

Comparing AI chatbot, tiered, and textual support in physics problem solving: Effects on cognitive load and affective-motivational outcomes

Eleonore Becker^{1,*}, Johannes Wünsche¹, Joaquin Veith¹,
Johanna Schrader², and Philipp Bitzenbauer¹

¹*Institute of Physics Education, Leipzig University, Vor dem Hospitaltore 1, 04103, Leipzig, Germany*

²*Leibniz University Hannover, L3S Research Center, Hannover, Germany
and CAIMed–Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine,
Hannover, Germany*



(Received 20 March 2026; accepted 19 August 2026; published 1 October 2026)

Providing appropriate support is important for helping students manage conceptual difficulties in physics. Established scaffolding methods, such as structured tiered support, can be effective but are often resource-intensive to develop and implement broadly. AI chatbots may offer a more easily scalable alternative, but their potential as learning support tools needs to be examined carefully and in direct comparison with established forms of support. As a first step, this study compares a custom-configured AI chatbot, a tiered support system, and traditional textbook-style textual support with respect to students' perceived cognitive load and affective-motivational outcomes during physics problem solving. In a cluster randomized field study, 268 ninth-grade students in Germany worked individually on a buoyancy problem using one of the three support formats. We measured intrinsic and extraneous cognitive load and affective-motivational outcomes (enjoyment, hope, hopelessness, self-efficacy, and situational interest) via research-validated questionnaires. Results showed that the AI chatbot and the tiered support system were associated with significantly lower perceived intrinsic and extraneous cognitive load than textual support. Post-test comparisons showed that students in the chatbot condition reported more favorable affective-motivational outcomes than students in the textual-support condition, while differences between the chatbot and tiered-support conditions were not statistically significant. Overall, the findings suggest that interactive support formats may shape students' short-term perceived cognitive and affective-motivational experiences during physics problem solving; future research should examine how these experiences relate to solution quality, conceptual learning, and longer-term outcomes.

DOI: [10.1103/nvf1-zrq8](https://doi.org/10.1103/nvf1-zrq8)

I. INTRODUCTION

Physics is widely recognized as a conceptually challenging subject. A primary difficulty for students is translating theoretical principles into concrete conceptual models, a skill that typically requires deliberate pedagogical support [1]. Without such guidance, many novices resort to *formula-based problem solving* rather than developing a deep understanding of the underlying physics [1,2]. To address this challenge, Differentiated Instruction (DI) is a well-established approach designed to tailor teaching to the diverse needs of all learners [3,4]. A pedagogical strategy for achieving DI is employing tiered

activities [5], a form of instructional scaffolding, which provides temporary, adaptive support to help a student succeed in a task that would otherwise be out of reach [2,6]. In physics education, “tiered support” (in German: *gestufte Lernhilfen*) exemplifies this by providing students with a sequence of optional, increasingly explicit learning aids, allowing them to individually adapt the level of assistance they receive [7]. The effectiveness of such scaffolding approaches is well documented, positively impacting not only academic achievement but also student motivation [8–10]. However, widespread application poses a significant practical challenge. While teachers recognize student heterogeneity, they are often constrained by high workloads and limited time [11]. This makes the extensive effort required to develop high-quality tiered materials for every task a prohibitive undertaking. Generative AI chatbots, especially when pedagogically tailored through specific configurations, may offer one possible response to this pedagogical dilemma. They have the potential to act as personalized tutors, potentially reducing teacher workload while offering scalable, on-demand support to

*Contact author: eleonore.becker@uni-leipzig.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

students engaged in parallel, individualized learning cf. [12,13]. Current research suggests that AI chatbots can support cognitive and affective-motivational outcomes in science, technology, engineering, and mathematics education, for example, by acting as intelligent tutors or providing personalized scaffolding [14–16], although findings depend on the specific system, task, and outcome measures considered. Other studies have explored the application of generative AI in open-ended tasks such as programming, creating teaching aids, or design thinking [15,17–19]. However, it remains an open question how affective outcomes, specifically achievement emotions [20], self-efficacy [21], and interest [22], are affected alongside cognitive load [23] when using an interactive AI chatbot for physics problem solving, particularly in direct comparison to established pedagogical support systems. Specifically, the distinct effects of dynamic, interactive AI assistance versus more static aids, such as structured tiered hints or traditional textbook-style explanations, on students' affective and cognitive experiences are not yet well understood.

This study directly addresses this gap by systematically comparing the effects of AI chatbot assistance with those of tiered support and traditional textbook-style explanations. To contextualize this comparison, the following section establishes a theoretical framework by examining different forms of instructional support.

II. RESEARCH BACKGROUND

A. Support in physics education

A central concept in physics education is instructional scaffolding: temporary, interactive support that helps a student build on prior knowledge to complete tasks that would otherwise be out of reach [2,6]. Crucially, effective scaffolding is not merely the provision of information but a dynamic process. Support is offered contingently—in direct response to a student's progress—and is gradually faded as the learner's competence grows [6,24].

This responsive nature distinguishes scaffolding from passive instructional aids. A textbook, for example, is a static resource that cannot adapt to a learner's immediate needs and therefore does not qualify as scaffolding [2]. Thus, instructional support can be viewed on a spectrum of interactivity and adaptivity. To evaluate the potential of new, highly interactive systems, it is necessary to compare them against established support methods across this spectrum, beginning with the most traditional form of assistance.

1. Textual support

Traditional textual support, most commonly embodied in textbooks, remains an important instructional tool in science classes [25,26]. In many science courses, such written materials are a fundamental requirement, and

instructors expect students to engage with them to enhance academic performance [27]. They are designed to be accessible resources that students can utilize according to their individual learning styles [25]. However, a significant discrepancy exists between the presumed importance of textual resources and their actual use by students. Student-reported compliance with assigned readings is varied, ranging from approximately 20%–70% [28]. This underutilization can be partly attributed to a perceived gap between the complexity of homework tasks and the relatively low complexity of explanations found in textbooks [27].

Learning from physics texts is an intrinsically challenging cognitive task [29]. The difficulty is compounded by the integration of two distinct language systems: natural language and mathematical-symbolic notation [29]. Furthermore, pedagogical features intended to aid comprehension, such as analogies, can sometimes increase cognitive processing demands, thereby hindering rather than helping learning [29]. From a constructivist perspective, a textual resource should be viewed not as a container of knowledge, but as a tool for students to construct meaning [30]. Yet, the perceived authority of texts can constrain this process, as students may be less inclined to elaborate on their own thinking [31].

Empirical work by Boxtel *et al.* [31] supports this, finding that student pairs with access to a textbook engaged in less collaborative elaboration compared to those without. Students often consulted the text without a specific goal, used inefficient search strategies, and failed to integrate the information with their own conceptions. Interestingly, despite the reduction in elaborative interaction, the study found no significant difference in post-test scores, suggesting that individual acquisition of factual knowledge from the text may have compensated for the lack of deeper, collaborative meaning-making. This positions textual support as a relevant, albeit potentially passive, baseline for comparison with more interactive learning support.

2. Tiered support

A more dynamic form of instructional assistance can be offered via scaffolding, where temporary support is provided to help a learner accomplish a task that would otherwise be out of reach [24]. One specific implementation of this principle is “tiered support” (in German: “gestufte Lernhilfen,” cf. [32]). In this approach, a complex problem is accompanied by a sequence of written hints that learners can access according to their perceived needs, allowing the support to be adaptive [7]. These hints are designed to prompt general learning strategies, such as paraphrasing the task, focusing on relevant information, activating prior knowledge, or visualizing the problem [8]. Typically, each hint consists of a question or prompt, with the corresponding answer made accessible in a second step [7]. The answer to the last hint is a complete sample

Learning Support

Hint 1
Explain the task in your own words. What is the main goal, and what parts are still unclear to you?

Solution to Hint 1

Hint 2
Why does an ice floe float in the first place?

An ice floe floats because ice is less dense than the water it's in.

Hint 3
Try to describe the exact point when a loaded ice floe is about to sink.

When an ice floe is floating normally, part of it is above the water. You can add weight to it until its top surface is perfectly level with the water's surface. At that point, it's holding the maximum possible load.

Hint 4
Imagine a tiny 1 cm³ cube of ice. How much extra weight can it support before it's completely submerged?

Solution to Hint 4

Show Hint 5

I'm done

FIG. 1. The digital interface for the tiered support system used in this study. Students could sequentially get a series of pre-written hints and optionally reveal their answer.

solution that allows students to verify their answer [8]. The tiered-support materials used in the present study are based on a physics problem-solving task on Archimedes' principle developed in the context of a DFG (German Research Foundation) project by Franke-Braun [8]. To illustrate this structure, Fig. 1 shows the digital interface used in this study, and the full text of the hints is provided in the supplemental material [33].

The principles of tiered support are closely aligned with the extensive research on worked examples, which provide a problem statement, solution steps, and a final answer [8]. The efficacy of both approaches can be explained by cognitive load theory (CLT, cf. [34]). CLT posits that working memory has a limited capacity. When novice learners attempt to solve a problem without guidance, they often engage in inefficient search strategies (e.g., means-ends analysis) that impose a high extraneous cognitive load—effort that is not productive for learning [34,35]. By providing a clear solution path, worked examples and tiered support eliminate this counterproductive search, thereby reducing extraneous load and freeing up cognitive resources [34].

Furthermore, structured methods such as tiered support help manage a material's intrinsic cognitive load. Intrinsic cognitive load refers to the cognitive demands imposed by the complexity of the learning task, which can be influenced by factors such as the learner's prior knowledge [36]. Although instructional design cannot remove the

conceptual complexity of the content, it can manage its impact on novices [37]. For example, segmenting a problem into smaller steps or presenting interacting elements in isolation before combining them facilitates serial processing over simultaneous processing. This strategy significantly reduces the demand on working memory [37,38].

This process is further optimized in what is known as faded worked examples, where parts of the solution are gradually omitted, requiring the learner to fill in the gaps [34]. This method creates a smooth transition from fully guided example study to independent problem solving [39]. Tiered support can be conceptualized as a learner-controlled fading procedure; by choosing which hints to access, students dynamically adjust the level of scaffolding, effectively creating their own incomplete example. Research has shown that such fading procedures lead to better near-transfer performance and fewer errors during learning when compared to traditional example-problem pairs [34].

Empirical studies on tiered support systems have demonstrated significant motivational benefits. Students working with tiered hints reported higher feelings of competence, higher intrinsic motivation, and a tendency toward greater social inclusion during partner work compared to groups receiving nonsequential help [8–10]. Furthermore, an incremental presentation of a worked example leads to higher feelings of competence and a reduced perceived cognitive load compared to presenting the solution as a whole [40]. However, while incremental presentation alone improves the affective experience, the addition of strategic prompts appears necessary to significantly improve the correctness of the final solution [40]. This suggests that while structured, sequential support is beneficial, the quality and function of the guidance play a key role in learning outcomes.

3. Chatbot support

The third form of support, AI chatbots, offer an increase in interactivity and adaptivity. Unlike tiered support, which follows a predefined, sequential path, AI chatbots provide dynamic and interactive scaffolding [41]. Modern LLM-based chatbots differ from earlier rule-based chatbots because they can generate flexible, context-sensitive dialogue and can be configured through prompt or role engineering to support specific pedagogical interaction styles [42]. Their primary pedagogical potential lies in their capacity for individualized support, which allows them to address specific student queries and guide a personalized learning process [12,43].

