International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 6, pp. 954-982, June 2026
https://doi.org/10.26803/ijlter.25.6.41
Received Apr 19, 2026; Revised Jun 5, 2026; Accepted Jun 9, 2026

From Digital Practice to Pedagogical Expertise: Does Interaction with an AI-Simulated Student Impact Preservice Physics Teachers' Skills?

Kuanysh Zhakpayev, **Yerlan Andasbayev*** , **Nazym Zhanatbekova**, **Aigerim Abdulayeva** and **Anastassiya Shendel**
Department of Physics and Mathematics,
Zhetysu University, Taldykorgan, Kazakhstan

Abstract. High-quality physics instruction depends not only on what teachers know but also on how they can translate disciplinary ideas into intelligible, responsive explanations for learners who struggle. However, student-teachers often have limited opportunities for repeated, low-stakes rehearsal of dialogic explanations that target misconceptions in real time. This study, which was conducted in Kazakhstan, examined whether sustained weekly practice with a generative artificial intelligence (GAI) chatbot built on the Llama 3.1 model, simulated struggling secondary school students, and augmented preservice physics teachers' professional development across one academic semester. A nonequivalent quasiexperimental pretest–posttest design with historical controls was implemented in an undergraduate physics teacher education program. The participants (n = 86) either received access to a purpose-built Telegram chatbot (n = 41) or completed the same semester without chatbot access (n = 45). Outcomes were assessed via an externally coded observational measure of explaining performance, a standardized test of physics-related pedagogical content knowledge (PCK), and a physics teaching self-efficacy scale. The findings indicated clear advantages for the chatbot condition in explaining performance (d = 0.90), suggesting that repeated, misconception-centered dialog with a simulated student can sharpen enacted explanatory practice. In contrast, differences favoring the chatbot group in terms of PCK (d = 0.35) and physics teaching self-efficacy (d = 0.20) were modest and not conclusive within the semester timeframe. The study contributes early quantitative evidence that scalable AI-driven student-persona simulations can bolster practice-near instructional communication while also delineating boundaries of impact for broader knowledge and belief outcomes. The results support the positioning of chatbot-based rehearsal as a targeted supplement within physics teacher education.

Citation: Zhakpayev, K., Andasbayev, Y., Zhanatbekova, N., Abdulayeva, A., & Shendel, A. (2026). From Digital Practice to Pedagogical Expertise: Does Interaction with an AI-Simulated Student Impact Preservice Physics Teachers' Skills? *International Journal of Learning, Teaching and Educational Research*, *25*(6), 954–982. https://doi.org/10.26803/ijlter.25.6.41

*Corresponding author: Yerlan Andasbayev; *andasbayev@aol.com*

Keywords: artificial intelligence; dialogic explaining; explaining skills; generative AI; large language models; pedagogical content knowledge; physics teacher education; preservice teachers; quasi-experimental design; self-efficacy; simulation-based learning

1. Introduction

Physics poses distinctive pedagogical demands in this regard: many of its core ideas are counterintuitive and entangled with robust, persistent learner preconceptions so that effective physics teaching hinges on a teacher's capacity to diagnose naive reasoning in the moment and reconstruct it through responsive, dialogic explanations. These topic-specific challenges make repeated rehearsal of misconception-focused explanations especially valuable for prospective physics teachers (Fukaya et al., 2025). Overall, the quality of science education rests, in no small part, on the caliber of its teachers. The classroom influence of teachers stands out as one of the key elements driving student learning outcomes (Xiong, 2025).

However, translating this recognition into systematic improvements in teacher education demands that researchers understand not only what prospective teachers know but also how they communicate that knowledge, how confidently they believe themselves capable of teaching it, and what pedagogical tools might sharpen all three. The present study addresses this challenge in the specific context of preservice physics teacher education, asking whether deliberate practice with an artificial intelligence (AI)-powered chatbot that simulates a struggling secondary school student can yield measurable gains in explaining skills, pedagogical content knowledge (PCK), and physics teaching self-efficacy. The following sections situate these three constructs within the broader literature, identify the gaps that motivated this investigation, and articulate the theoretical framework that guided its design.

Inspired by historical milestones such as the publication of Darwin's theory of evolution, speculative discussions about self-reproducing and even evolving machines emerged as early as the 1860s (Brandao, 2025). What began as a speculative philosophy has since crystallized into a defining technological force. Research on the application of AI in education has been a long-standing area of scholarly interest, gaining formal momentum in the late 1990s. Liu et al. (2026) situated this momentum within a broader scoping review of AI in experiential learning, underscoring the field's ambition to move beyond generalized instruction toward deeply personalized learning environments.

AI can be understood as systems that mimic human cognitive functions, such as reasoning and pattern recognition, via data-driven methods (Golec et al., 2025); critically, such systems learn from data and improve their performance over time without being explicitly programmed for each specific task (Cajueiro & Celestino, 2026). A subfield of particular contemporary relevance is generative AI (GAI), which is designed to produce novel outputs rather than merely replicating, sorting, or analyzing user inputs (Peikos & Stavrou, 2025). Through their capacity to maintain contextual continuity in extended interactions and foster inquiry-based environments, GAI agents are believed to transform education from traditional instruction-centered approaches to student-centered inquiry and discovery (Shi et al., 2026).

In recent years, research on the use of AI in teacher education has grown considerably, particularly since the release of ChatGPT in November 2022 (Aqazade et al., 2026; Molina-Soria et al., 2026; Shrestha & Yi, 2026). A typical instantiation of AI in educational settings is the GAI chatbot, which is a computer agent that processes user input and generates contextually responsive output via AI technology (Su et al., 2025). Chatbots are making a vigorous entry into education, and their prevalent application thus far is tutoring students, with empirical studies being conducted to enrich learners' knowledge through AI-mediated interactions (Calvo-Utrilla et al., 2026).

One particularly generative line of experimentation involves engaging participants in role-play scenarios with GAI-powered chatbots to rehearse occupational skills in synthetic but realistic contexts. For example, Littell and Peterson (2025) reported that students interact in real time with AI-generated chatbot "characters" acting as project stakeholders, using these conversations to inform the development of a communication management plan; participants reported more in depth, realistic interactions than with traditional methods.

Within teacher education specifically, a recent investigation by He et al. (2025) engaged 28 third-year student-teachers in tasks requiring them to mentor AI learners in mathematical problem solving and reported that AI teaching practices significantly enhanced participants' knowledge acquisition and that high-performing groups demonstrated complex mentoring patterns emphasizing procedural guidance. A quasiexperimental study by Jiang et al. (2026) investigated the use of GAI to provide feedback among preservice physics teachers; the main findings indicated that the GAI models provided feedback on instructional designs comparable in quality to human-generated feedback, effectively supporting preservice teachers' reflective practice and collaborative knowledge building.

Although these developments are encouraging, prior work has overwhelmingly cast the preservice teacher as a learner tutored by AI, as a designer who receives AI feedback on lesson plans, or as an observer in mixed-reality classrooms; comparatively little research has positioned the student-teacher as the teacher in a sustained, misconception-driven dialog with an AI student persona, and almost none has tested whether such rehearsal produces measurable gains in externally scored performance, knowledge, and belief outcomes simultaneously. This emerging line of work draws on Grossman et al.'s (2009) seminal concept of approximations of practice, which is termed by the authors "opportunities for novices to engage in practices that are more or less proximal to the practices of a profession" (ibid., p. 2058).

Recent studies have begun to realize such approximations with AI chatbots: Lee and Yeo (2022) designed a chatbot allowing student-teachers to practice assessing student thinking and responding to misconceptions by interacting with a virtual student agent; Son et al. (2024) reported that elementary preservice teachers interact with a dual-agent AI chatbot featuring both a virtual student and a virtual mentor to develop responsive teaching skills in mathematics; Lee et al. (2025) subsequently demonstrated statistically significant gains in responsive teaching skills attributable to chatbot interaction in a controlled study; and Mikeska et al. (2023) showed, in a quasiexperimental evaluation using online simulated classrooms, that repeated approximations of practice can produce statistically reliable growth in targeted instructional behaviors among preservice teachers.

However, this body of evidence remains concentrated in mathematics education and, more broadly, in general pedagogical skills; controlled, quantitative investigations of whether GAI-enabled teaching rehearsal can develop competencies specific to physics instruction, including dialogic explanations and misconception diagnoses that are central to this domain, are conspicuously absent. The present exploration targets this void by evaluating a self-designed GAI chatbot in which the preservice teacher must diagnose and remediate a simulated student's physics misconceptions and by quantifying its effects on three complementary outcomes — enacted explaining skills, standardized PCK, and physics teaching self-efficacy. In doing so, the study contributes among the first quantitative, effect-size-based evidence on GAI-enabled approximations of teaching practice in physics teacher education.

Together, the theoretical framework and the literature gaps described above motivate the following overarching research aim: to conduct a preliminary, quantitatively grounded investigation into whether access to a GAI chatbot designed to simulate a struggling secondary school student augments the professional development of preservice physics teachers over a single academic semester, as indexed by explaining skills, physics-related PCK, and physics teaching self-efficacy. Recognizing the exploratory character of this inquiry, the study adopts exclusively quantitative instrumentation; this choice prioritizes replicability, minimizes interpretive subjectivity, and enables formal effect-size estimation — features that are especially valuable when establishing an empirical baseline in a domain where controlled intervention evidence is absent. To the best of the authors' knowledge, this study is among the first to use a behavioral observational measure of explaining performance, a standardized paper–pencil PCK test, and a concept–specific self-efficacy scale within a single pretest–posttest design to examine the effects of GAI chatbot interaction on preservice physics teachers' professional competencies.

The investigation was guided by the following research questions (RQs):
RQ1. Do preservice physics teachers who engage with the GAI chatbot student persona over a semester demonstrate greater improvement in explaining skills than do historical controls who receive standard teacher education without chatbot access?
RQ2. Do preservice physics teachers in the chatbot condition show greater preto-post gains in physics-related PCK than their historical counterparts?
RQ3. Do preservice physics teachers in the chatbot condition show greater preto-post gains in physics teaching self-efficacy than their historical counterparts?

