scales for recording teacher and student characteristics: Documentation of psychometric procedures as part of the scientific monitoring of the pilot project Self-Effective Schools]*, edited by R. Schwarzer and M. Jerusalem (1999), ISBN 978-3-00-003708-5.

[68] M. Jerusalem and L. Satow, Schulbezogene Selbstwirksamkeitserwartung [School-related self-efficacy expectations], in *Skalen zur Erfassung von Lehrer- und Schülermerkmalen [Scales for Recording Teacher and Student Characteristics]* (1999), p. 15.

[69] R Core Team, R: A Language and Environment for Statistical Computing (2024), https://www.R-project.org/.

[70] W. Revelle, psych: Procedures for Psychological, Psychometric, and Personality Research (2024), https://CRAN.R-project.org/package=psych.

[71] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[72] M. Blanz, *Forschungsmethoden und Statistik für die Soziale Arbeit: Grundlagen und Anwendungen [Research Methods and Statistics for Social Work: Basics and Applications]* (2021), 10.17433/978-3-17-039819-1.

[73] M. Gamer, J. Lemon, I. Fellows, and P. Singh, irr: Various Coefficients of Interrater Reliability and Agreement (2019), https://CRAN.R-project.org/package=irr.

[74] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. **20**, 37 (1960).

[75] J. Cohen, Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, Psychol. Bull. **70**, 213 (1968).

[76] N. Döring, *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften [Research methods, and evaluation in the social, and human sciences]*, 6th ed. (Springer, Berlin, Heidelberg, 2023), ISBN 978-3-662-64762-2.