Empirical findings on GenAI-supported learning are promising but heterogeneous. Meta-analytic evidence suggests that AI chatbots and GenAI-supported interventions can have positive effects on learning-related outcomes, but these effects are highly heterogeneous and

depend on educational level, subject domain, intervention duration, outcome type, and instructional design [44,45]. With respect to motivational outcomes, a scoping review by Huang *et al.* [41] suggests that chatbot-supported learning often shows advantages over conditions without chatbot support, while effects tend to be weaker than those of interventions involving human feedback or carefully designed pedagogical approaches. A similar pattern is emerging in physics education, where LLM-based tools have been associated with improved learning performance, motivation, and student engagement [14,16,46–49]. For example, Lademann *et al.* [50] found that AI-generated explanations produced by a custom chatbot in a mathematical and physical learning context enhanced positive-activating emotions, situational interest, and self-efficacy, and reduced intrinsic and extraneous cognitive load, while effects on learning performance remained unclear. Similarly, Kestin *et al.* [46] found that a carefully structured AI tutor in an undergraduate physics course was associated with higher post-test performance and greater reported engagement and motivation than an active-learning class, while Chen and Chang [14] reported that integrating ChatGPT into a game-based physics learning environment improved students' performance and perceived competence and reduced intrinsic cognitive load compared with game-based learning alone.

However, this promising body of research is subject to important caveats. A central concern is that chatbot-supported problem solving may blur the distinction between producing a correct or polished response and developing durable understanding, particularly when quick answers or coherent explanations are treated as evidence of learning [51]. For example, Bastani *et al.* [52] showed in a field experiment with high school mathematics students that access to a standard GPT-4 tutor improved performance during assisted practice, yet students performed worse than controls on a subsequent unassisted exam when access to the tool was removed; these negative effects were largely mitigated when the tutor was constrained to provide teacher-designed hints rather than direct solutions. Such findings point to a broader risk that GenAI use may encourage students to off-load cognitive and metacognitive work rather than engage in the reasoning processes needed for learning [51,53,54]. For example, Stadler *et al.* [55] found that using an LLM for a research task significantly reduced students' cognitive load but also led to a decline in the quality of their reasoning and argumentation. In physics education, Krupp *et al.* [53] similarly showed that unguided ChatGPT-assisted problem solving can lead students toward copy-and-paste strategies and uncritical acceptance of incorrect AI-generated solutions. Against this background, cognitive load and affective-motivational responses to chatbot support should be interpreted as indicators of students' experience during problem solving, rather than as direct evidence of learning effectiveness.

III. RESEARCH RATIONALE

A central challenge in physics education is effectively addressing student heterogeneity through differentiated support [4]. This study compares three instructional tools that offer varying degrees of individualization. Textbook-style support, serving as a baseline, is a static and standardized resource that cannot adapt to individual learner needs, potentially imposing a high cognitive load [2,29]. An improvement is tiered support, which offers learner-directed differentiation but is too resource-intensive for educators to create for every task [7,11]. In contrast, AI chatbots represent another potentially scalable form for dynamic individualization [12]. Building on prior work from the DFG project by Wodzinski and colleagues on structured support for physics problem solving, the present study extends this design by adding an AI-chatbot condition and comparing it with tiered and textual support in a common physics problem-solving context.

However, there is still limited comparative evidence on how these support modalities affect students' self-reported cognitive load and affective-motivational experiences during physics problem solving. Existing research on AI in physics education has primarily focused on learning performance [46,48,49], or has been situated in gamified learning environments [14] or at the university level [56]. Other studies have explored general AI tutors [16] or examined user experience with task-specific tutors [57].

To our knowledge, no prior study has directly compared a task-specific AI chatbot explicitly designed to assist with physics problem solving for middle school-aged students in a regular classroom context against both tiered support and traditional textual support. While previous work often emphasizes final performance outcomes, it frequently overlooks the underlying cognitive processes (intrinsic and extraneous cognitive loads) and affective experiences that shape student learning and engagement during problem solving.

These dimensions are central for evaluating learning support systems because different forms of guidance may shape how students experience the problem-solving process. From a cognitive-load perspective, support systems can differ in the extent to which they help learners manage the conceptual demands of a task, for example, by segmenting complex information, guiding attention to relevant elements, or prompting students to articulate their reasoning [36,37,58,59]. From an affective-motivational perspective, support may also shape students' task-related situational interest, self-efficacy, and achievement emotions, which, for instance, can be influenced by mastery experiences and verbal persuasion [21,60].

This study, therefore, provides a systematic classroom-based comparison of three support types—static textbook-style explanations, tiered support, and an AI chatbot—with respect to students' perceived cognitive load and concurrent affective-motivational states during physics problem

TABLE I. Overview of experimental conditions.

Group	Description of support condition	N
G1: Chatbot	Participants interacted with a custom-configured AI chatbot for task-related guidance and feedback. The chatbot was designed to promote independent problem solving and conceptual understanding.	103
G2: Tiered support	Participants were provided with a series of structured, step-by-step hints on a website, which could be revealed sequentially (adapted from Franke-Braun [8]).	64
G3: Textual support (control)	Participants received a static digital textbook-style page, which contained all the information required to solve the problem.	101

solving. By doing so, we seek to generate practical, evidence-based insights into how different support tools shape students' short-term problem-solving experiences. Thus, this investigation is guided by the following research questions:

- RQ1: To what extent do AI chatbot support, tiered support, and traditional textual support differ in their effects on students' perceived cognitive load during physics problem solving?
- RQ2: To what extent do AI chatbot support, tiered support, and traditional textual support differ in their effects on students' situational interest, self-efficacy, and achievement emotions (hope, hopelessness, and enjoyment) during physics problem solving?

IV. MATERIALS AND METHODS

A. Study design and sample

We conducted a cluster randomized¹ field study using a pre-post-test design to compare the effects of different learning support systems on student motivation and cognitive load during physics problem solving. The final sample consisted of 268 ninth-grade students who had completed all questionnaires and explicitly confirmed their use of the support system for problem solving during the study. The study was conducted within the students' classrooms and followed a standardized protocol to ensure consistency across groups. Students were instructed to work silently and individually.

During the intervention phase, each student individually solved a physics task on buoyancy, taken from Franke-Braun [8]. The task was presented on a paper worksheet, and students were allotted 20 min for completion and an additional 5 min to compare their work with a sample solution. Each worksheet also contained a brief guide for the accompanying support tool, which offered practical instructions for its use and guidance on applying metacognitive strategies. Students recorded their answers and reasoning directly on the worksheet.

All participants performed the same core task, but the learning support they received varied across three experimental conditions illustrated in Table I.

To assess the effects of the intervention, data were collected via questionnaires administered directly before (pretest) and after (post-test) the intervention. Individual student responses were linked across the two time points using an anonymous identifier. The dependent variables were measures of interest, self-efficacy, cognitive load, and achievement emotions (hope, hopelessness, and enjoyment). To account for potential confounding by prior physics knowledge, we assessed participants' prior physics knowledge. A detailed overview of all measured variables, including their abbreviations, measurement times, and scale properties, can be found in Table II. The study design is visualized in Fig. 2.

B. Variable selection

To investigate our research questions, we developed a questionnaire to assess key affective and cognitive variables related to the intervention. Affective and motivational variables were assessed before and after the intervention, whereas cognitive load was assessed after the intervention. The selection of these variables was guided by established theoretical frameworks from educational psychology and physics education research.

To account for the effects of the different support conditions, we controlled for prior conceptual understanding of density, because it shapes how students experience task difficulty and correlates with motivation, achievement emotions, and cognitive load [36,61–63].

Our first research question addressed the cognitive demands of the task, which we assessed through the lens of cognitive load theory (CLT) [23,64]. Intrinsic cognitive load was included to capture how demanding students perceived the conceptual content and structure of the buoyancy problem to be. Extraneous cognitive load was included to capture cognitive demands that students perceived as arising from the way the problem and the accompanying support were presented.

For the second research question, we assessed students' task-related affective and motivational experiences. Situational interest was included because it captures

¹Cluster randomization was performed at the class level.

TABLE II. Overview of all assessed variables in terms of construct abbreviation, scale length, measurement times, and internal consistency expressed via Cronbach's α . If scales have been administered at both pre- and post-test points in time, two values for Cronbach's α are reported.

Variable (Abbr., No. of Items)	Pretest	Posttest	Cronbach's α	References
Prior knowledge (PK, 9)	✓	✗	0.63	Zenger and Bitzenbauer [66]
Self-efficacy (SE, 7)	✓	✓	0.88 & 0.90	Duncan <i>et al.</i> [67]
Enjoyment (EE, 4)	✓	✓	0.79 & 0.88	Bieleke <i>et al.</i> [68]
Hope (EH, 4)	✓	✓	0.88 & 0.89	Bieleke <i>et al.</i> [68]
Hopelessness (EHL, 4)	✓	✓	0.84 & 0.82	Bieleke <i>et al.</i> [68]
Intrinsic CL (ICL, 2)	✗	✓	0.58	Klepsch <i>et al.</i> [69]
Extraneous CL exercise (ECLA, 3)	✗	✓	0.61	Klepsch <i>et al.</i> [69]
Extraneous CL support (ECLL, 3)	✗	✓	0.81	Klepsch <i>et al.</i> [69]
Situational interest (SI, 6)	✗	✓	0.85	Woithe [70]

students' immediate interest in the specific activity and influences attention and the depth of learning [22]. Self-efficacy was included because students' beliefs about their ability to solve the task are relevant for persistence and engagement in challenging physics problems [65]. Enjoyment, hope, and hopelessness were included as achievement emotions that capture positive and negative affective experiences during the activity, as these are highly predictive of academic outcomes and are linked to the cognitive resources students deploy during a task [20].

Together, these variables allowed us to compare the three support conditions with respect to students' short-term perceived cognitive demands and task-related affective-motivational experiences during physics problem solving.