2. Literature Review

2.1 Explaining Skills in Physics Teacher Education

Explaining is among the most elemental acts of teaching. Kulgemeyer and Geelan (2024) conceptualize instructional explanations from a constructivist perspective as a communicative bridge that helps students synthesize formal classroom knowledge with their own real-world experiences to build a more coherent and personal understanding of scientific concepts. In physics specifically, the quality of a teacher's explanations has been linked to whether students develop durable, accurate conceptual models rather than surface-level procedural fluency. However, the development of this skill during initial teacher education is far from guaranteed: Kulgemeyer et al. (2020) reported that preservice physics teachers' explanatory skills did not change significantly over semester-long supervised school-based field experience—a sobering finding that calls into question the assumption that practicum exposure alone is sufficient to cultivate expert explanatory practice.

This stagnation is particularly consequential given that explaining skills is not merely a performative nicety but is empirically tied to student outcomes. If novice teachers enter the profession without having had opportunities to rehearse targeted, dialogic explanations under conditions that require them to diagnose and address misconceptions in real time, neither observation nor standard coursework may close the gap. Deliberate, low-stakes practice contexts (ideally ones that provide iterative, immediate feedback) are therefore theoretically motivated as supplements to conventional teacher education. This need is sharpened by the demands of adaptive teaching: responding effectively to physics misconceptions requires teachers to infer a learner's current reasoning and recalibrate their explanations in real time, without the luxury of extended reflection.

Without repeated, deliberately structured encounters with student misconceptions, prospective teachers frequently default to transmissive, monologic explanations and struggle to address real-world student errors when they surface in authentic classroom situations (Chen et al., 2020; Lee & Yeo, 2022). A simulated, always-available student simulacrum whose reasoning is persistently naive is therefore precisely calibrated to fill this gap, furnishing the repeated encounters with misconception-driven dialog that standard programmes rarely provide. To the best of the authors' knowledge, the present study is among the first to investigate systematically whether interactions with an AI pupil persona simulator can strengthen preservice teachers' explanatory skills, whether in physics or in other school subjects.

2.2 Pedagogical Content Knowledge in Science Teacher Education

PCK has been described as "a unique blending of content and pedagogy that enables teachers to transform their content knowledge into forms that are pedagogically powerful and yet adaptive to the variations in ability and background presented by the students" (Park & Chan, 2026, p. 1). Since Shulman (1986) first introduced the construct, its complex and "amorphic" nature has led science education scholars to make significant efforts to articulate and clarify it, resulting in various definitions and conceptual models.

Despite this conceptual proliferation, all major models share the same theoretical foundation and emphasize PCK as essential for transforming content knowledge into forms accessible to students in viable instruction (Boedeker et al., 2025; Chaitidou & Peikos, 2026). In Physics Education, the number of investigations on PCK for the teaching of physics has increased in recent years, although it remains low compared with chemistry, biology, and technology; most of this work characterizes PCK from qualitative studies, with relatively few studies measuring its development quantitatively. Teachers' professional knowledge is considered one of the most important predictors of instructional quality. A meta-analysis of mathematics and science PCK research confirmed that PCK is positively correlated with teacher content knowledge and teacher efficacy (Fukaya et al., 2025). This evidence underscores the practical stakes of interventions targeting its development.

Crucially, a recent systematic review spanning nearly four decades of teacher research synthesizing 217 empirical studies from 1986--2023 confirmed PCK's strong connections with content knowledge and teaching practices but revealed inconclusive relationships with teacher affective–motivational variables and student outcomes while identifying a heavy reliance on self-reported data as a persistent methodological limitation in the literature (Park & Chan, 2026). This overreliance on self-report instruments in PCK research has limited the field's capacity to draw firm inferences about what preservice teachers actually know and can do. Exploratory research on how GAI-mediated activities impact prospective teachers' PCK, as measured through objective performance-based instruments, is also notably scarce.

Scholars such as Jiang and Li (2026) engaged preservice teachers in an AI-saturated module with scaffolded learning-by-design tasks and collaborative lesson planning; the experimental group demonstrated significantly greater gains across technological, pedagogical, and content knowledge domains than did controls who engaged with AI within a standard educational technology course. However, this work operates knowledge through self-reports and lesson plan artifacts, leaving open the question of whether such gains are detectable with standardized, externally scored PCK instruments.

2.3 Physics Teaching Self-efficacy

Self-efficacy can be defined as an individual's conviction in their personal capabilities to motivate themselves and mobilize the requisite cognitive resources and action plans to effectively navigate a given situation (Stavropoulou, 2025). In teacher education, this general construct is applied domain-specifically: teaching self-efficacy refers to a teacher's belief in their capacity to execute the behaviors necessary to achieve the desired student learning outcomes in a particular subject area (Jerrim et al., 2025).

While nascent research has begun to explore preservice teachers' engagement with GAI-supported chatbots in virtual teaching or curriculum design scenarios, these studies have not yet rigorously examined how such bot-supported activities affect physics teaching self-efficacy. For example, Luo et al. (2025) employed a small-scale quasiexperimental design with only 26 participants and reported improvements in self-efficacy alongside content knowledge gains; however, the study's methodological constraints and mixed findings on related constructs limit the conclusiveness of its claims. This gap highlights the need for empirically robust investigations into whether practicing the act of teaching with a simulated AI student can cultivate the kind of mastery experience that Bandura's framework would predict to be most formative for self-efficacy development.

2.4 Theoretical Underpinning

The study draws on three complementary theoretical traditions. The first is Vygotsky's sociocultural theory of learning, which emphasizes learners' active role in constructing knowledge through interaction and experience (Han et al., 2026). By interacting with GAI-based chatbots, learners can engage in hands-on training, navigating synthetic scenarios that demand creative problem-solving—interactions that, when scaffolded appropriately, extend the learner's competence into regions that independent practice cannot reach. The chatbot in the present study operationalizes this principle by presenting a "student character" whose reasoning begins at a naive, intuition-based level and progressively approaches correctness as the preservice teacher supplies pedagogically substantive guidance, a structure that requires the teacher trainee to actively scaffold another learner's thinking, thereby exercising, in situ, the very diagnostic and explanatory processes that constitute PCK and explaining competence.

Specifically, this sociocultural premise maps directly onto the explanatory task examined in this study: each chatbot dialog places the teacher candidate in the more knowledgeable role, obliging them to locate the simulated student's current naive reasoning, supply a calibrated explanatory move, and reassess the student's response turn by turn. The teacher-trainee cannot simply transmit an answer but must track the simulated learner's evolving understanding and intervene at precisely the right level of abstraction, thereby mirroring, in compressed and repeatable form, the zone of proximal development dynamic that Vygotsky identified as the engine of cognitive growth.

The second theoretical pillar is Bandura's social learning theory, which posits that learning occurs through observation, imitation, and modeling (Sun & Wang, 2026). By interacting with GAI-based chatbots, learners can engage in "hands-on" training, navigating synthetic scenarios that demand empathy and creative problem solving. Moreover, relying on Bandura's conception of the self-efficacy construct, the authors suggest that repeated episodes of successful pedagogical interaction with the virtual student (essentially repeated mastery experiences) are posited to strengthen preservice teachers' belief in their capacity to explain physics concepts effectively and address student misconceptions, thereby cultivating self-efficacy (Zhang et al., 2026).

The third lens is the deliberate practice framework (Ericsson, 2008), which holds that expertise is not a straightforward product of time-on-task but of structured, feedback-rich practice directed at weaknesses (Johnson et al., 2026). Standard teacher education, as noted earlier, rarely affords student-teachers the volume of repeated, goal-directed practice episodes needed to refine explanatory skill. The GAI chatbot functions as a low-stakes rehearsal platform that is available on demand, which consistently presents fresh misconception scenarios and provides implicit feedback through the trajectory of the simulated student's understanding, approximating the conditions that deliberate practice theory identifies as necessary for skill acquisition.

3. Methodology

3.1 Research Design

A quasiexperimental pretest–posttest design with historical controls was employed, i.e., subjects were assigned to the control and treatment groups nonrandomly (Sohrabi et al., 2025) and with a year gap. The design was chosen because the central aim was to compare developmental trajectories in preservice physics teachers who had access to the GAI Chatbot with those of a comparable cohort who completed the identical program except the chatbot. No concurrent random assignment was possible; a simultaneous control condition would have created a high risk of social contamination, as all the students shared lectures, seminars, and practicum placements. The core physics curriculum and main teaching staff remained constant throughout the two academic years.

The intervention group participated during the winter semester of 2025–2026, whereas the comparison group was drawn from a cohort of student-teachers who completed identical pre- and post-measurements one academic year earlier, before the chatbot's development. This temporal separation of conditions was deliberate: concurrent randomization was not feasible within the institutional context, and recruiting a simultaneous waitlist control risked social contamination given that all student-teachers shared lecture courses, seminars, and practicum placements within the same department.

The use of historical controls, while limiting causal inference relative to a randomized controlled trial, preserved ecological validity and eliminated the ethical concern of withholding a potentially beneficial learning opportunity from a group of students enrolled in parallel. Both groups were assessed at equivalent calendar points within their respective academic years, and all measurement conditions were held constantly across cohorts.

3.2 Research participants

The participants were bachelor's-level students enrolled in the third/fourth year of a physics teacher education program at a single public university in an Asian country. Students at these stages have completed foundational coursework in both physics content and general pedagogy, affording them the conceptual vocabulary necessary to engage meaningfully with a virtual student's physics misconceptions while still being sufficiently early in their professional ontogeny to benefit from deliberate practice in explaining.

Recruitment was carried out through convenience sampling during face–to-face meetings organized by faculty staff, at which the research aims, participant rights and responsibilities, voluntary nature of participation, and anonymous character of all the data were explained. Eligible participants were required to meet two criteria: (a) active enrollment as a preservice physics teacher in the third or fourth year of study at the participating university and (b) self-declared ability and willingness to interact with the chatbot meaningfully at least once per week throughout the intervention semester. A total of 93 student-teachers provided informed consent across the two recruitment waves. Of these, 86 completed all the research activities, including both the baseline and postexperimental measurements in full; these 86 constituted the analyzed sample (intervention group, n = 41; historical control group, n = 45).

Among the 86 participants, 82 identified as female, a proportion consistent with the widely documented feminization of the teaching profession (Kammermeier et al., 2025). The experimental chatbot group comprised 23 third year and 18 fourth-year preservice teachers, whereas the historical control group consisted of 29 third year and 16 fourth-year students. The participants were aged between 19 and 23 years. A post hoc power analysis conducted in the Superpower R package (Lakens & Caldwell, 2021) confirmed that the achieved sample size was sufficient to detect effects of interest at conventional power and significance thresholds. All study procedures received approval from the university's institutional ethics committee before data collection.

3.3 Treatment

Both the experimental and control groups participated in the standard programmatic activities of the physics teacher education curriculum throughout the study period. These basic conditions included weekly subject-matter lectures covering core topics of secondary school physics, discipline-specific pedagogy seminars addressing instructional design and curriculum planning, and a structured school-based practicum in which student-teachers observed and gradually assumed teaching responsibilities under the supervision of cooperating teachers. These activities were substantively identical across the two cohorts and constituted the baseline educational experience upon which the experimental condition was layered.

The experimental group additionally received access to a researcher-designed GAI chatbot distributed via an invitation link. The chatbot simulated a secondary school student who approaches a physics teacher after a lesson, presenting an incorrect solution to a physics problem and requesting clarification. The participant, acting in the role of the teacher, was expected to guide the virtual student toward a correct understanding through a series of explanations and leading questions without simply providing the answer (a sample interaction is in Appendix 1). The chatbot was implemented as a Telegram bot, leveraging the Telegram application programming interface (API) for user interaction. The backend was built in Python via the Flask web framework and deployed as a web service on Render. The application's source code was hosted in a private repository on GitHub, and communication with Render was established via a webhook endpoint that received and processed incoming Telegram updates. For natural language generation, the chatbot relies on the Cerebras cloud infrastructure with the Llama 3.1 8B large language model (LLM).

Each conversational turn involved up to two LLM calls: one to evaluate the pedagogical quality of the teacher-trainee's input and one to generate the student character's response. The session state, including the assigned physics topic, the student's class level, the dialog history, and the number of substantively helpful teacher turn, was retained in JSON format to ensure continuity across server restarts. The Telegram was selected because it was already widely used by students in the program, accessible on both smartphones and computers, and permitted interaction outside the scheduled class time. The university's learning management system did not support the creation or execution of conversational AI bots.

The GAI chatbot simulated a struggling secondary school student across grades 7 through 11, presenting an incorrect solution to a physics problem drawn from a predefined bank of template tasks covering the secondary school physics curriculum; the student's topic and grade level were assigned randomly at the start of each session. The bot communicated entirely in Russian, the intervention participants' native language. The student character's responses were governed by a multistage scaffolding logic: on the first one to two teacher turns, the LLM was explicitly instructed to produce only colloquial, intuition-based reasoning without formulas or numerical substitution; from the third turn onward, the character could begin approximating a correct solution, with full correctness reached after four or more substantively helpful teacher interventions.

The chatbot was developed by the authors through "vibe coding", which is an AI-assisted, no-code development approach that does not require programming expertise, enabling users to generate code through intuitive, conversational prompts and iterative refinement (Chow & Ng, 2025). The code was generated and progressively refined through dialog with an AI coding assistant, with the authors directing the design logic, identifying behavioral failures, and specifying corrections. Prior to the main study, the system was piloted by the research team, who engaged the GAI chatbot in the role of teacher trainee across multiple physics topics. This pilot phase served to surface and resolve dialog scenario failures—including premature convergence to the correct answer, logically incoherent student responses, and code-switching to English mid-responses—and to verify the absence of major technical issues, including duplicate message delivery and session loss upon server restart.

The physics teacher education curriculum, course sequence, and prescribed textbooks remained unchanged across the cohorts. No major programmatic or staffing changes occurred in the intervening period. Implementation fidelity was monitored through server-side logging of session timestamps and per-user interaction counts, which were reviewed periodically to confirm that participants in the intervention group were engaging with the chatbot at the expected frequency throughout the semester.

3.4 Measures

Three instruments were employed to assess the response variables. The instruments were originally in English and were translated to Russian via forward-backward procedures. The translations were reviewed for semantic accuracy and naturalness of expression by three teacher educators external to the research team, and face validity was confirmed by consensus among these reviewers. Internal consistency was assessed via the Spearman–Brown prophecy formula on random split-half methods. A pilot study with a focus group of soon-to-be teachers (n = 7) refined the chatbot interactions and measurement procedures prior to full implementation. Participant data remained confidential throughout, with identifiers removed post analysis.

3.4.1 Explaining performance

Explaining skills was assessed via the observational instrument developed by Kulgemeyer and Riese (2018), which evaluates the quality of instructional explanations on the basis of 12 behavioral categories. Eleven categories capture appropriate explanatory strategies (e.g., diagnosing understanding), whereas one category records inappropriate behavior, defined as answering inadequately. Individual explanatory performance was quantified by summing the number of appropriate strategy categories observed for a given participant and subtracting one point if no appropriate behavior was also observed. The raw scores (two coders, agreement 87%) therefore ranged from -1, indicating the exclusive presence of no appropriate behavior, to 11, indicating the presence of all appropriate behaviors with no penalty. The split-half reliability was .77 at baseline and .81 at the final measurement.

3.4.2 PCK

Physics-related PCK was quantified via the PCK test developed by Kirschner et al. (2016). The instrument comprises 12 open-ended items (items 4–8, 10, and 12–15, including subitems) that assess two core facets of PCK: (a) knowledge of instructional strategies and representations and (b) knowledge of students' understanding of physics. Individual joint scores can range from 0 to 24 points (12 items × 0–2 points each), with higher scores indicating greater PCK. Each item was evaluated on a polytomous scale where responses could receive 0 points (incorrect), 1 point (partially correct), or 2 points (fully correct), depending on the completeness and accuracy of the answer. Two independent raters coded all the responses; the interrater agreement reached 91%. The split-half reliability was .90 for the pretest and .84 for the posttest.

3.4.3 Self-efficacy

Physics teaching self-efficacy was operationalized following this approach (Norton, 2019), in which participants rated their perceived capacity to teach the concept addressed in each of the 10 PCK items immediately after completing that item. Ratings were made on a five-point scale anchored at 1 (very low self-efficacy: the respondent feels that they would need to do a great deal of research before attempting to teach the concept) and 5 (very high self-efficacy: the respondent feels that they could teach questions of that structure with very little or no preparation). The remaining scale points represented gradations of confidence intermediate between these poles. Individual self-efficacy scores were computed as the mean rating across the 10 items, yielding a continuous score between 1 and 5. The split-half reliability was .88 at the first assessment and .93 postintervention.

3.5 Data collection

Data collection was administered in a paper-and-pencil format in a university lecture theater under standardized conditions, with the exception of the explanatory performance assessment, which required a separate procedural protocol. Both the baseline and post experimental assessments were conducted by research assistants who were blinded to the participants' group membership and to the phase of testing, thereby minimizing assessor bias. Pretesting took place over two weeks immediately preceding the start of the intervention semester, and post testing was administered over two weeks immediately following its conclusion, with the two assessment windows separated by approximately one academic year across the two cohorts. Pre- and posttests were administered under identical conditions with respect to venue, time of day, instructions provided, and time allowed for completion.

Explaining performance was assessed via a structured two-phase protocol based on the dialogic explanatory assessment framework (Kulgemeyer & Riese, 2018). In the first phase, the participants were given ten minutes to prepare an explanation of a standard school physics phenomenon assigned to them at the session. During preparation, the participants were instructed to tailor their explanations to a hypothetical student with no prior knowledge of the topic and were provided with standardized reference materials, which they were free to consult alongside their own notes. In the second phase, each participant delivered a ten-minute explanation in a one-on-one format to an explainee roleplayed by one of the undergraduate research assistants who were nonphysics majors and were entirely blind to the study's objectives and context.

To ensure standardized elicitation conditions across all assessments and all participants, the explainees were trained to deliver a fixed set of prompts at designated points in the interaction; these prompts were designed to elicit specific explanatory behaviors, such as requesting a simpler explanation, asking for a sketch, or prompting the use of appropriate physics terminology. All the sessions explained were video recorded. Two trained research assistants independently coded each recorded session according to the 12 behavioral categories of the original instrument to derive individual explanatory performance scores.

3.6 Data analysis

All the statistical analyses were conducted via the R programming environment (version 4.5.3). The alpha level for determining statistical significance was established at a priori at .05 for all omnibus tests and diagnostic procedures. Prior to conducting the primary inferential analyses, preliminary linear mixed-effects models (LMMs) were fitted for each dependent variable to evaluate the underlying statistical assumption. These preliminary models incorporated fixed effects for group, time, the group-by-time interaction, and the study year as covariates to control for baseline educational differences, alongside a random intercept for each participant to account for the nested nature of the repeated-measures data.

LMMs were preferred over nonparametric alternatives because nonparametric alternatives cannot accommodate the nested repeated-measures structure of the data simultaneously, include a continuous covariate for the study year, and provide a direct inferential test of the group-by-time interaction that is required to answer the three research questions. Moreover, LMMs are known to be remarkably robust to moderate departures from residual normality, as demonstrated by large-scale simulation evidence showing that fixed-effect estimates remain unbiased even when distributional assumptions are objectively violated (Schielzeth et al., 2020).