C. Instrument development

The questionnaire used in this study was developed using established and research-validated scales to measure the variables of interest. Scales, originally developed in English, were translated into German, with meticulous attention to preserving their conceptual meaning. To ensure relevance to the specific physics context, namely buoyancy, the wording of some items and the introductory text for each scale were systematically adapted. These adapted items were then reviewed by two physics education experts to ensure their meaning was preserved. For consistency across measures and to simplify the response process for students, all variables except prior physics knowledge were assessed using a uniform five-point rating scale, ranging from 1 ("Does not apply at all") to 5 ("Fully applies"). A comprehensive overview of all constructs,

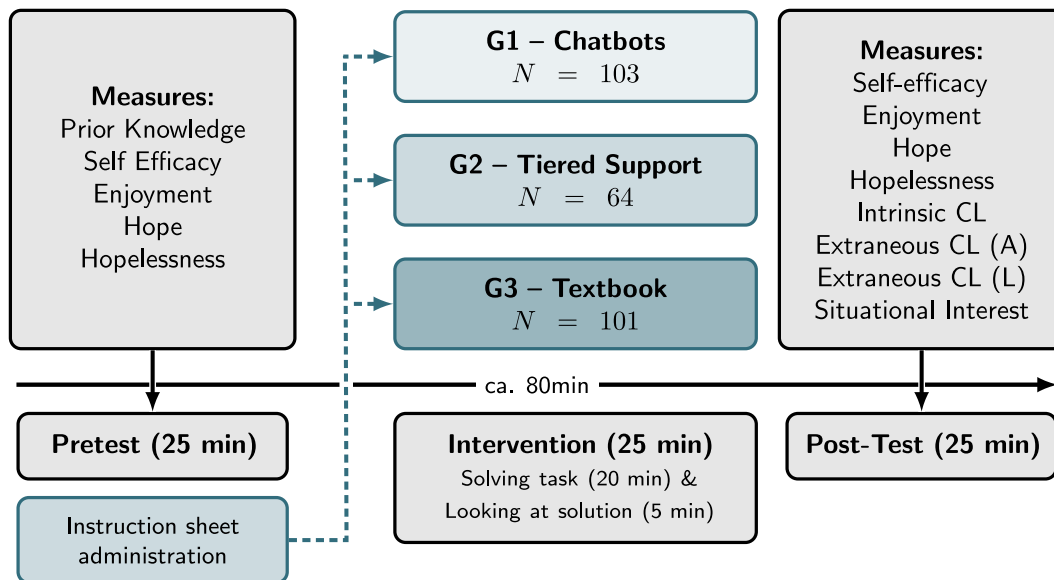


FIG. 2. Overview of the study design, including the three different groups and the corresponding sample sizes as well as the assessed variables and the dedicated time slots.

their corresponding scales, source references, and estimates of internal consistency (Cronbach's α) is provided in Table II. While most scales demonstrated good internal consistency ($\alpha > 0.70$), the brief scales for intrinsic cognitive load ($\alpha = 0.58$) and extraneous cognitive load of the exercise ($\alpha = 0.61$) exhibited lower internal consistency, alongside the prior knowledge test ($\alpha = 0.63$).

D. Materials

The final materials for all three conditions—the textual support page, the tiered support task, and the chatbot's system prompt—are provided in Supplemental Material [33] alongside the buoyancy task. Here, we provide an overview of the selection of the physics problem and development support materials.

1. Physics problem selection

The physics task asked students to determine how large an ice floe would have to be in order to carry one person without sinking, drawing on the concepts of density and buoyancy in the context of Archimedes' principle [8]. Prior research, primarily with university students, has documented persistent difficulties in reasoning about buoyancy and Archimedes' principle. These include conflating mass, volume, and density, assuming that the buoyant force depends on an object's mass or depth, and failing to distinguish the volume of an object from the volume of fluid it displaces [71,72]. In a related task involving a piece of ice floating in water, students sometimes failed to equate the buoyant force with the object's weight and instead focused directly on determining the displaced volume [71]. More generally, students have also been found to apply Newtonian equilibrium conditions inconsistently in hydrostatic contexts, for example by attributing a nonzero net force to an object at rest or inferring force magnitudes from an object's vertical position [71]. Such difficulties provide a rationale for structured guidance that helps learners coordinate the relevant quantities and physical relationships.

In the present study, however, this content served as the common physics problem-solving context for comparing support formats; students' conceptual understanding and conceptual change were not assessed. The task was selected because it provided a physics problem-solving context for which an established sequence of tiered hints and a model solution had already been developed within prior research on tiered support [8]. Using this task allowed the present study to extend prior work on tiered support to the context of AI-based learning support. The task is a typical school-physics problem that foregrounds the conceptual application of buoyancy in an age-appropriate way while deliberately reducing real-world complexity, which limits the extent to which the findings generalize to more open-ended modeling tasks.

2. Textual support development

The textual support material for the control group was developed through a systematic process of design, including expert feedback. Its creation was guided by a set of explicit design principles derived from textbook research, ensuring the material featured (i) a clear structure and didactically appropriate content for the ninth-grade target group; (ii) precise, linguistically correct text; and (iii) high-quality, supportive visualizations [73,74]. To verify its pedagogical quality and content validity, the resulting digital A4 page was reviewed and approved by two experts in physics education research. The final, self-contained document provided all necessary information for a student to solve the problem.

3. Tiered support development

For the scaffolding group, the intervention utilized a specific tiered support (*gestufte Lernhilfen*) task taken directly from the materials developed by Franke-Braun [8]. While the original design was intended for collaborative pair work, the materials were modified for individual problem solving to align with our study's protocol. The support was implemented digitally on a custom website where students could sequentially reveal hints on a step-by-step basis at their own discretion. The structure mirrored the original design: each tier consisted of a prompt or question that students could reveal, followed by a corresponding answer that could be unlocked in a second step. This learner-controlled process concluded with a final tier containing the complete model solution, allowing students to verify their work.

4. Chatbot development

The AI chatbot used in this study was developed using OpenAI's o3-mini model accessed via the Fobizz platform.² Its pedagogical behavior was meticulously prompt-engineered through a rigorous iterative refinement cycle, a methodology established in the development of educational technology [46,75,76]. During each cycle, the chatbot's responses to simulated student inputs were evaluated against three key pedagogical criteria derived from research on expert tutoring: (i) its ability to accurately assess a student's current understanding; (ii) its capacity to pose guiding, reflective questions rather than providing direct corrections; and (iii) its adherence to a "build on what's right" principle by first affirming correct elements of a student's reasoning before prompting them to reconsider flawed parts [77,78]. This iterative process was continued until the chatbot consistently met these standards, resulting in the final system prompt (included in the supplemental material [33]), which instructed the

²Accessible via <https://fobizz.com/en/>.

chatbot to act as a supportive tutor, primarily designed to guide students toward independent problem solving and conceptual understanding. Importantly, the chatbot was configured using a template-based design. The reusable component was the model-specific pedagogical interaction prompt; the task-specific component consisted of the problem statement and sample solution. This split reduces the adaptation effort for new tasks, because teachers do not need to author a complete sequence of hints and answers for each problem but can reuse the validated interaction template with task materials that are typically already available.

E. Data analysis

The data were analyzed in three steps: In the first step, pre-analytical checks were performed, e.g., by examining correlations between variables and verifying requirements for mixed linear effects modeling. In the next step, longitudinal variables (those with pretest and posttest measurements) were analyzed using mixed linear effects models, including the interaction between measurement time and experimental group as the primary effect, while accounting for the hierarchical nesting of students within classes. Finally, in the third step, post-only variables were evaluated using cross-sectional linear mixed-effects models. The following subsections describe each step in greater detail.

To complement the mixed effects modeling results and provide a more holistic picture of the data, we further report general descriptive statistics, including response distribution, mean values, and standard deviation (SD) as well as median values for all variables, split by intervention group. As mentioned above, we additionally provide Cronbach's α as a measure of internal consistency for each scale, where values above 0.70 are considered acceptable [79].

Data analysis was conducted using R (version 4.4.2) and its packages lme4 (version 2.0-1), lmerTest (version 3.2-1), performance (version 0.16.0), effectsize (version 1.0.1), and emmeans (version 1.11.1).

1. Pre-analytical checks

We first examined the relationship between all variables (cf. Table II), separated by pre- and post-test, using Spearman's rank-order correlation coefficient ρ —which is suitable for correlating variables even if the normality assumption is violated—in conjunction with the classification into weak ($|\rho| < 0.20$), medium ($0.20 \leq |\rho| < 0.30$) and strong ($|\rho| \geq 0.30$) correlations according to Hemphill [80]. Furthermore, since the assignment was performed at the class level, the requirements for mixed linear effects models were assessed by computing intraclass correlation coefficients (ICCs) and design effects (DEFF) for all outcome variables. Commonly used thresholds for multi-level modeling are $\text{ICC} \geq 0.15$ according to [81] and $\text{DEFF} \geq 2$ according to [82]. ICCs indicated non-trivial

clustering at the class level (ICC between 0.02 and 0.17 across all variables) alongside respective design effects (DEFF between 1.33 and 3.77 across variables), demonstrating between-class variance and justifying the use of mixed linear effects models (MLM). Model assumptions for MLM were evaluated for all mixed effects models through visual inspection of residual plots and Q - Q plots, and, additionally, fitted-versus-residual plots were examined to assess homoscedasticity [83,84]. No substantial deviations from normality or heteroscedasticity were observed. The cluster structure contained 16 different classes with a mean cluster size of $N_{\text{mean}} = 16.75$.

2. Data analysis of longitudinal variables

For outcome variables where both pretest and post-test measurements were available (i.e., enjoyment, hope, hopelessness, and self-efficacy), linear mixed effects modeling was performed. Specifically, in the Wilkinson-Rogers notation cf. [85], the following model specification was utilized:

$$\begin{aligned} \text{Outcome} \sim & \text{time} + \text{group} + \text{time} : \text{group} \\ & + \text{PriorKnowledge} \\ & + (1|\text{class}) \\ & + (1|\text{class} : \text{participant}). \end{aligned} \quad (1)$$

Here, the $\text{time} : \text{group}$ interaction term serves as the primary effect, evaluating whether the experimental groups exhibited differential trajectories over time—the chatbot condition (G1) was designated as the reference group. Prior knowledge was incorporated as a fixed effect to serve as a covariate. To account for the hierarchical structure of the data, random effects were specified for school classes, as well as for individual participants nested within those classes. Because participants remained within the same class across both pretest and post-test, the model was explicitly set up to ensure individuals could not cross clusters between measurement times. Random slopes for the time trend resulted in a singular fit and failed model convergence; thus, they were omitted from further analysis. Model parameters were estimated using restricted maximum likelihood (REML) to ensure unbiased estimates of the variance components [86].

To evaluate overall model goodness-of-fit, marginal and conditional R^2 metrics were computed as proposed in Ref. [87]:

- Marginal R^2 (R^2_{marg}) quantifies the proportion of variance explained exclusively by the fixed effects, thus reflecting population-level explanatory power.
- Conditional R^2 (R^2_{cond}) quantifies the proportion of variance explained by both fixed and random effects combined, thereby reflecting the full model fit including the group-level cluster structure.

TABLE III. Spearman's rank-order correlation coefficient ρ for all pairs of pretest variables: self-efficacy (SE), enjoyment (EE), hope (EH), hopelessness (EHL), prior knowledge (PK). Significant codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Diagonal self-correlations are omitted.

	SE	EE	EH	EHL	PK
SE					
EE	0.48***				
EH	0.70***	0.63***			
EHL	-0.52***	-0.28***	-0.49***		
PK	0.27***	0.27***	0.27***	-0.24***	

3. Data analysis of post-only variables

Variables assessed exclusively at post-test (i.e., intrinsic CL, extraneous CL exercise (ECLA), extraneous CL support (ECLL), and situational interest) were analyzed separately using cross-sectional mixed effects models. These models specified experimental condition as the sole fixed factor, prior knowledge as a covariate, and school class as a random intercept to account for the nested data structure:

$$\text{Outcome} \sim \text{group} + \text{PriorKnowledge} + (1|\text{class}). \quad (2)$$

Following the approach outlined in Ref. [88], inferential testing for fixed effects was based on type III analysis of covariance (ANCOVA) F -tests applying Satterthwaite approximations for degrees of freedom. Since the primary focus for the post-only outcomes—in contrast to the longitudinal variables—lies explicitly on exploring specific post-test group differences, these effects were further investigated via *post hoc* pairwise comparisons using estimated marginal means (EMMs) with Tukey adjustments to control for type I error inflation [89]. Effect sizes for all pairwise comparisons were operationalized as standardized mean differences (equivalent to Cohen's d). These estimates were computed via the `eff_size` function within the `emmeans` package, utilizing the

model's residual standard deviation as the standardizer and the model's residual degrees of freedom. Following standard conventions, we use established ranges of small ($d < 0.5$), medium ($0.5 \leq d < 0.8$) and large ($0.8 \leq d$) effect sizes [90], relative to the model's unexplained variance.