The normality of the model residuals was tested via the Shapiro–Wilk test. The assumption of homoscedasticity was assessed by submitting the absolute residuals to an analysis of variance to check for equal variances across the group-by-time design cells. Additionally, the presence of excessively influential observations was evaluated via Cook's distance, with a predetermined flagging threshold of 4/n; the maximum acceptable proportion of such influential cases was set at 5% of the total sample.

The diagnostic evaluations indicated that the proportion of influential observations remained within the acceptable 5% limit across all domains, precluding the need for outlier removal. However, the diagnostic tests revealed that the assumption of homoscedasticity was significantly violated across all three outcome variables. Furthermore, the assumption of normally distributed residuals was violated for the explanatory skills measure. To robustly accommodate these widespread violations of equal variances without sacrificing statistical power or ignoring the hierarchical data structure, an adaptive modeling strategy was implemented.

Standard LMMs were eschewed in favor of heteroscedastic LMMs, which were constructed via the *nlme* package. These heteroscedastic models retained the original fixed and random effects structures but were augmented with a variance weighting function that explicitly allowed for distinct residual variance estimates between the chatbot and control groups. The primary inferential focus was directed at the group-by-time interaction term, which captured the differential developmental trajectories of the two cohorts. To quantify the magnitude of the intervention, estimated marginal means were extracted from the fitted heteroscedastic models via the *emmeans* package.

A difference-in-differences (DiD) effect size, expressed as Cohen's d, was subsequently computed by calculating the difference in the preto-post marginal mean gains between the treatment and comparison groups, standardized by the pooled baseline standard deviation. These effect sizes were interpreted via conventional thresholds, where values approaching 0.20, 0.50, and 0.80 were designated as representing small, moderate, and large effects, respectively.

4. Results

The descriptive statistics are listed in Table 1. An examination of the descriptive statistics revealed a pronounced developmental trajectory for the preservice teachers interacting with the GAI chatbot, whereas the historical control cohort exhibited relative stagnation (Figure 1).

Table 1: Pre- to posttest descriptive statistics by group

Outcome	Group	Pretest, mean (SD)	Posttest, mean (SD)
Explaining skills	Control	5.20 (1.16)	5.53 (0.89)
	Experimental	4.61 (1.14)	7.34 (1.04)
Pedagogical content knowledge	Control	9.11 (2.86)	13.69 (3.75)
	Experimental	11.46 (3.80)	17.22 (3.92)
Teaching self-efficacy	Control	3.26 (0.56)	3.51 (0.55)
	Experimental	3.77 (0.40)	3.92 (0.49)

The subsequent mixed-model analysis, detailed in Table 2, confirmed this observation. The interaction effect between group and time was statistically significant (p = .001), with a large effect size.

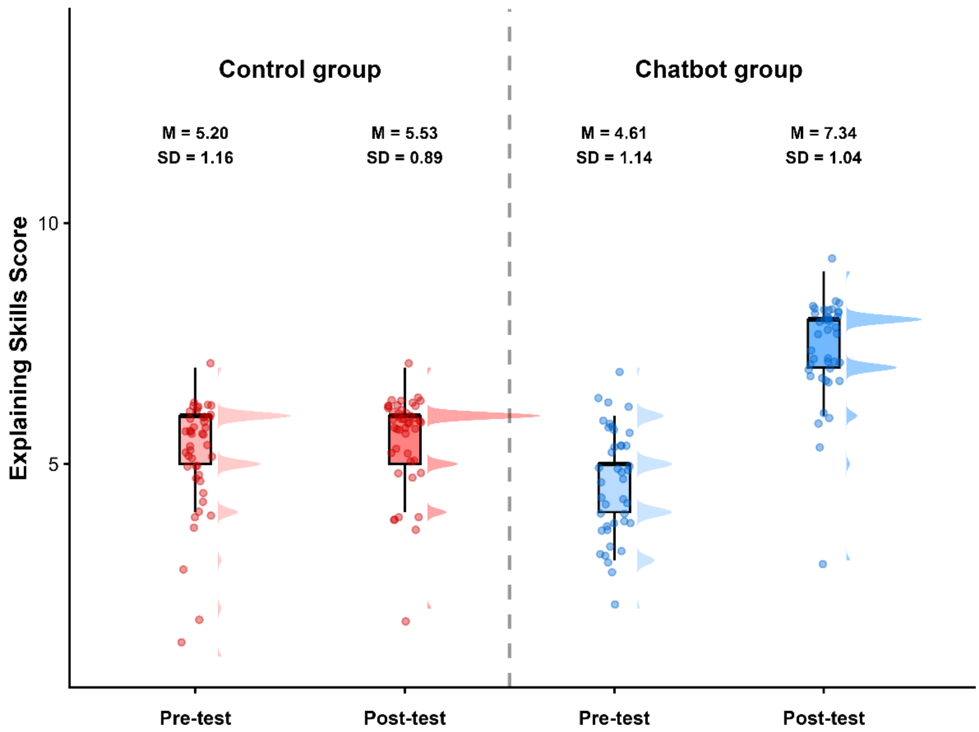


Figure 1: Preto-postdevelopmental trajectory of explanatory skills

Table 2: Mixed-model analysis of variance for explaining skills, physics-related PCK, and teaching self-efficacy

Variable	F	df_1	df_2	p	Cohen's d
Explaining skills	115.66	1	84	.001	0.90
Pedagogical content knowledge	2.29	1	84	.134	0.35
Teaching self-efficacy	0.42	1	84	.516	0.20

Turning to physics-related PCK, the descriptive patterns similarly indicated substantial cognitive gains in experimental condition compared with a noticeably more modest baseline drift in the control group (Figure 2). The inferential model yielded a nonsignificant interaction effect ($p = .134$) of a small magnitude.

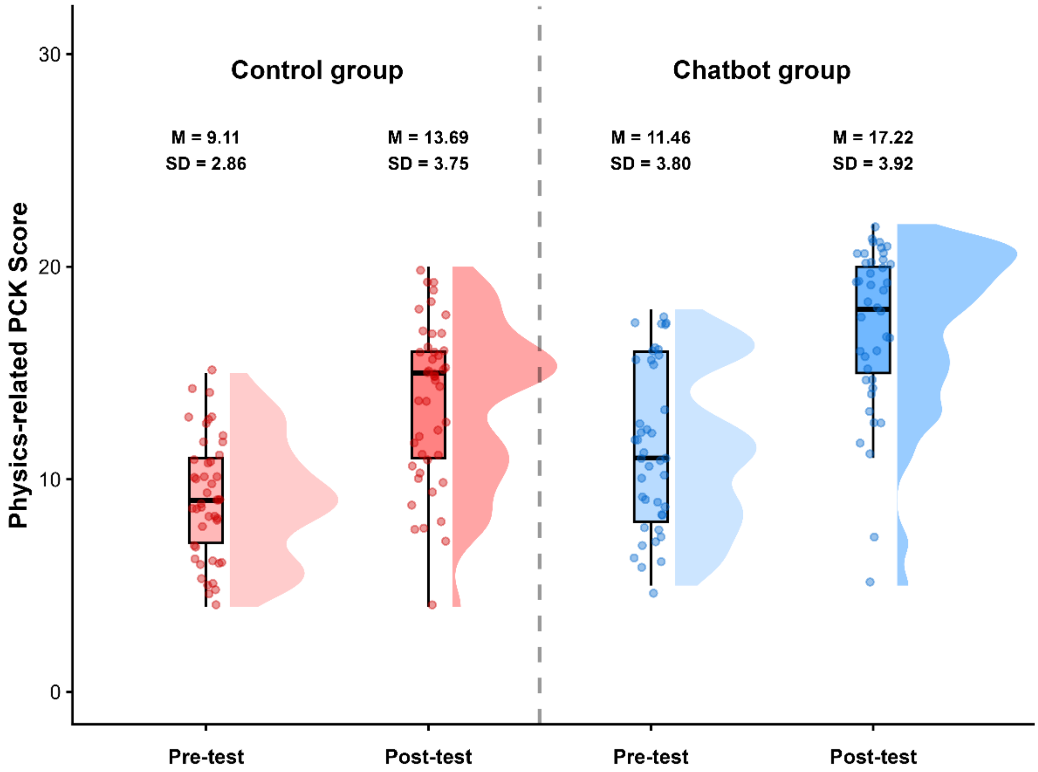


Figure 2: Preto-post developmental trajectory of physics-related pedagogical content knowledge

Finally, concerning physics teaching self-efficacy, the visual distributions highlighted an enhancement in the chatbot group juxtaposed against a parallel upward trend in self-declared confidence among the historical controls (Figure 3). The statistical evaluation assessed this divergence, yielding an interaction of $p = .516$, with a small effect size.

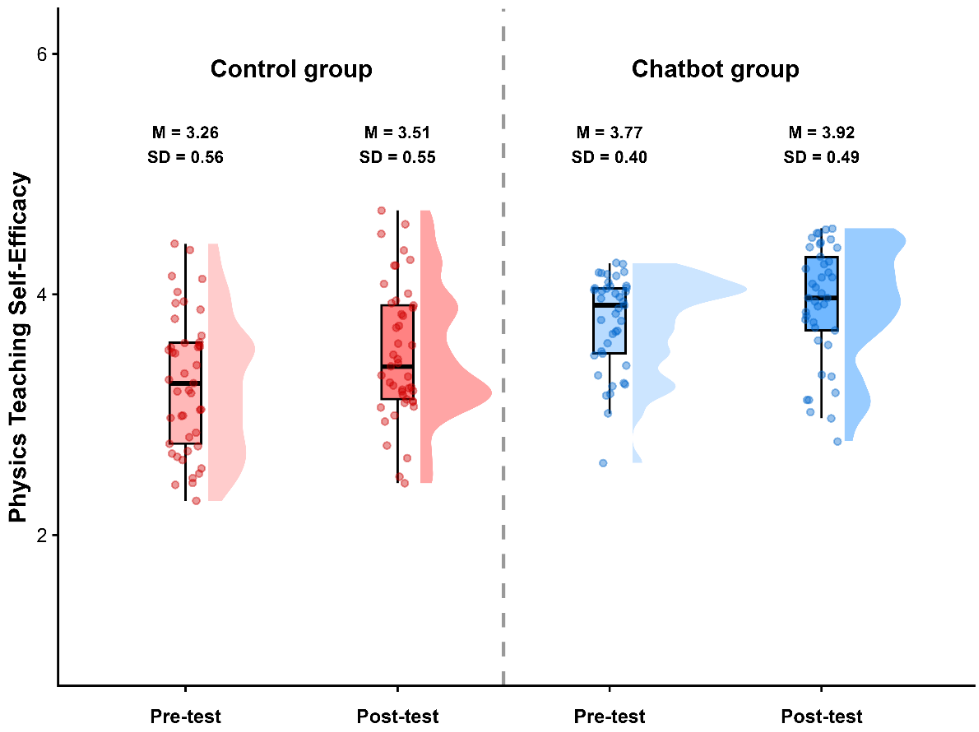


Figure 3: Preto-post developmental trajectory of physics teaching self-efficacy

5. Discussion

This study examined whether structured, semester-long rehearsal with a GAI chatbot designed to role-play a struggling secondary physics student can strengthen the core competencies of preservice physics teachers beyond what is typically observed in standard program experiences. The intervention was intentionally positioned as practice in "being the teacher" during a misconception-driven dialog rather than as GAI-supported study, lesson planning, or feedback on artifacts.

In line with deliberate-practice framing, the chatbot created a repeatable setting in which preservice teachers repeatedly had to diagnose a learner's reasoning, select representations, and adjust explanations turn by turn. The study further sought to quantify potential effects via three complementary outcomes that map onto professional performance, professional knowledge, and professional beliefs. In doing so, the design aimed to provide an early quantitative baseline for a form of GAI-enabled approximation of teaching practice that is relatively easy to scale compared with human-led simulations.

In brief, semester-long rehearsal with the GAI student-persona chatbot produced a large and statistically discernible advantage in enacted explanatory skills ($d = 0.90$, $p = .001$) but only small and nonsignificant advantages in physics-related PCK ($d = 0.35$, $p = .134$) and physics teaching self-efficacy ($d = 0.20$, $p = .516$). This outcome pattern is in line with a nascent body of practice-based and simulation-based teacher education research showing that dialog-coupled instructional behaviors are rather responsive to repeated approximations of practice (Kavanagh et al., 2020; Lee et al., 2025; Mikeska et al., 2023; Sapkota & Max, 2024).

5.1 Explaining Skills

The intervention cohort improved substantially more than the historical control cohort did in terms of the observational explanatory measure. Interpreted in practical terms, access to a simulated student who persistently presented incorrect reasoning appears to have strengthened participants' enacted explanatory repertoire, not merely their declarative talk about explaining. This pattern aligns with practice-based teacher education arguments that improvements in complex teaching moves typically require repeated approximations of practice rather than exposure alone (Cohen et al., 2026). It also aligns with simulation-based teacher education findings that structured rehearsals can strengthen targeted instructional behaviors, especially when the rehearsal environment involves repeated attempts without high interpersonal cost or classroom risk (Alkan et al., 2026).

Several mechanisms plausibly account for the magnitude of the explained gain. One mechanism is that the chatbot interaction repeatedly prompts generative "learning-by-explaining," which tends to strengthen understanding and the ability to express causal and mechanistic relationships because explanations require selecting, organizing, and integrating ideas into a coherent model; generative learning theory predicts that these overt constructive activities should yield performance benefits beyond passive exposure. A second mechanism is conversational contingency (Carolus et al., 2024): because the simulated student persona responds turn by turn, the teacher candidate is repeatedly cued to monitor comprehension signals and adjust explanations, which resembles the cognitive demands of adaptive teaching and may accelerate the acquisition of flexible explanation routines.

A third mechanism is retrieval-based strengthening: each session forces repeated retrieval and re-expression of physics concepts and instructional moves across various topics, which could add to the retention and application of complex ideas; recent evidence indicates that retrieval practice benefits not only memory but also the application of complex educational concepts (Corral & Carpenter, 2025), whereas other work shows that retrieval practice can enhance new learning even when it does not generalize to all downstream task types (Pastötter et al., 2022). Taken together, these mechanisms suggest that the intervention's strongest leverage point may be the sheer density of prompted explanatory retrieval and adaptive reformulation rather than any single feature of the platform.

5.2 Physics-related PCK

The chatbot group showed a positive but nonsignificant advantage on the standardized PCK test, so the evidence does not support a firm claim of PCK improvement attributable to chatbot access under the present design. A conservative interpretation is that the intervention may have provided some PCK-relevant learning opportunities, but those opportunities either did not accumulate strongly enough over one semester or were not well captured by the particular standardized PCK instrument used. This pattern is plausible because PCK is often described as knowledge that is highly topic- and context-specific (Boedeker et al., 2025; Hartmuth et al., 2025; Yip, 2025), and the chatbot's randomized assignment of topics across sessions may involve dispersed learning across a wider set of concepts than those sampled by the test. Additionally, improvements in enacted explanations may not automatically translate into gains on open-ended written PCK items, which require a different representation of knowledge than interactive dialog does.

Compared with other chatbot-based teacher education interventions, the direction of the effect is compatible with findings that GAI chatbots can support preservice teachers' pedagogical thinking when embedded in structured instructional designs. For example, a quasiexperimental "chatbot-based 5E learning approach" reported benefits for preservice teachers' theoretical knowledge and instructional design skills, suggesting that chatbot interaction can support pedagogically relevant knowledge building when tasks are tightly scaffolded and aligned with targeted constructs (Liu et al., 2024). Additionally, recent work in mathematics teacher education has used a GAI chatbot functioning as both a virtual student and mentor to examine responsive teaching practices, highlighting that chatbot architectures that combine simulated learners with pedagogical coaching may be particularly potent for knowledge growth that resembles PCK (Son et al., 2024).

5.3 Physics Teaching Self-efficacy

Teaching self-efficacy increased slightly more in the chatbot group than in the historical controls, but the interaction was not statistically significant, and the estimated effect was small. Interpreted conservatively, the results do not support the claim that chatbot-based teaching rehearsal meaningfully outperformed the ordinary semester experience in shifting self-efficacy, at least as measured by concept-specific ratings attached to the PCK items. The parallel upward trend in both cohorts suggests that typical program experiences, including practicum exposure, may already provide salient confidence cues over a semester, potentially leaving limited additional room for a GAI chatbot to produce a distinct advantage. Another possibility is that the intervention created mixed experiences: some sessions may have felt successful, whereas others may have highlighted uncertainty, and produce net modest change when aggregate.

In addition, the self-efficacy measure was tightly bound to concept-specific judgments tied to the PCK items, which may have made it especially sensitive to test-related feelings of competence rather than to a more stable sense of instructional capability in authentic classroom interaction. This finding accords with the evidence from large-scale observational work, which has suggested that self-efficacy and instruction do not invariably move in lockstep, implying that performance gains can occur without proportionate shifts in self-efficacy scores (Jerrim et al., 2025). Accordingly, the present results are interpretable as a case where enacted explanations improved substantially, whereas confidence metrics did not clearly diverge from the normal developmental trajectory observed in a standard program semester.

Finally, the chatbot likely provided mastery-like episodes for some participants, but it also may have delivered ambiguous cues because "student improvement" depended partly on model behavior rather than only on participant competence.

In such settings, efficacy information may be noisier than in supervised practicum contexts where feedback from mentors and real student responses may be perceived as more diagnostic. Overall, the outcome pattern resembles a "skill-first" profile: the chatbot may be especially effective at strengthening interactive discourse moves that can be rehearsed at high frequency, whereas deeper professional knowledge and beliefs may require additional scaffolds or longer exposure.

This study makes several contributions that are unusually concrete for the fast-moving "GenAI in teacher education" literature. Empirically, this study provides quantitative evidence that a scalable chatbot designed as a struggling student persona can meaningfully improve preservice physics teachers' enacted explanatory skills over a semester, as measured by a behavioral observational coding instrument rather than self-reports alone. Methodologically, it adds an example of combining an observational performance measure, a standardized open-ended PCK assessment scored by independent raters, and a concept-specific efficacy measure in a single coherent pretest–posttest intervention evaluation, enabling differentiated inferences about performance, knowledge, and beliefs. Conceptually, it advances the design space of AI in teacher education by emphasizing AI-enabled approximations of practice in which preservice teachers repeatedly practice diagnosing and responding to misconceptions in real time.

For practice, the results imply that teacher education programs that struggle to provide sufficient opportunities for repeated, individualized teaching rehearsals may be able to supplement coursework with chatbot-based simulations specifically for explanation and dialogic responsiveness. This approach is potentially more scalable than mixed-reality environments that require specific facilities and scheduling, although mixed-reality research has demonstrated complementary strengths for approximating classroom complexity (Dalinger et al., 2020). Relatedly, some explorations have begun integrating generative AI with virtual reality simulations, but these studies often rely heavily on qualitative evidence, which makes direct comparisons of effect sizes and measurement sensitivity difficult (Hong et al., 2025; Zhang et al., 2024; Zhang et al., 2026).

The current study's quantitative effect on an externally coded explanatory outcome therefore strengthens the empirical case that GAI-mediated role play can move beyond engagement and perceived usefulness toward measurable improvement in enacted instructional communication. The present work also adds to the emerging evidence base that chatbot interventions can be constructed not only as tutors for students but also as practice partners for future teachers, with measurable effects when the target competency is closely coupled with dialog.

5.4 Limitations

Several limitations should temper interpretation while still allowing the results to be informative. First, the comparison using historical controls constrains causal certainty and leaves open the possibility that cohort differences contributed to some observed changes, even when measurement windows and procedures were standardized. Moreover, the intervention's effectiveness depends on actual engagement with the GAI chatbot across the semester; variability in usage intensity, timing, or severity of participation could have produced heterogeneous "dosage," which would blur relationships between the intervention and outcomes in ways not captured by group assignment.