V. RESULTS

A. Descriptive results

Scale length and reliabilities are reported in Table II as part of the instrument overview.

The relationship between all variables in terms of Spearman's ρ is provided in Table III for the pretest and in Table IV for the post-test. Bivariate correlations across all measures revealed statistically significant associations ($p < 0.001$). In the pretest, their absolute values range from the medium correlation $\rho = 0.27$ (between prior knowledge and self-efficacy) to the strong correlation $\rho = 0.70$ (between hope and self-efficacy), while in the post-test, they exhibit a similar range from $\rho = 0.17$ (between self-efficacy and intrinsic CL) to $\rho = 0.71$ (between hope and enjoyment).

Descriptive statistics for all longitudinal variables are provided in Table VIII in the Appendix and for all post-only variables in Table IX, including means, standard deviations, medians, minima, and maxima.

Overall, the responses on each measure average on the scale's centers with means and medians clustering roughly between 2 and 3.5. The least average agreement can be found in group 3 (textbook) with regards to situational interest (SI; $m = 2.11$) and hope (EH; $m = 2.14$) while the highest average agreement is also observed in group 3 and relates to the intrinsic cognitive load (ICL; $m = 3.74$) as well as the extraneous cognitive load of the exercise (ECLA; $m = 3.42$). Notably, for each variable, responses are distributed along the entire scale length, as evidenced by all minima being 1 and all maxima being 5.

Since a between-subject multilevel ANCOVA was conducted for all post-only variables, we provide a more

TABLE IV. Spearman's rank-order correlation coefficient ρ for all pairs of post-test variables: enjoyment (EE), hope (EH), hopelessness (EHL), intrinsic CL (ICL), extraneous CL exercise (ECLA), extraneous CL support (ECLL), situational interest (SI), self-efficacy (SE). Significant codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Diagonal self-correlations are omitted.

	EE	EH	EHL	ICL	ECLA	ECLL	SI	SE
EE								
EH	0.71***							
EHL	-0.42***	-0.47***						
ICL	-0.22***	-0.25***	0.37***					
ECLA	-0.27***	-0.26***	0.45***	0.45***				
ECLL	-0.38***	-0.30***	0.44***	0.37***	0.51***			
SI	0.69***	0.60***	-0.29***	-0.06	-0.22***	-0.31***		
SE	0.50***	0.62***	-0.49***	-0.17**	-0.19**	-0.23***	0.51***	

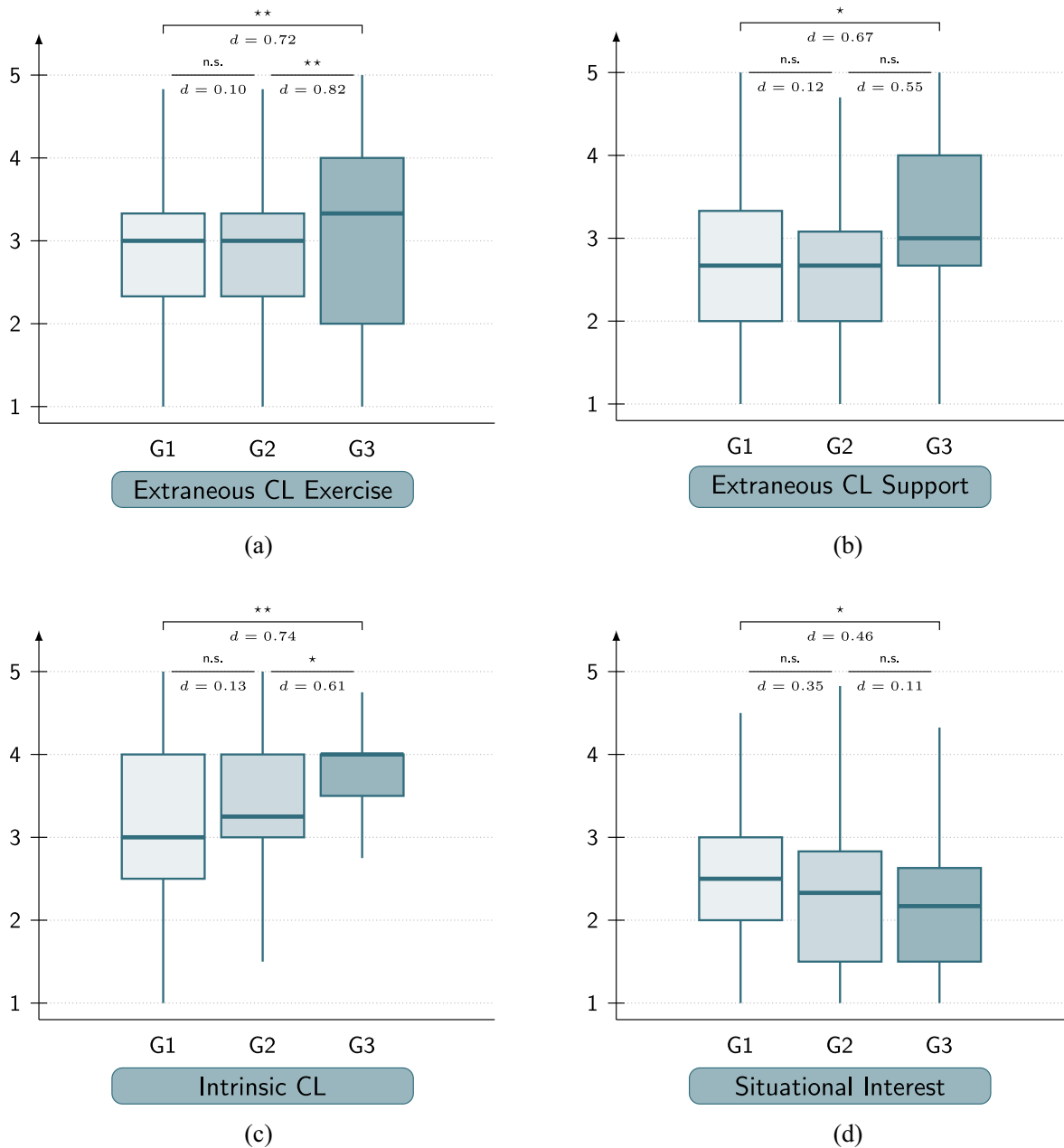


FIG. 3. Boxplots for the variables ECLA (a), ECLL (b), ICL (c), and SI (d), split by group—chatbots (G1), tiered support (G2), textbook (G3). In addition, absolute values of standardized mean differences as measures of effect sizes are reported alongside the respective significance level of the multilevel ANCOVA for each comparison.

detailed overview of the response distributions via boxplots in Fig. 3. Here, we followed the established standard of restricting the whisker length to 1.5 of the interquartile range. The boxplots reveal a consistent downward trend across the variables enjoyment, hope, situational interest, and self-efficacy, progressing from group G1 (chatbots) to group G3 (textbook). Conversely, an upward trend can be observed regarding hopelessness and all cognitive load variants. However, not all differences are statistically significant or have considerable effect sizes. A more

detailed, granular exploration is warranted, which will be presented in the next section.

B. Results regarding the longitudinal variables

Fixed effects estimates derived from the MLM, alongside their corresponding standard errors, degrees of freedom, t statistics and p values, are presented in Table V. Interaction terms are evaluated relative to the chatbot condition (G1). In addition, R^2 values are provided to indicate model fit. Random effect variances on the

TABLE V. Results of the linear mixed effects models for all longitudinal variables: self-efficacy (SE), enjoyment (EE), hope (EH), hopelessness (EHL). Fixed effects estimates (β) with standard errors (SE), approximated degrees of freedom (d.o.f.), t values, and p values are reported. Conditional (R^2_{cond}) and marginal (R^2_{marg}) goodness-of-fit indices are provided for each model. Significant effects are highlighted in bold font. G2: tiered support condition, G3: textbook support condition.

Variable	Predictor	β (SE)	d.o.f.	t	p	R^2_{cond}	R^2_{marg}
SE	Prior Knowledge	0.10 (0.02)	252.94	4.53	< 0.001	0.650	0.098
	time	−0.10 (0.06)	271.09	−1.68	0.093		
	G2	−0.18 (0.14)	22.11	−1.24	0.229		
	G3	−0.11 (0.14)	13.58	−0.83	0.422		
	time $\times$ G2	−0.05 (0.10)	271.93	−0.52	0.606		
	time $\times$ G3	−0.27 (0.09)	271.09	−3.09	0.002		
EE	Prior knowledge	0.09 (0.02)	249.75	3.90	< 0.001	0.533	0.170
	time	−0.35 (0.08)	272.00	−4.18	< 0.001		
	G2	0.09 (0.16)	23.38	0.54	0.596		
	G3	−0.07 (0.15)	14.09	−0.48	0.639		
	Time $\times$ G2	−0.34 (0.14)	272.00	−2.49	0.014		
	Time $\times$ G3	−0.43 (0.12)	272.00	−3.61	< 0.001		
EH	Prior knowledge	0.08 (0.02)	251.84	3.48	< 0.001	0.570	0.165
	Time	−0.37 (0.08)	272.00	−4.84	< 0.001		
	G2	−0.02 (0.15)	26.52	−0.12	0.910		
	G3	−0.12 (0.14)	16.22	−0.86	0.405		
	Time $\times$ G2	−0.19 (0.12)	272.00	−1.53	0.127		
	Time $\times$ G3	−0.38 (0.11)	272.00	−3.50	< 0.001		
EHL	Prior knowledge	−0.13 (0.03)	266.39	−4.71	< 0.001	0.632	0.106
	Time	0.00 (0.08)	272.00	0.03	0.977		
	G2	0.22 (0.20)	28.54	1.09	0.285		
	G3	0.03 (0.20)	15.96	0.15	0.882		
	Time $\times$ G2	0.17 (0.13)	272.00	1.32	0.188		
	Time $\times$ G3	0.55 (0.17)	272.00	4.70	< 0.001		

participant level (level 2) and class level (level 3), along with the level-1 residual variances, are provided in Table X in the Appendix.

Inferential testing revealed a significant interaction between measurement time and the textbook group (G3) for every dependent variable—relative to the chatbot reference group, the textbook condition exhibited a significantly sharper decline from pretest to post-test in self-efficacy (SE), enjoyment (EE), and hope (EH). An inverse pattern emerged for hopelessness (EHL); for this outcome, the textbook condition demonstrated a significantly greater increase over time compared to the chatbot group.

In contrast, the tiered support condition (G2) did not manifest any statistically significant interactions with time apart from enjoyment (EE), thus performing comparably to the chatbot reference group.

Additionally, prior knowledge served as a statistically significant predictor across all analyzed longitudinal variables.

No statistically significant differences are observed between the conditions prior to the intervention, as evidenced by the nonsignificant interaction terms of the predictors G2 and G3 across all variables.