Measurement and inference limitations also remain. The explanatory measure relies on human coding, and while interrater agreement is reported, agreement metrics can be misleading if the construct is complex and multifaceted (Conry-Murray et al., 2024). As mentioned above, the self-efficacy measure was tightly linked to the PCK items and administered immediately after each item; this coupling may have heightened context-dependence and test-related affect, making efficacy scores partly a reflection of perceived test success rather than a stable sense of competence for teaching in authentic classroom contexts.

Additionally, the intervention's chatbot logic intentionally advanced the student toward correctness after "helpful" turns. Since generative models can be inconsistent, participant experiences may vary in terms of perceived realism and contingency, potentially dampening effects on knowledge and beliefs even if the explanation practice frequency is high. Finally, the sample was highly gender-imbalanced (82 of 86 participants identified as female). Although this distribution reflects the participating program, it limits the generalizability of the findings to more gender-balanced populations of preservice physics teachers.

5.5 Implications and Suggestions

For practice, the results support embedding chatbot-based student persona rehearsals as a structured supplement aimed at strengthening explanatory communication and responsiveness to misconceptions. Given evidence that simulation learning benefits depend strongly on sequencing and content alignment (Lebert & Vilarroya, 2025), to increase the likelihood of measurable PCK gains, teacher education programs may consider pairing the chatbot practice with explicit decomposition activities, such as structured reflection prompts that require preservice teachers to name the misconception, justify the representation used, and articulate an instructional next step; research on feedback in teacher learning contexts (Asregid et al., 2023) suggests that structured feedback processes can strengthen reflective practice and potentially deepen learning from rehearsal episodes.

For self-efficacy, the design implication is to make mastery cues more interpretable. For example, adding stable performance dashboards, exemplars, and calibrated feedback may help participants attribute student improvement to specific instructional domains rather than opaque model behavior. Faculty adoption strategy also matters. A realistic pathway for institutional uptake is to treat chatbot simulations as staged innovations rather than all-or-nothing changes. In particular, teacher education faculties could frame implementation through the four levels of GenAI adoption described by Kurtz et al. (2024), beginning with exploratory use, moving toward routine utility, then toward copilot integration for instructional decision support, and finally toward transformative redesign in which AI-enabled approximations of practice are systematically integrated into methods coursework and practice support. In early stages, the safest practice is likely constrained, well-scaffolded chatbot rehearsal tasks that target specific talk moves and explanation routines; later stages could integrate analytics and supervision structures.

Future research should pursue designs that can sharply isolate mechanisms and optimize outcomes. One promising line of research is comparative experimentation on chatbot architectures: a student-only persona versus a dual-agent design that includes a virtual mentor, as dual-agent designs may better support the explicit pedagogical reasoning associated with PCK growth. Another line of alignment research involves mapping chatbot scenario banks directly onto the blueprint of the PCK test to ensure content overlaps and then testing whether that alignment increases the sensitivity of the PCK outcome.

6. Conclusion

This study addresses the problem that preservice physics teachers often have insufficient opportunities to exercise responsive, misconception-centered explanations before encountering the full complexity of real classrooms. To investigate whether the GAI can help bridge this gap, a semester-long intervention was implemented in which preservice physics teachers repeatedly interacted with a purpose-built GAI chatbot simulating a struggling secondary school physics student. The strongest finding was a large, statistically significant improvement in explaining performance for the chatbot group relative to historical controls. Conversely, the advantages of PCK and self-efficacy were small and statistically nonsignificant.

The results therefore indicate a differentiated pattern of impact: repeated chatbot-mediated teaching rehearsal appears especially effective for strengthening practice-near communicative skills, whereas benefits to broader professional knowledge and confidence are less certain within a single term. The design limitations inherent in historical cohort comparisons, together with measurement constraints around PCK and efficacy, further limit claims about causality and breadth of impact. The central takeaway for teacher education practice is that GAI chatbot simulations offer a scalable, low-cost supplement for rehearsing dialogic explanations and responsiveness to misconceptions and should be embedded as a targeted, well-scaffolded complement to field-based learning rather than a replacement for it.

For researchers, the “skills-first” profile observed here suggests that broader knowledge and belief outcomes require tighter alignment between simulation content and target constructs, enhanced architectures that combine student simulation with explicit reflection prompts or virtual mentor feedback, and longer exposure. Randomized, mechanism-focused designs with more diverse samples are needed to establish causal warrants and to calibrate the conditions under which GAI-enabled approximations of teaching practice transfer to a wider range of professional competencies.

AI Use Declaration
Grammarly was used to check the language and grammar of the manuscript, followed by human proofreading.

7. References

Alkan, Ö. K., Kulaksız, T., Aksoy, N. C., Durmaz, B. N., Uçar, F. M., Özcan, M., & Kalkavan, B. (2026). Simulation-based learning approach in teaching practicum to support pre-service teachers' teaching skills. *Education and Information Technologies*. Advance online publication. https://doi.org/10.1007/s10639-026-13951-z

Aqazade, M., Mauntel, M., & Atabas, S. (2026). Empowering mathematics teacher educators: Exploring artificial intelligence-driven mathematical tasks. *School Science and Mathematics, 126*(1), 24–43. https://doi.org/10.1111/ssm.18339

Asregid, D., Mihiretie, D. M., & Kassa, S. A. (2023). Teacher educators use of feedback to facilitate reflective practice among pre-service teachers during microteaching. *Cogent Education, 10*(2). https://doi.org/10.1080/2331186X.2023.2257121

Boedeker, P. J., Moreno, N. P., Newell, A. D., Maheshwari, A., McMillin, M. A., Tippen, S. P., Pillow, M. T., & Bergemann, A. D. (2025). Pedagogical content knowledge as a lens for transforming teaching in medical and health professions education. *Medical Science Educator, 35*(3), 1707–1714. https://doi.org/10.1007/s40670-025-02325-8

Brandao, P. R. (2025). The impact of artificial intelligence on modern society. *AI, 6*(8), 190. https://doi.org/10.3390/ai6080190

Cajueiro, D. O., & Celestino, V. R. R. (2026). A comprehensive review of artificial intelligence regulation: Weighing ethical principles and innovation. *Journal of Economy and Technology, 4*, 77–91. https://doi.org/10.1016/j.ject.2025.07.001

Calvo-Utrilla, M., Paños, E., & Ruíz-Gallardo, J. (2026). Chatbots in science education: A scoping review of early empirical evidence. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-025-10260-x

Carolus, A. E., McLaughlin, K. A., Lengua, L. J., Rowe, M. L., Sheridan, M. A., Zalewski, M., Moran, L., & Romeo, R. R. (2024). Conversation disruptions in early childhood predict executive functioning development: A longitudinal study. *Developmental Science, 27*(1), e13414. https://doi.org/10.1111/desc.13414

Chaitidou, M., & Peikos, G. (2026). Pedagogical content knowledge in science education. *Encyclopedia, 6*(2), 43. https://doi.org/10.3390/encyclopedia6020043

Chen, C., Sonnert, G., Sadler, P. M., & Sunbury, S. (2020). The impact of high school life science teachers' subject matter knowledge and knowledge of student misconceptions on students' learning. *CBE—Life Sciences Education, 19*(1), 9. https://doi.org/10.1187/cbe.19-08-0164

Chow, M., & Ng, O. (2025). From technology adopters to creators: Leveraging AI-assisted vibe coding to transform clinical teaching and learning. *Medical Teacher, 47*(12), 1927–1929. https://doi.org/10.1080/0142159X.2025.2488353

Cohen, J., Marisa Yonas, A., & Elizabeth Wilson, K. (2026). Approximating teaching: A systematic review of the research. *Review of Educational Research*. Advance online publication. https://doi.org/10.3102/00346543251368383

Conry-Murray, C., Waltzer, T., DeBernardi, F. C., Fossum, J. L., Haasova, S., Matthews, M. S., O'Mahony, A., Moreau, D., Baum, M. A., Karhulahti, V., McCarthy, R. J., Paterson, H. M., McSweeney, K., & Elsherif, M. M. (2024). Validity and transparency in quantifying open-ended data. *Advances in Methods and Practices in Psychological Science, 7*(4). https://doi.org/10.1177/25152459241275217

Corral, D., & Carpenter, S. K. (2025). Effects of retrieval practice on retention and application of complex educational concepts. *Learning and Instruction, 100*, 102219. https://doi.org/10.1016/j.learninstruc.2025.102219

Dalinger, T., Thomas, K. B., Stansberry, S., & Xiu, Y. (2020). A mixed reality simulation offers strategic practice for preservice teachers. *Computers and Education, 144*, 103696. https://doi.org/10.1016/j.compedu.2019.103696

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine, 15(*11), 988–994. https://doi.org/10.1111/j.1553-2712.2008.00227.x

Fukaya, T., Nakamura, D., Kitayama, Y., & Nakagoshi, T. (2025). A systematic review and meta-analysis of research on mathematics and science pedagogical content knowledge: Exploring its associations with teacher and student variables. *Teaching and Teacher Education, 155*, 104881. https://doi.org/10.1016/j.tate.2024.104881

Golec, M., Hatay, E. S., Gill, S. S., & Buyya, R. (2025). Artificial intelligence (AI): Foundations, trends and future directions. *Telematics and Informatics Reports, 20*, 100265. https://doi.org/10.1016/j.teler.2025.100265

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record: The Voice of Scholarship in Education, 111*(9), 2055-2100. https://doi.org/10.1177/016146810911100905

Han, C., Lin, C., Ku, H., & Huang, Y. (2026). An integrative educational framework to enhance logical reasoning and cognitive skills in undergraduate nursing students: A constructivist grounded theory approach. *Nurse Education Today, 160*, 107004. https://doi.org/10.1016/j.nedt.2026.107004

Hartmuth, D., Amin, R. M., Schlichting, D., Schött, S., Haudenschild, J., Küchenhoff, H., & Neuhaus, B. J. (2025). Developing science teachers' enacted pedagogical content knowledge through integration of student feedback into the Refined Consensus Model of pedagogical content knowledge. *Cogent Education, 12*(1), 2470563. https://doi.org/10.1080/2331186X.2025.2470563

He, X., Zhou, R., Fan, Q., Xiao, X., Yu, Y., & Yan, Z. (2025). Preparing student teachers for professional development: Mentoring generative artificial intelligence (AI) learners in mathematical problem solving. *IEEE Transactions on Learning Technologies, 18*, 458–469. https://doi.org/10.1109/tlt.2025.3557037

Hong, S., Moon, J., Eom, T., Awoyemi, I. D., & Hwang, J. (2025). Generative AI-enhanced virtual reality simulation for pre-service teacher education: A mixed-methods analysis of usability and instructional utility for course integration. *Education Sciences, 15*(8), 997. https://doi.org/10.3390/educsci15080997

Jerrim, J., Prieto-Latorre, C., Marcenaro-Gutierrez, O. D., & Shure, N. (2025). Teacher self-efficacy, instructional practice, and student outcomes: Evidence from the talis video study. *American Educational Research Journal, 62*(2), 378-413. https://doi.org/10.3102/00028312241300265

Jiang, S., & Li, J. (2026). Triadic instructional design: The impact of structured AI training on pre-service teachers' intelligent-TPACK, attitudes, and lesson planning skills. *Education Sciences, 16*(2), 315. https://doi.org/10.3390/educsci16020315

Jiang, Y., Feng, X., Liu, Y., Wang, Y., Xie, L., & Bao, L. (2026). Generative AI for feedback and collaborative knowledge construction in preservice physics teacher education. *Physical Review Physics Education Research, 22*, 010116. https://doi.org/10.1103/hm13-jv98

Johnson, M., Maitland, E., & Sofka, W. (2025). Developing judgement for business: An AI-based model of independent management learning. *Journal of Business Research, 204*, 115842. https://doi.org/10.1016/j.jbusres.2025.115842

Kammermeier, M., Muckenthaler, M., Weiß, S., & Kiel, E. (2025). Feminization of teaching: Gender and motivational factors of choosing teaching as a career. *Frontiers in Education, 10*, 1471015. https://doi.org/10.3389/feduc.2025.1471015

Kavanagh, S. S., Metz, M., Hauser, M., Fogo, B., Taylor, M. W., & Carlson, J. (2020). Practicing responsiveness: Using approximations of teaching to develop teachers' responsiveness to students' ideas. *Journal of Teacher Education, 71*(1), 94-107. https://doi.org/10.1177/0022487119841884

Kirschner, S., Borowski, A., Fischer, H. E., Gess-Newsome, J., & von Aufschnaiter, C. (2016). Developing and evaluating a paper-and-pencil test to assess components of physics teachers' pedagogical content knowledge. *International Journal of Science Education, 38*(8), 1343–1372. https://doi.org/10.1080/09500693.2016.1190479

Kulgemeyer, C., & Geelan, D. (2024). Towards a constructivist view of instructional explanations as a core practice of science teachers. *Science Education, 108*, 1034–1050. https://doi.org/10.1002/sce.21863

Kulgemeyer, C., & Riese, J. (2018). From professional knowledge to professional performance: The impact of CK and PCK on teaching quality in explaining situations. *Journal of Research in Science Teaching, 55*(10), 1393–1418. https://doi.org/10.1002/tea.21457

Kulgemeyer, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Reinhold, P., Riese, J., Schecker, H., Schröder, J., & Vogelsang, C. (2020). Professional knowledge affects action-related skills: The development of preservice physics teachers' explaining skills during a field experience. *Journal of Research in Science Teaching, 57*(10), 1554–1582. https://doi.org/10.1002/tea.21632

Kurtz, G., Amzalag, M., Shaked, N., Zaguri, Y., Kohen-Vacs, D., Gal, E., Zailer, G., & Barak-Medina, E. (2024). Strategies for integrating generative AI into higher education: navigating challenges and leveraging opportunities. *Education Sciences, 14*(5), 503. https://doi.org/10.3390/educsci14050503

Lakens, D., & Caldwell, A. R. (2021). Simulation-Based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920951503

Lee, D., & Yeo, S. (2022). Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Computers and Education, 191*, 104646. https://doi.org/10.1016/j.compedu.2022.104646

Lee, D., Son, T., & Yeo, S. (2025). Impacts of interacting with an AI chatbot on preservice teachers' responsive teaching skills in math education. Journal of Computer Assisted Learning, 41(1), e13091. https://doi.org/10.1111/jcal.13091

Littell, W. N., & Peterson, B. L. (2025). AI-powered chatbots to simulate executive interactions for students performing stakeholder analysis. *Communication Teacher, 39*(1), 18–25. https://doi.org/10.1080/17404622.2024.2405188

Liu, C. C., Wang, H. J., Wang, D., Tu, Y. F., Hwang, G. J., & Wang, Y. (2024). An interactive technological solution to foster preservice teachers' theoretical knowledge and instructional design skills: A chatbot-based 5E learning approach. *Interactive Learning Environments, 32*(10), 6698–6716. https://doi.org/10.1080/10494820.2023.2277761

Liu, Z., Cheon, S., Stanbury, A., Jiao, X., Xing, W., & Kang, H. (2026). Towards contextual-based AI: A scoping review of artificial intelligence in X reality for personalized learning. *Computers and Education Artificial Intelligence, 10*, 100523. https://doi.org/10.1016/j.caeai.2025.100523

Luo, J., Zhu, C., Hu, L., & Sun, M. (2025). Empowering preservice teachers through textbook design activities with GAI-based chatbot. *IEEE Transactions on Learning Technologies, 18*, 822–832. https://doi.org/10.1109/tlt.2025.3606757

Mikeska, J. N., Howell, H., & Kinsey, D. (2023). Do simulated teaching experiences impact elementary preservice teachers' ability to facilitate argumentation-focused discussions in mathematics and science? *Journal of Teacher Education, 74*(5), 422-436. https://doi.org/10.1177/00224871221142842

Molina-Soria, M., Aparicio-Herguedas, J. L., Fuentes-Nieto, T., & López-Pastor, V. M. (2026). The combination of artificial intelligence and formative assessment in teacher education: A systematic review. *Encyclopedia, 6*(3), 66. https://doi.org/10.3390/encyclopedia6030066

Norton, S. (2019). Middle school mathematics pre-service teachers' content knowledge, confidence and self-efficacy. *Teacher Development, 23*(5), 529–548. https://doi.org/10.1080/13664530.2019.1668840

Park, S., & Chan, K. K. H. (2026). Unpacking PCK relationships and development in science teachers: A systematic synthesis of empirical research. *Review of Educational Research*. Advance online publication. https://doi.org/10.3102/00346543251394404

Pastötter, B., Urban, J., Lötzer, J., & Frings, C. (2022). Retrieval practice enhances new learning but does not affect performance in subsequent arithmetic tasks. *Journal of Cognition, 5*(1), 22. https://doi.org/10.5334/joc.216

Peikos, G., & Stavrou, D. (2025). ChatGPT for science lesson planning: An exploratory study based on pedagogical content knowledge. *Education Sciences, 15*(3), 338. https://doi.org/10.3390/educsci15030338

Sapkota, B., & Max, B. (2024). A conceptual synthesis on approximations of practice in mathematics teacher education. *Research in Mathematics Education, 26*(3), 569–595. https://doi.org/10.1080/14794802.2023.2207088

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. Methods in Ecology and Evolution, 11(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Shi, Y., Yu, K., Dong, Y., & Chen, F. (2026). Large language models in education: A systematic review of empirical applications, benefits, and challenges. *Computers and Education Artificial Intelligence, 10*, 100529. https://doi.org/10.1016/j.caeai.2025.100529

Shrestha, S., & Yi, J. (2026). TPACK-based professional development for the AI era: Fostering pre-service teachers' acceptance of generative AI in mathematics classrooms. *Journal of Mathematics Teacher Education*. Advance online publication. https://doi.org/10.1007/s10857-026-09747-0

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Sohrabi, Z., Ghasemi, S., Nouri Khaneghah, Z., Barry, A., Zhianifard, A., & Nosrati, S. (2025). A scoping review of quasi-experimental studies on simulation-based learning in medical education: Trends and insights from the last decade. *Medical Journal of the Islamic Republic of Iran, 39*, 70. https://doi.org/10.47176/mjiri.39.70

Son, T., Yeo, S., & Lee, D. (2024). Exploring elementary preservice teachers' responsive teaching in mathematics through an artificial intelligence-based chatbot. *Teaching and Teacher Education, 146*, 104640. https://doi.org/10.1016/j.tate.2024.104640

Stavropoulou, G., Daniilidou, A., & Nerantzaki, K. (2025). Exploring the Interplay of motivation, self-efficacy, critical thinking, and self-regulation in predicting academic achievement among university students. *F1000Research, 14*, 344. https://doi.org/10.12688/f1000research.161821.2

Su, Y., Luo, M., & Zhong, C. (2025). To chat or not: Pre-service English teachers' perceptions of and needs in chatbot's educational application. *SAGE Open, 15*(1). https://doi.org/10.1177/21582440251321853

Sun, Q., & Wang, W. (2025). Peer effects and learning mechanisms of digital innovation: Evidence from manufacturing industry in China. *Journal of Engineering and*

Technology Management, 79, 101938. https://doi.org/10.1016/j.jengtecman.2025.101938

Xiong, X. (2025). Influence of teaching styles of higher education teachers on students' engagement in learning: The mediating role of learning motivation. *Education for Chemical Engineers, 51*, 87–102. https://doi.org/10.1016/j.ece.2025.02.005