The fixed effects accounted for between 9.8% (SE) and 17.0% (EE) of the total variance, as indicated by the R^2_{marg} values. When accounting for the hierarchical cluster structures, however, the total proportion of explained variance increased substantially. The full conditional models explained between 53.3% (EE) and 65.0% (SE) of the total variance, signaling prominent participant- and class-specific clustering.

C. Results regarding the post-only variables

The results of the type III F tests from the linear mixed-effects model are compiled in Table VI. The table details the F -statistics, numerator degrees of freedom, Satterthwaite-approximated denominator degrees of freedom, and corresponding p values. These statistics are reported separately for the main effect of the experimental condition (group) and the covariate prior knowledge across each variable. In addition, conditional (R^2_{cond}) and marginal (R^2_{marg}) goodness-of-fit indices are provided.

Random effect variances on the class level (level 2), along with the level-1 residual variances (representing the participant level), are provided in Table XI.

TABLE VI. Overall multilevel ANCOVA results for all post-only variables: intrinsic cognitive load (ICL), extraneous cognitive load exercise (ECLA), extraneous cognitive load support (ECLL), and situational interest (SI). For each model, effects are reported for the experimental condition (group) and the covariate prior knowledge. Statistically significant p values ($p < 0.05$) are shown in bold.

Variable	Predictor	F (NumDF, DenDF)	p	R^2_{cond}	R^2_{marg}
ICL	Group	$F(2, 16.41) = 5.96$	0.011	0.17	0.10
	Prior knowledge	$F(1, 262.64) = 0.79$	0.376		
ECLA	Group	$F(2, 15.82) = 8.70$	0.002	0.19	0.13
	Prior knowledge	$F(1, 255.95) = 9.03$	0.003		
ECLL	Group	$F(2, 16.97) = 4.41$	0.029	0.17	0.09
	Prior knowledge	$F(1, 264.81) = 4.10$	0.044		
SI	Group	$F(2, 12.04) = 4.22$	0.041	0.06	0.04
	Prior knowledge	$F(1, 219.79) = 1.40$	0.239		

The results of the respective *post hoc* pairwise comparisons for all post-only variables are compiled in Table VII. The table details the model-based estimated marginal mean differences (ΔM) alongside their standard errors (SEs), Tukey-adjusted p values, and standardized mean differences (Cohen's d) with corresponding 95% confidence intervals. Each comparison evaluates the directional differences between the chatbot support (G1), tiered support (G2), and textbook support (G3) conditions.

1. Intrinsic cognitive load

Regarding intrinsic cognitive load (ICL), the linear mixed model revealed a statistically significant omnibus effect of experimental condition [$F(2, 16.41) = 5.96$, $p = 0.011$, $R^2_{\text{cond}} = 0.17$, $R^2_{\text{marg}} = 0.10$]. Prior knowledge did not serve as a significant predictor within this model. *Post hoc* pairwise comparisons demonstrated a statistically significant difference between the chatbot condition and the textbook condition, favoring the chatbot framework [$\Delta M = -0.565$, $p = 0.006$, $d = -0.74$]. The tiered-support condition also differed significantly from the textbook condition [$\Delta M = -0.466$, $p = 0.027$, $d = -0.61$]. The chatbot and tiered-support conditions did not differ significantly. All group contrasts as well as all multilevel ANCOVA results are provided in Tables VI and VII, and the distribution profiles for all groups are illustrated via the boxplots in Fig. 3.

2. Extraneous cognitive load exercise

Regarding extraneous cognitive load exercise (ECLA), the linear mixed model revealed a statistically significant omnibus effect of experimental condition [$F(2, 15.82) = 8.70$, $p = 0.002$, $R^2_{\text{cond}} = 0.19$, $R^2_{\text{marg}} = 0.13$]. Prior knowledge also served as a significant predictor within

this model [$F(1, 255.95) = 9.03$, $p = 0.003$]. *Post hoc* pairwise comparisons demonstrated a statistically significant difference between the chatbot condition and the textbook condition, favoring the chatbot framework [$\Delta M = -0.517$, $p = 0.004$, $d = -0.72$]. Likewise, a statistically significant difference was observed between the tiered support condition and the textbook condition, favoring the tiered support framework [$\Delta M = -0.590$, $p = 0.003$, $d = -0.82$]. However, the pairwise comparison between the chatbot condition and the tiered support

TABLE VII. Results of the *post hoc* pairwise comparisons for all post-only variables: intrinsic cognitive load (ICL), extraneous cognitive load exercise (ECLA), extraneous cognitive load support (ECLL), and situational interest (SI). Model-based estimated marginal mean differences (ΔM) with standard errors (SEs), adjusted p values, and standardized mean differences (d) are reported for each group comparison. p values are adjusted using the Tukey method to control for type I error inflation. Significant effects are highlighted in bold font.

Variable	Comparison	ΔM (SE)	p	d
ICL	G1–G2	−0.100 (0.181)	0.594	−0.13
	G1–G3	−0.565 (0.172)	0.006	−0.74
	G2–G3	−0.466 (0.185)	0.027	−0.61
ECLA	G1–G2	0.073 (0.159)	0.653	0.10
	G1–G3	−0.517 (0.148)	0.004	−0.72
	G2–G3	−0.590 (0.161)	0.003	−0.82
ECLL	G1–G2	−0.106 (0.221)	0.638	−0.12
	G1–G3	−0.598 (0.211)	0.015	−0.67
	G2–G3	−0.491 (0.227)	0.051	−0.55
SI	G1–G2	0.274 (0.145)	0.086	0.35
	G1–G3	0.362 (0.130)	0.018	0.46
	G2–G3	0.087 (0.145)	0.558	0.11

condition did not reach statistical significance. All group contrasts as well as all multilevel ANCOVA results are provided in Tables VI and VII, and the distribution profiles for all groups are illustrated via the boxplots in Fig. 3.

3. Extraneous cognitive load support

Regarding extraneous cognitive load support (ECLL), the linear mixed model revealed a statistically significant omnibus effect of experimental condition [$F(2, 16.97) = 4.41$, $p = 0.029$, $R^2_{\text{cond}} = 0.17$, $R^2_{\text{marg}} = 0.09$]. Prior knowledge also served as a significant predictor within this model [$F(1, 264.81) = 4.10$, $p = 0.044$]. *Post hoc* pairwise comparisons demonstrated a statistically significant difference between the chatbot condition and the textbook condition, favoring the chatbot framework [$\Delta M = -0.598$, $p = 0.015$, $d = -0.67$]. No other pairwise comparison reached statistical significance. All group contrasts as well as all multilevel ANCOVA results are provided in Tables VI and VII, and the distribution profiles for all groups are illustrated via the boxplots in Fig. 3.

4. Situational interest

Regarding situational interest (SI), the linear mixed model revealed a statistically significant omnibus effect of experimental condition [$F(2, 12.04) = 4.22$, $p = 0.041$, $R^2_{\text{cond}} = 0.06$, $R^2_{\text{marg}} = 0.04$]. Prior knowledge did not serve as a significant predictor within this model. *Post hoc* pairwise comparisons demonstrated a statistically significant difference between the chatbot condition and the textbook condition, favoring the chatbot framework [$\Delta M = 0.362$, $p = 0.018$, $d = 0.46$]. No other pairwise comparison reached statistical significance. All group contrasts as well as all multilevel ANCOVA results are provided in Tables VI and VII, and the distribution profiles for all groups are illustrated via the boxplots in Fig. 3.

VI. DISCUSSION

This study aimed to compare the effects of an AI chatbot, tiered support, and a traditional digital textbook-style page on students' cognitive and affective-motivational experiences during physics problem solving. Overall, the results indicate a consistent pattern across the three conditions. Students in the AI chatbot and tiered support conditions generally reported lower intrinsic and extraneous cognitive load than students in the static textual support condition, although the difference between tiered support and textual support did not reach statistical significance for extraneous cognitive load related to the support material. For the longitudinal affective-motivational variables, the problem-solving activity appeared demanding across conditions: positive affective-motivational task-related states tended to decrease from pretest to post-test, while hopelessness tended to increase. Relative to the chatbot condition, the

textual-support condition showed a statistically significantly less favorable trajectory for these variables, whereas differences between chatbot and tiered support were mostly not statistically significant. For the post-only variable situational interest, which refers to immediate interest in the specific buoyancy task with the help of a support tool, students in the chatbot condition reported significantly higher situational interest than students in the textual-support condition, whereas the other pairwise comparisons were not statistically significant. Descriptively, however, situational interest followed the same general ordering as the longitudinal affective-motivational variables, with the most favorable values in the chatbot condition, followed by tiered support and then textual support.

The findings concerning perceived extraneous cognitive load are consistent with assumptions from cognitive load theory. However, given the relatively low internal consistencies of the cognitive load scales, these differences should be interpreted as suggestive patterns in students' perceived task demands rather than as precise estimates of differences in cognitive processing. One possible explanation for differing cognitive load across conditions could be that students in the textual-support condition were required to coordinate several demanding processes at once, simultaneously attempting to understand the physics problem while searching, sorting, and evaluating a dense block of text for relevant information. In particular, the need to connect information from the problem statement with different parts of the textbook-style page may have created a split-attention effect, thereby consuming working memory resources that could otherwise have been directed toward the conceptual aspects of the buoyancy problem [91].

In contrast, both the tiered support and chatbot systems were designed to reduce this need for searching for information by providing more structured guidance [92]. Although the difference between tiered support and textual support did not reach statistical significance for extraneous cognitive load related to the support material, the descriptive pattern pointed in the same direction as the other cognitive-load measures. The tiered support system structured the solution path, presenting information in predigested, sequential steps that eliminated the need for students to search or plan. Similarly, the chatbot provided on-demand, targeted guidance, allowing students to offload the "search and sort" process to the AI. However, such reduction should not be understood as beneficial in every respect. Searching, selecting, and organizing information within a text are valuable scientific and metacognitive skills that are worth fostering in physics education. Whether these demands should be reduced or deliberately maintained depends on the learning goal for the specific lesson. In the present task, where the focus was on applying the concept of buoyancy rather than on developing research or information-search competence, reducing some of these demands may have allowed students to allocate

more cognitive resources to the conceptual aspects of the problem [93].

To add nuance, the AI chatbot condition may have been subject to two competing effects on extraneous load. On the one hand, the chatbot may have reduced extraneous load when students used it as intended: by eliciting targeted self-explanations [36,59] and leveraging the collective working memory effect [58]. On the other hand, the very nature of this interaction might have introduced a “hidden load of interactivity” that can increase extraneous load. Students must formulate a coherent text, type it, interpret the AI’s potentially imperfect response, and critically evaluate its trustworthiness.

Beyond extraneous load, students in the chatbot and tiered-support conditions reported lower intrinsic cognitive load than students in the textual-support condition. One possible explanation is that textual support may have left students to manage the conceptual complexity of the buoyancy problem largely on their own. Without external support to help them activate prior knowledge or structure the task, the complexity of the topic may have imposed a higher demand on their working memory, resulting in a high perceived intrinsic load. In contrast, both interactive tools may have acted as a bridge for this knowledge gap, for instance, through isolating interacting elements [37], breaking the complex topic down into manageable sub-tasks without altering the core conceptual demands.