Yip, S. Y. (2025). Transition into teaching: Examining the pedagogical content knowledge (PCK) development of pre-service STEM career-change teachers. *Journal of Education for Teaching, 51*(3), 525–540. https://doi.org/10.1080/02607476.2025.2471901

Zhang, C., Hu, M., Wu, W., Chen, Y., Wang, K., Gao, T., & Fu, Q. (2026). AI versus teacher feedback in developing pre-service teachers' teaching self-efficacy: A quasi-experimental study in simulated teaching. *Teaching and Teacher Education, 172*, 105355. https://doi.org/10.1016/j.tate.2025.105355

Zhang, N., Ke, F., Dai, C., Southerland, S. A., & Barrett, A. (2026). Science and mathematics preservice teachers' perceptions and experiences of practicing dialogic teaching in generative AI-powered virtual reality simulation. *Teaching and Teacher Education, 171*, 105349. https://doi.org/10.1016/j.tate.2025.105349

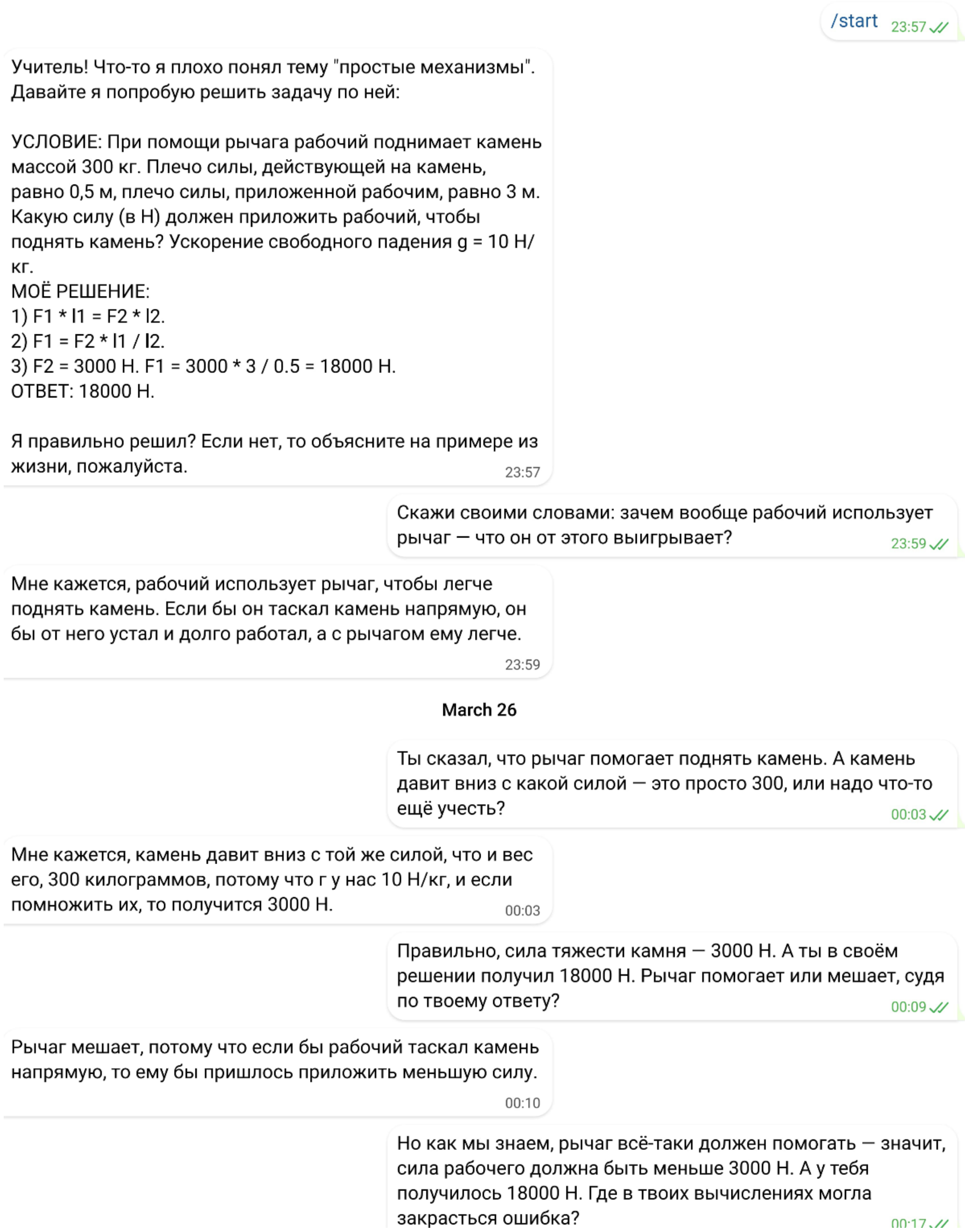
Zhang, N., Ke, F., Dai, C., Southerland, S. A., & Yuan, X. (2024). Seeking to support preservice teachers' responsive teaching: Leveraging artificial intelligence-supported virtual simulation. *British Journal of Educational Technology, 56*(3), 1148–1169. https://doi.org/10.1111/bjet.13522

Appendix 1

A sample participant–chatbot dialog (translated into English for publication purposes). Screenshots of the original Russian dialog are shown below.

Student-teacher:/start
Bot: Teacher! I didn't quite understand the topic "simple machines." Let me try to solve a problem on it:
CONDITION: When a lever is used, a worker lifts a stone with a mass of 300 kg. The lever arm of the force acting on the stone is 0.5 m, and the lever arm of the force applied by the worker is 3 m. What force (in N) must the worker apply to lift the stone? Acceleration due to gravity g = 10 N/kg.
MY SOLUTION: F1 × l1 = F2 × l2.
F1 = F2 × l1/l2.
F2 = 3000 N. F1 = 3000 × 3/0.5 = 18000 N.
ANSWER: 18000 N.
Did I solve it correctly? If not, please explain it with a real-life example.
Student-teacher: Explain in your own words: why does the worker use a lever at all — what does he gain from it?
Bot: It seems to me that the worker uses the lever to lift the stone more easily. If he were carrying the stone directly, he would get tired, and it would take a long time, but with the lever, it would be easier for him.
Student-teacher: You said the lever helps lift the stone. However, what force is the stone pushing down with — is it just 300, or is there something else we need to consider?
Bot: I think the stone pushes down with the same force as its weight, 300 kilograms, because g is 10 N/kg, and if you multiply them, you get 3000 N.
Student-teacher: Correct, the gravitational force on the stone is 3000 N. However, in your solution you get 18000 N. According to your answer, does the lever help or get in the way?
Bot: The lever gets in the way, because if the worker were carrying the stone directly, he would have to apply less force.
Student-teacher: However, as we know, the lever should actually help — which means the worker's force should be less than 3000 N. However, you got 18000 N. Where might the error be in your calculations?
Bot: Here's where the mistake is. This is likely because I mixed the forces and lengths. I thought that F2 = 3000 N and F1 = F2 × l1/l2, but that leads to the lever moving in the way instead of helping.
Student-teacher: So what needs to be fixed in your formula so that the worker's force comes out less than 3000 N?
Bot: Most likely, because I set up the proportion wrong: F1 × l1 = F2 × l2. I messed up when rearranging the formula. I think we need to set F2 = 3000 × l1/l2 so that the lever helps.
Student-teacher: Almost! Only F2 is already the known force of the stone, 3000 N. You need to find F1—the worker's force. Try plugging in the numbers.
Bot: I think we need to set F1 = 3000 × 3/0.5 so that the lever helps. Then, I'll get F1 = 18000 N, but that's probably wrong again, because the lever should help, not get in the way...

Student-teacher: Let's clarify: the lever arm for the stone (its gravitational force of 3000 N) is the short one, 0.5 m; the lever arm for your force F1 is the long one, 3 m. According to the lever rule: F1 × 3 = 3000 × 0.5. So what is F1?
Bot: F1 × 3 = 3000 × 0.5, which means F1 × 3 = 1500, and since 3 is greater than 0.5, F1 should be less than 1500, so F1 = 1500 ÷ 3.
Student-teacher: Excellent! F1 × 3 = 1500 N·m, so F1 = 1500/3 = 500 N. You solved it correctly — the worker applies only 500 N instead of 3000 N. Great job!



Вот, где ошибка. Наверное, потому что у меня неправильно связал силы и длины. Я подумал, что F2 = 3000 Н, а F1 = F2 * l1 / l2, но меня это приводит к тому, что рычаг мешает, а не помогает. 00:17

Так что же нужно исправить в твоей формуле, чтобы сила рабочего получилась меньше 3000 Н? 00:22

Наверное, потому что у меня неправильно поставлена пропорция: F1*л1 = F2*л2. Я ошибся при перестановке формулы, мне кажется, нужно поставить F2 = 3000*л1/л2, чтобы рычаг помогал. 00:22

Почти! Только F2 — это уже известная сила камня, 3000 Н. Тебе нужно найти F1 — силу рабочего. Попробуй подставить числа. 00:24

Мне кажется, нужно поставить F1 = 3000 * 3 / 0.5, чтобы рычаг помогал. Затем я получу F1 = 18000 Н, но это, скорее всего, опять будет неправильно, потому что рычаг должен мешать, не помогать... 00:24

Давай уточним: плечо камня (его сила тяжести 3000 Н) — короткое, 0,5 м; плечо твоей силы F1 — длинное, 3 м. По правилу рычага: F1 × 3 = 3000 × 0,5. Так чему равно F1? 00:26

Ф1 × 3 = 3000 × 0,5, значит Ф1 × 3 = 1500, и, наверное, потому что 3 больше 0,5, Ф1 должно быть меньше 1500, значит Ф1 = 1500 ÷ 3. 00:26

Отлично! F1 × 3 = 1500 Н·м, F1 = 1500 / 3 = 500 Н. Ты правильно решил — рабочий прикладывает всего 500 Н вместо 3000 Н. Молодец! 00:30