A complementary or alternative explanation is that lower perceived intrinsic load may partly reflect cognitive off-loading rather than more productive management of task complexity [94]. In the chatbot condition, students may have gamed the system, using the available help in a way that reduced their own cognitive effort, for example, by requesting hints, procedural instructions, or solution steps prematurely [95]. In such cases, lower perceived intrinsic load would not indicate that students developed a more coherent understanding of the buoyancy problem but could instead reflect that demanding parts of the problem-solving process were shifted to the support system. A related possibility is that the chatbot induced an illusion of understanding: coherent explanations, immediate feedback, and a sense of being guided may have made the task feel more comprehensible even if this did not translate into deeper conceptual understanding. From this perspective, the chatbot may have increased students’ subjective sense of comprehension by making the task appear easier to process, a mechanism that has been discussed in research on metacognitive illusions and fluency-based judgments of learning [96,97]. This critical interpretation is in line with work by Krupp *et al.* [53] and Stadler *et al.* [55]. From this perspective, the higher perceived intrinsic load in the textual-support condition may not only reflect overload but could also partly indicate more direct engagement with the conceptual complexity of the task.

The findings of this study reflect the interplay between the cognitive and affective-motivational dimensions of learning. Across conditions, lower perceived cognitive load was associated with more favorable task-related affective-motivational responses. The chatbot group reported the least unfavorable affective-motivational profile, while the textual support group reported the least positive profile. This relationship is further substantiated by significant correlations, which show that higher extraneous cognitive load of the support tool correlates with lower feelings of enjoyment ($\rho = -0.38, p < 0.001$) and hope ($\rho = -0.30, p < 0.001$), while associated with higher hopelessness ($\rho = 0.44, p < 0.001$). This connection is well supported theoretically; high extraneous cognitive load can deplete cognitive resources and induce frustration, while positive emotions may in turn act as a buffer, assisting students in regulating their response to a demanding task [93,98–100].

A nuance emerges when comparing the two interactive support systems. While both were generally associated with lower cognitive load than the textual support, the chatbot condition showed a significantly less unfavorable affective-motivational profile than the textual-support condition. This pattern suggests that the chatbot may have helped buffer the affective-motivational challenges of the task. One possible explanation is that the chatbot provided interactional cues that made the support experience feel more responsive and socially engaging. This interpretation aligns with research on AI tutors, chatbots, and pedagogical agents, suggesting that interactive, personalized, and affectively supportive learning cues—such as timely feedback, adaptive guidance, enthusiastic or encouraging agent behaviors, and socially responsive interaction—can contribute to students’ motivation and engagement [14,16,46,101–103].

Our findings on the affective-motivational pattern of the tiered support group complement earlier research by highlighting the importance of the social learning context: Seminal studies by Franke-Braun [8], Schmidt-Weigand *et al.* [10] or Hänze *et al.* [9] found that students in tiered support conditions reported more favorable affective-motivational outcomes than students in control conditions, including perceived competence, intrinsic motivation, and enjoyment. In our study, the tiered-support condition tended to fall between the chatbot and textual-support conditions descriptively. One possible explanation for this difference lies in the social learning context. Whereas earlier studies were conducted with students working in collaborative pairs, our study focused on individual problem solving. It is plausible that the affective benefits previously observed were tied not only to the hints themselves but also to the successful peer interaction they facilitated.

The chatbot’s conversational responsiveness provides a theoretical reason to expect a more favorable affective-motivational profile than in less interactive support

formats. However, the chatbot did not show a statistically significant advantage over tiered support. This finding suggests that, in a short single-problem intervention, receiving functional and structured help may have been more important for students' affective-motivational experience than the specific interaction style through which this help was provided. Both interactive formats may therefore have buffered the unfavorable affective-motivational profile observed during the task, whereas the textual-support condition offered less immediate guidance. At the same time, the absence of a significant chatbot–tiered-support difference should not be interpreted as evidence of equivalence. The smaller tiered-support group may have limited the sensitivity to detect small differences between these two conditions.

VII. LIMITATIONS

While this study provides valuable insights, its conclusions should be considered in light of several limitations that also frame avenues for future research. First, the intervention's brief, single-session design limits our conclusions to short-term effects. This snapshot approach precludes the observation of longitudinal dynamics, such as the decay of novelty effects or the potential emergence of familiarization effects [104–106]. The “good enough” principle we observed may be an artifact of this short duration. Second, our reliance on self-report scales introduces measurement limitations. While standard practice, such measures are susceptible to common method biases such as social desirability [107]. Furthermore, the measures for intrinsic cognitive load (ICL, $\alpha = 0.58$), extraneous cognitive load exercise (ECLA, $\alpha = 0.61$), and prior knowledge (PK, $\alpha = .63$) fell below the conventional Cronbach's α threshold of 0.70 [79]. The findings involving these variables should therefore be interpreted with caution. In particular, the group differences in perceived intrinsic cognitive load and extraneous cognitive load related to the exercise may partly reflect measurement error, and the role of prior knowledge as a covariate may have been estimated less precisely than would be possible with a more reliable measure. Third, the sample of $N = 268$ may have been insufficient to detect small effect sizes. The consistent but nonsignificant trends favoring the chatbot over the tiered support on several affective measures might represent such small, yet potentially meaningful, effects. Future research using larger samples or more targeted designs would be needed to resolve these subtle differences definitively. Finally, the findings are specific to the particular physics task and to the specific technological implementations used in this study. The findings may therefore not generalize directly to more open-ended modeling tasks. In addition, the AI chatbot was built on a particular LLM (OpenAI's o3-mini) with a tailored system prompt, and its

performance is inherently tied to this design. This is a critical consideration, as recent work suggests that the pedagogical design of an AI tutor is a key variable, even if the optimal approach for fostering student learning remains an open and unresolved question [57].

VIII. OUTLOOK

This study offers foundational insights into the differential impacts of AI chatbots, tiered, and traditional support in physics education, laying the groundwork for a targeted future research agenda. Methodologically, future work should aim to increase both rigor and duration. While our cluster-randomized design maximized ecological validity, controlled laboratory experiments with true randomization of individuals would strengthen causal claims. To better explore effects over time, the intervention could be extended with a second problem-solving phase and a follow-up assessment. Such a pre-post-follow-up design would not only allow for the measurement of learning outcomes and retention but would also critically test our “good enough” hypothesis by revealing whether the nuanced affective benefits of chatbots emerge with prolonged use. A deeper understanding of the underlying learning processes is also essential. Future studies should analyze how different scaffolding methods shape students' applied problem-solving strategies, for instance, through think-aloud protocols. For the chatbot condition specifically, analyzing chat logs to identify the types of prompts students use can offer invaluable insight into their help-seeking behaviors and conceptual hurdles. Furthermore, it is crucial to investigate how student factors—including general self-regulation skills, text-based self-regulation strategies, and reading strategies—predict motivational, emotional, and learning outcomes, and how they moderate the effectiveness of different support formats, particularly in text-heavy support conditions [108], thereby informing the development of more adaptive and equitable support systems. Finally, future research should disentangle the effects of the support tool from the social context of its use by employing a 2×2 factorial design, crossing the support type (tiered hints vs chatbot) with the learning setting (individual vs collaborative). Such research, combined with systematic explorations into chatbot pedagogical design (e.g., prompt engineering and feedback style), is paramount for harnessing the potential of AI chatbots.

ACKNOWLEDGMENTS

This work is supported by the Open Access Publishing Fund of Leipzig University.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo [109].

APPENDIX

TABLE VIII. Descriptive statistics for all longitudinal variables, split by measurement time and group—chatbots (G1), tiered support (G2), textbook (G3).

(a) Pre/post-descriptives for the variable self-efficacy (SE).					
Group	Mean (pre/post)	SD (pre/post)	Median (pre/post)	Minimum (pre/post)	Maximum (pre/post)
G1	2.99/2.89	0.64/0.77	3.00/3.00	1/1	5/5
G2	2.88/2.72	0.63/0.72	3.00/2.86	1/1	5/5
G3	2.93/2.55	0.80/0.81	3.00/2.64	1/1	5/5
(b) Pre/post-descriptives for the variable enjoyment (EE).					
Group	Mean (pre/post)	SD (pre/post)	Median (pre/post)	Minimum (pre/post)	Maximum (pre/post)
G1	3.00/2.65	0.70/0.93	3.00/2.50	1/1	5/5
G2	3.15/2.47	0.72/0.88	3.25/2.50	1/1	5/5
G3	2.99/2.20	0.84/0.89	3.00/2.25	1/1	5/5
(c) Pre/post-descriptives for the variable hope (EH).					
Group	Mean (pre/post)	SD (pre/post)	Median (pre/post)	Minimum (pre/post)	Maximum (pre/post)
G1	2.99/2.62	0.71/0.87	3.00/2.75	1/1	5/5
G2	3.01/2.45	0.64/0.87	3.00/2.50	1/1	5/5
G3	2.90/2.14	0.78/0.84	3.00/2.25	1/1	5/5
(d) Pre/post-descriptives for the variable hopelessness (EHL).					
Group	Mean (pre/post)	SD (pre/post)	Median (pre/post)	Minimum (pre/post)	Maximum (pre/post)
G1	2.60/3.18	0.89/0.82	2.50/3.00	1/1	5/5
G2	2.78/3.27	0.96/0.75	2.75/3.25	1/1	5/5
G3	2.61/3.74	0.90/0.77	2.50/4.00	1/1	5/5

TABLE IX. Descriptive statistics for prior knowledge as well as for all post-only variables, split by group—chatbots (G1), tiered support (G2), textbook (G3).

(a) Descriptives for the variable prior knowledge.					
Group	Mean	SD	Median	Minimum	Maximum
G1	5.56	1.79	5.50	1	9
G2	5.85	1.92	6.00	2	9
G3	6.04	2.00	6.00	1	9
(b) Descriptives for the variable ICL.					
Group	Mean	SD	Median	Minimum	Maximum
G1	3.18	0.82	3.00	1	5
G2	3.27	0.75	3.25	1	5
G3	3.74	0.77	4.00	1	5
(c) Descriptives for the variable ECLA.					
Group	Mean	SD	Median	Minimum	Maximum
G1	2.92	0.76	3.00	1	5
G2	2.84	0.70	3.00	1	5
G3	3.42	0.75	3.33	1	5
(d) Descriptives for the variable ECLL.					
Group	Mean	SD	Median	Minimum	Maximum
G1	2.69	0.93	2.67	1	5
G2	2.74	0.98	2.67	1	5
G3	3.26	0.92	3.00	1	5
(e) Descriptives for the variable SI.					
Group	Mean	SD	Median	Minimum	Maximum
G1	2.46	0.83	2.50	1	5
G2	2.21	0.77	2.33	1	5
G3	2.11	0.77	2.17	1	5

TABLE X. Random effect variance components with standard deviations in parentheses for the linear mixed effects models of the longitudinal variables.

Variable	Student level σ^2	Class level σ^2	Residual σ^2
SE	0.29 (0.54)	0.03 (0.16)	0.20 (0.45)
EE	0.26 (0.51)	0.03 (0.17)	0.38 (0.61)
EH	0.27 (0.52)	0.03 (0.16)	0.32 (0.56)
EHL	0.43 (0.66)	0.07 (0.27)	0.35 (0.59)

TABLE XI. Random effect variance components with standard deviations in parentheses for the multilevel ANCOVAs of the post-only variables.

Variable	Class level σ^2	Residual σ^2
ICL	0.05 (0.23)	0.58 (0.76)
ECLA	0.03 (0.19)	0.51 (0.72)
ECLL	0.08 (0.29)	0.81 (0.90)
SI	0.01 (0.12)	0.62 (0.79)

- [1] A. Maries, S.-Y. Lin, and C. Singh, Challenges in designing appropriate scaffolding to improve students' representational consistency: The case of a Gauss's law problem, *Phys. Rev. Phys. Educ. Res.* **13**, 020103 (2017).
- [2] B. R. Belland, *Instructional Scaffolding in STEM Education: Strategies and Efficacy Evidence* (Springer, Cham, Switzerland, 2017).
- [3] UNESCO, *Global Education Monitoring Report 2020: Inclusion and Education: All Means All* (UNESCO, 2020), p. 502.
- [4] C. A. Tomlinson, *The Differentiated Classroom: Responding to the Needs of all Learners*, 2nd ed. (ASCD, Alexandria, VA, 2014).
- [5] C. A. Tomlinson, *How to Differentiate Instruction in Academically Diverse Classrooms*, 3rd ed. (ASCD, Alexandria, VA, 2017).
- [6] D. Wood, J. S. Bruner, and G. Ross, The role of tutoring in problem solving, *J. Child Psychol. Psychiatry* **17**, 89 (1976).
- [7] L. Stäudel and R. Wodzinski, Komplexität erhalten und gezielt unterstützen: Aufgaben mit gestuften Lernhilfen im naturwissenschaftlichen Unterricht, in *Selbstbestimmung und Classroom-Management: Empirische Befunde und Entwicklungsstrategien zum guten Unterricht*, edited by T. Bohl, K. Kansteiner-Schänzlin, M. Kleinknecht, B. Kohler, and A. Nold (Verlag Julius Klinkhardt, Bad Heilbrunn, 2010), pp. 236–253.
- [8] G. Franke-Braun, *Aufgaben mit gestuften Lernhilfen: Ein Aufgabenformat zur Förderung der sachbezogenen Kommunikation und Lernleistung für den naturwissenschaftlichen Unterricht*, Studien zum Physik- und Chemielernen (Logos Verlag, Berlin, 2008), Vol. 88.
- [9] M. Hänze, F. Schmidt-Weigand, and S. Blum, Mit gestuften Lernhilfen selbständig lernen und arbeiten, in *Kooperatives und selbständiges Lernen von Schülern*, edited by K. Rabenstein and S. Reh (VS-Verlag, Wiesbaden, 2007), pp. 197–208.
- [10] F. Schmidt-Weigand, G. Franke-Braun, and M. Hänze, Erhöhen gestufte Lernhilfen die Effektivität von Lösungsbeispielen? Eine Studie zur kooperativen Bearbeitung von Aufgaben in den Naturwissenschaften, *Unterrichtswissenschaft* **36**, 365 (2008).
- [11] Robert Bosch Stiftung, Deutsches Schulbarometer: Befragung Lehrkräfte. Ergebnisse zur aktuellen Lage an allgemein- und berufsbildenden Schulen, <https://www.bosch-stiftung.de/de/publikation/deutsches-schulbarometer> (2024).
- [12] K. Neumann, J. Kuhn, and H. Drachsler, Generative Künstliche Intelligenz in Unterricht und Unterrichtsforschung—Chancen und Herausforderungen, *Unterrichtswissenschaft* **52**, 227 (2024).
- [13] Y. Liang, D. Zou, H. Xie, and F. L. Wang, Exploring the potential of using ChatGPT in physics education, *Smart Learn. Environ.* **10**, 52 (2023).
- [14] C.-H. Chen and C.-L. Chang, Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior, *Educ. Inf. Technol.* **29**, 18621 (2024).
- [15] Y. Ji, X. Zou, T. Li, and Z. Zhan, The effectiveness of ChatGPT on pre-service teachers' STEM teaching literacy, learning performance and cognitive load in STEM teacher training courses, in *Proceedings of the 2023 14th International Conference on Educational Technology and Management (ICETM)* (ACM, Guangzhou, China, 2023).
- [16] D. T. K. Ng, C. W. Tan, and J. K. L. Leung, Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study, *Br. J. Educ. Technol.* **55**, 1328 (2024).
- [17] D. M. Johnson, W. Doss, and C. M. Estep, Using ChatGPT with novice Arduino programmers: Effects on performance, interest, self-efficacy, and programming ability, *J. Res. Tech. Careers* **8**, 1 (2024).
- [18] K.-L. Huang, Y.-C. Liu, and M.-Q. Dong, Incorporating AIGC into design ideation: A study on self-efficacy and learning experience acceptance under higher-order thinking, *Thinking Skills Creativity* **52**, 101508 (2024).
- [19] R. Yilmaz and F. G. Karaoglan Yilmaz, The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation, *Comput. Educ. Artif. Intell.* **4**, 100147 (2023).
- [20] R. Pekrun, The control-value theory of achievement emotions: Assumptions, corollaries, and implications

- for educational research and practice, *Educ. Psychol. Rev.* **18**, 315 (2006).
- [21] A. Bandura, Self-efficacy: Toward a unifying theory of behavioral change, *Adv. Behav. Res. Therapy* **1**, 139 (1978).
- [22] S. Hidi and K. A. Renninger, The four-phase model of interest development, *Educ. Psychol.* **41**, 111 (2006).
- [23] J. Sweller, Cognitive load during problem solving: Effects on learning, *Cognit. Sci.* **12**, 257 (1988).
- [24] A. Collins, J. S. Brown, and S. E. Newman, Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics, in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, edited by L. B. Resnick (Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1989), pp. 453–494.
- [25] F. Ogan-Bekiroglu, To what degree do the currently used physics textbooks meet the expectations?, *J. Sci. Teach. Educ.* **18**, 599 (2007).
- [26] G. Jonas-Ahrend, M. Kapanadze, A. Mazzolini, and F. Joubran, Ergebnisse einer Reviewstudie zur Evaluation von Physiklehrbüchern, in *Proceedings of GDGP Jahrestagung 2023* (GDGP, Hamburg, 2023).
- [27] M. French, F. Taverna, M. Neumann, L. P. Kushnir, J. Harlow, D. Harrison, and R. Serbanescu, Textbook use in the sciences and its relation to course performance, *Coll. Teach.* **63**, 171 (2015).
- [28] C. M. Burchfield and J. Sappington, Compliance with required reading assignments, *Teach. Psychol.* **27**, 58 (2000).
- [29] P. A. Alexander and J. M. Kulikowich, Learning from physics text: A synthesis of recent research, *J. Res. Sci. Teach.* **31**, 895 (1994).
- [30] D. J. Cunningham, T. M. Duffy, and R. A. Knuth, The textbook of the future, in *Hypertext: A Psychological Perspective*, edited by C. McKnight, A. Dillon, and J. Richardson (E. Horwood, New York, 1993), pp. 19–50.
- [31] C. van Boxtel, J. van der Linden, and G. Kanselaar, The use of textbooks as a tool during collaborative physics learning, *J. Exp. Educ.* **69**, 57 (2000).
- [32] J. Leisen, *Methoden-Handbuch Deutschsprachiger Fachunterricht (DFU)* (Varus, Bonn, 1999).
- [33] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/nvfl-zrq8> for the complete instructional materials used in the three experimental conditions, including the buoyancy task, the chatbot system prompt, the tiered support, and the textbook-style support.
- [34] A. Renkl and R. K. Atkinson, Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective, *Educ. Psychol.* **38**, 15 (2003).
- [35] R. K. Atkinson, S. J. Derry, A. Renkl, and D. Wortham, Learning from examples: Instructional principles from the worked examples research, *Rev. Educ. Res.* **70**, 181 (2000).
- [36] J. Sweller, J. J. G. van Merriënboer, and F. Paas, Cognitive architecture and instructional design: 20 years later, *Educ. Psychol. Rev.* **31**, 261 (2019).
- [37] E. Pollock, P. Chandler, and J. Sweller, Assimilating complex information, *Learn. Instr.* **12**, 61 (2002).
- [38] R. E. Mayer and R. Moreno, Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning, in *Cognitive Load Theory*, edited by J. L. Plass, R. Moreno, and R. Brünken (Cambridge University Press, Cambridge, England, 2010), p. 152.
- [39] R. K. Atkinson, A. Renkl, and M. M. Merrill, Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps, *J. Educ. Psychol.* **95**, 774 (2003).
- [40] F. Schmidt-Weigand, M. Hünze, and R. Wodzinski, Complex problem solving and worked examples: The role of prompting strategic behavior and fading-in solution steps, *Z. Pädagog. Psychol.* **23**, 129 (2009).
- [41] W. Huang, J. Jiang, R. B. King, and L. K. Fryer, Chatbots and student motivation: A scoping review, *Int. J. Educ. Technol. Higher Educ.* **22**, 26 (2025).
- [42] E. Tufino and B. Gregorcic, Creating a customisable Socratic AI physics tutor, *Phys. Educ.* **60**, 065037 (2025).
- [43] E. Kasneci *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Diff.* **103**, 102274 (2023).
- [44] X. Han, H. Peng, and M. Liu, The impact of GenAI on learning outcomes: A systematic review and meta-analysis of experimental studies, *Educ. Res. Rev.* **48**, 100714 (2025).
- [45] R. Wu and Z. Yu, Do AI chatbots improve students' learning outcomes? Evidence from a meta-analysis, *Br. J. Educ. Technol.* **55**, 10 (2024).
- [46] G. Kestin, K. Miller, A. Klales, T. Milbourne, and G. Ponti, *AI Tutoring Outperforms Active Learning* (Research Square, 2024), [10.21203/rs.3.rs-4243877/v1](https://doi.org/10.21203/rs.3.rs-4243877/v1).
- [47] T. Li, Y. Ji, and Z. Zhan, Expert or machine? Comparing the effect of pairing student teacher with in-service teacher and ChatGPT on their critical thinking, learning performance, and cognitive load in an integrated-STEM course, *Asia Pac. J. Educ.* **44**, 45 (2024).
- [48] K. Alarbi, M. Halaweh, H. Tairab, N. R. Alsahhi, 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**, ep448 (2023).
- [50] J. Lademann, J. Henze, and S. Becker-Genschow, Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables, *Phys. Rev. Phys. Educ. Res.* **21**, 010147 (2025).
- [51] P. Wulff and M. Kubsch, Learning against the machine: The double edged sword of (Gen)AI in STEM education, *Int. J. STEM Educ.* **12**, 66 (2025).
- [52] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakçı, and R. Mariman, Generative AI without guardrails can harm learning: Evidence from high school mathematics, *Proceed. Natl. Acad. Sci.* **122**, e2422633122 (2025).
- [53] L. Krupp, S. Steinert, M. Kiefer-Emmanouilidis, K. E. Avila, P. Lukowicz, J. Kuhn, S. Küchemann, and

- J. Karolus, 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* (IOS Press, Amsterdam, Netherlands, 2024), pp. 199–212.
- [54] J. Zheng, L. Hao, K. Lu, A. Garg, M. Reese, M.-J. Yap, I.-J. Wang, X. Wu, W. Huang, J. Hoffman, A. Kelly, M. Le, R. Zhang, Y. Lin, M. Faayez, and A. Liu, Do students rely on AI? Analysis of student-ChatGPT conversations from a field study (2025).
- [55] M. Stadler, M. Bannert, and M. Sailer, Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry, *Comput. Hum. Behav.* **160**, 108386 (2024).
- [56] Y. Wang and S. Wang, The advantages and disadvantages of ChatGPT in humanities, social sciences and STEM education: From a comparative perspective, in *Proceedings of the International Conference on Global Politics and Socio-Humanities* (2024), pp. 108–114.
- [57] A. Lieb and T. Goel, Student interaction with NewtBot: An LLM-as-tutor chatbot for secondary physics education, in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24* (Association for Computing Machinery, New York, NY, USA, 2024).
- [58] F. Kirschner, F. Paas, and P. A. Kirschner, A cognitive load approach to collaborative learning: United brains for complex tasks, *Educ. Psychol. Rev.* **21**, 31 (2009).
- [59] A. Renkl, R. Stark, H. Gruber, and H. Mandl, Learning from worked-out examples: The effects of example variability and elicited self-explanations, *Contemp. Educ. Psychol.* **23**, 90 (1998).
- [60] A. Maries and C. Singh, Helping students become proficient problem solvers part I: A brief review, *Educ. Sci.* **13**, 156 (2023).
- [61] P. R. Pintrich and E. V. De Groot, Motivational and self-regulated learning components of classroom academic performance, *J. Educ. Psychol.* **82**, 33 (1990).
- [62] R. Pekrun, Control-value theory: From achievement emotion to a general theory of human emotions, *Educ. Psychol. Rev.* **36**, 83 (2024).
- [63] A. Dong, M. S.-Y. Jong, and R. B. King, How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking, *Front. Psychol.* **11**, 591203 (2020).
- [64] J. J. G. van Merriënboer and J. Sweller, Cognitive load theory and complex learning: Recent developments and future directions, *Educ. Psychol. Rev.* **17**, 147 (2005).
- [65] P. N. Iwuanyanwu, Facilitating problem solving in a university undergraduate physics classroom: The case of students' self-efficacy, *Interdiscip. J. Environ. Sci. Educ.* **18**, e2270 (2022).
- [66] T. Zenger and P. Bitzenbauer, Exploring German secondary school students' conceptual knowledge of density, *Sci. Educ. Int.* **33**, 86 (2022).
- [67] T. Duncan, P. R. Pintrich, D. A. F. Smith, and W. J. McKeachie, Motivated strategies for learning questionnaire (MSLQ) manual (2015). [10.13140/RG.2.1.2547.6968](https://doi.org/10.13140/RG.2.1.2547.6968).
- [68] M. Bieleke, K. Gogol, T. Goetz, L. Daniels, and R. Pekrun, The AEQ-S: A short version of the achievement emotions questionnaire, *Contemp. Educ. Psychol.* **65**, 101940 (2021).
- [69] M. Klepsch, F. Schmitz, and T. Seufert, Development and validation of two instruments measuring intrinsic, extraneous, and Germane cognitive load, *Front. Psychol.* **8**, 1997 (2017).
- [70] J. Woithe, Designing, Measuring and Modelling the Impact of the Hands-on Particle Physics Learning Laboratory S'Cool LAB at CERN: Effects of Student and Laboratory Characteristics on High-School Students' Cognitive and Affective Outcomes, Doctoral dissertation, Technische Universität Kaiserslautern, 2020.
- [71] M. E. Loverude, C. H. Kautz, and P. R. L. Heron, Helping students develop an understanding of Archimedes' principle. I. Research on student understanding, *Am. J. Phys.* **71**, 1178 (2003).
- [72] P. R. L. Heron, M. E. Loverude, P. S. Shaffer, and L. C. McDermott, Helping students develop an understanding of Archimedes' principle. II. Development of research-based instructional materials, *Am. J. Phys.* **71**, 1188 (2003).
- [73] I. Devetak and J. Vogrinc, The criteria for evaluating the quality of the science textbooks, in *Critical Analysis of Science Textbooks*, edited by M. S. Khine (Springer, Dordrecht, 2013), pp. 3–15.
- [74] E. Fuchs, I. Niehaus, and A. Stoletzki, *Das Schulbuch in der Forschung: Analysen und Empfehlungen für die Bildungspraxis* (V&R Unipress, Göttingen, 2014), Vol. 4.
- [75] P. Bassner, E. Frankford, and S. Krusche, Iris: An AI-driven virtual tutor for computer science education, in *LAK '24: Proceedings of the 14th International Conference on Learning Analytics & Knowledge* (ACM, 2024) p. 400.
- [76] T. Wan and Z. Chen, Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning, *Phys. Rev. Phys. Educ. Res.* **20**, 010152 (2024).
- [77] M. R. Lepper, M. F. Drake, and T. O'Donnell-Johnson, Scaffolding techniques of expert human tutors, in *Scaffolding Student Learning: Instructional Approaches and Issues*, edited by K. Hogan and M. Pressley (Brookline Books, Cambridge, MA, 1997), pp. 108–144.
- [78] S. Lemmrich, T. Ehmke, and K. Reusser, Adaptive Lernunterstützung durch fachliche Präzision und interaktionale Qualität: Ein Handlungsmodell zu adaptiver Lernunterstützung, *Praxis Forschung Lehrerinnen Bildung, Z. Schul- Professionsentwicklung (PFLB)* **6**, 2 (2024).
- [79] K. S. Taber, The use of Cronbach's alpha when developing and reporting research instruments in science education, *Res. Sci. Educ.* **48**, 1273 (2018).
- [80] J. F. Hemphill, Interpreting the magnitudes of correlation coefficients, *Am. Psychol.* **58**, 78 (2003).
- [81] K. M. Irimata and J. R. Wilson, Identifying intraclass correlations necessitating hierarchical modeling, *J. Appl. Stat.* **45**, 626 (2018).
- [82] M. H. C. Lai, and O.-M. Kwok, Examining the rule of thumb of not using multilevel modeling: The "Design Effect Smaller than Two" rule, *J. Exp. Educ.* **83**, 423 (2015).

- [83] J. Santos Nobre and J. da Motta Singer, Residual analysis for linear mixed models, *Biom. J.* **49**, 863 (2007).
- [84] J. M. Singer, F. M. Rocha, and J. S. Nobre, Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures, *Int. Stat. Rev.* **85**, 290 (2017).
- [85] G. N. Wilkinson and C. E. Rogers, Symbolic description of factorial models for analysis of variance, *J. R. Stat. Soc., Ser. C* **22**, 392 (1973).
- [86] R. R. Corbeil and S. R. Searle, Restricted maximum likelihood (REML) estimation of variance components in the mixed model, *Technometrics* **18**, 31 (1976).
- [87] S. Nakagawa, H. Schielzeth, and R. B. O'Hara, A general and simple method for obtaining R^2 from generalized linear mixed-effects models, *Methods Ecol. Evol.* **4**, 133 (2013).
- [88] E. M. Smith, M. M. Stein, C. Walsh, and N. G. Holmes, Direct measurement of the impact of teaching experimentation in physics labs, *Phys. Rev. X* **10**, 011029 (2020).
- [89] H. Abdi and L. J. Williams, Tukey's honestly significant difference (HSD) test, in *Encyclopedia of Research Design*, edited by N. J. Salkind (SAGE Publications, Inc., Thousand Oaks, CA, 2010).
- [90] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge, New York, 1988).
- [91] J. Sweller, Element interactivity and intrinsic, extraneous, and Germane cognitive load, *Educ. Psychol. Rev.* **22**, 123 (2010).
- [92] A. J. Martin, *Using Load Reduction Instruction (LRI) to Boost Motivation and Engagement* (British Psychological Society, Leicester, 2016).
- [93] P. Evans, M. Vansteenkiste, P. Parker, A. Kingsford-Smith, and S. Zhou, Cognitive load theory and its relationships with motivation: A self-determination theory perspective, *Educ. Psychol. Rev.* **36**, 7 (2024).
- [94] S. D. Shaw and G. Nave, Thinking—Fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender, [10.2139/ssrn.6097646](https://arxiv.org/abs/10.2139/ssrn.6097646) (2026).
- [95] V. Alevén, I. Roll, B. M. McLaren, and K. R. Koedinger, Help helps, but only so much: Research on help seeking with intelligent tutoring systems, *Int. J. Artif. Intell. Educ.* **26**, 205 (2016).
- [96] A. Koriát and R. A. Bjork, Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test, *Mem. Cognit.* **34**, 959 (2006).
- [97] C. Yang, R. Yu, X. Hu, L. Luo, T. S.-T. Huang, and D. R. Shanks, How to assess the contributions of processing fluency and beliefs to the formation of judgments of learning: Methods and pitfalls, *Metacognit. Learn.* **16**, 319 (2021).
- [98] B. S. Hawthorne, D. A. Vella-Brodrick, and J. Hattie, Well-being as a cognitive load reducing agent: A review of the literature, *Front. Educ.* **4**, 121 (2019).
- [99] A. J. Martin and P. Evans, Load reduction instruction: Exploring a framework that assesses explicit instruction through to independent learning, *Teach. Teach. Educ.* **73**, 203 (2018).
- [100] A. J. Martin, P. Ginns, E. C. Burns, R. Kennett, and J. Pearson, Load reduction instruction in science and students' science engagement and science achievement, *J. Educ. Psychol.* **113**, 1126 (2021).
- [101] A. M. Canonigo, Levering AI to enhance students' conceptual understanding and confidence in mathematics, *J. Comput. Assist. Learn.* **40**, 3215 (2024).
- [102] Y. Wang, X. Feng, J. Guo, S. Gong, Y. Wu, and J. Wang, Benefits of affective pedagogical agents in multimedia instruction, *Front. Psychol.* **12**, 797236 (2022).
- [103] M. Beege and S. Schneider, Emotional design of pedagogical agents: The influence of enthusiasm and model-observer similarity, *Educ. Technol. Res. Dev.* **71**, 859 (2023).
- [104] L. M. Jenö, V. Vandvik, S. Eliassen, and J.-A. Grytnes, Testing the novelty effect of an m-learning tool on internalization and achievement: A self-determination theory approach, *Comput. Educ.* **128**, 398 (2019).
- [105] T. Long, K. I. Gero, and L. B. Chilton, Not just novelty: A longitudinal study on utility and customization of an AI workflow, in *Proceedings of the 2024 ACM Designing Interactive Systems Conference, DIS '24* (Association for Computing Machinery, New York, NY, USA, 2024), pp. 782–803.
- [106] M. D. Proença, V. G. Motti, K. R. Rodrigues, and V. P. Neris, Coping with diversity: A system for end-users to customize web user interfaces, *Proceed. ACM Human-Comput. Interact.* **5**, 1 (2021).
- [107] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, Common method biases in behavioral research: A critical review of the literature and recommended remedies, *J. Appl. Psych.* **88**, 879 (2003).
- [108] E. S. Marzola, Interrogating the text: Questioning strategies designed to improve reading comprehension, *J. Reading Writing Learn. Disabil. Int.* **4**, 243 (1988).
- [109] E. Becker, J. Wünsche, J. Veith, J. Schrader, and P. Bitzenbauer, Comparing AI Chatbot, Tiered, and Textual Support in Physics Problem Solving: Effects on Cognitive Load and Affective-Motivational Outcomes, [10.5281/zenodo.20832587](https://zenodo.org/record/20832587) (2026).