Check for updates

RESEARCH ARTICLE

# Artificial intelligence in everyday classrooms: Effects of a purpose-built AI study assistant on secondary students' learning, motivation, and engagement

[version 1; peer review: 1 approved, 1 not approved]

Theodoros Karafyllidis[1], Anna Vacalopoulou[2], Spyridoula Stamouli[2], Marios Papaevripidou[1], George Paraskevopoulos[2], Dimitra Dimitrakopoulou[3], Sofoklis Sotiriou[3], Vassilis Katsouros[2]

[1]Department of Educational Sciences, University of Cyprus, Nicosia, Cyprus
[2]Institute for Language and Speech Processing, Athena Research Center, Marousi, Attica, Greece
[3]Research and Development Department, Ellinogermaniki Agogi, Pallini, Greece

**v1** **First published:** 29 Jul 2026, **6**:267
https://doi.org/10.12688/openreseurope.23295.1
**Latest published:** 29 Jul 2026, **6**:267
https://doi.org/10.12688/openreseurope.23295.1

## Abstract

### Background

Recent advances in artificial intelligence (AI) have accelerated the development of educational technologies intended to support students' learning processes. However, classroom-based evidence on the educational impact of purpose-built AI applications remains limited, particularly across national contexts.

### Methods

This study examined the effects of an AI conversational study assistant (Study Buddy) on secondary students' academic performance, motivation, and engagement in authentic school settings. A quasi-experimental design was employed across two case studies conducted in Cyprus (Grade 10 Physics, N = 47) and Greece (Grade 7 History, N = 70). In each context, intact classes were assigned to experimental and control conditions. Academic performance was assessed using curriculum-aligned teacher-developed tests, while motivation and engagement were measured using the Motivation and Engagement Survey. In addition, system-generated usage logs were analyzed descriptively to document student access, interaction intensity, and patterns of tool use during the intervention. Statistical

**Open Peer Review**

**Approval Status** ✓ ✗

| | 1 | 2 |
|---|---|---|
| **version 1** 29 Jul 2026 | ✓ view | ✗ view |

1. **Samuel Ocen**, Mountains of the Moon University, Fort Portal, Uganda
2. **Andy Nguyen**, University of Oulu, Oulu, Finland

Any reports and responses or comments on the article can be found at the end of the article.

analyses included descriptive statistics and parametric or non-parametric group comparisons depending on distributional assumptions.

**Results**

Students who used Study Buddy demonstrated significantly higher learning gains in Physics and higher post-test scores in History compared to peers in control groups. In contrast, no statistically significant differences were observed for motivation or engagement in either case study. Usage log analysis indicated that students actively engaged with the application and that interaction patterns reflected the instructional design of each case study.

**Conclusions**

The findings suggest that purpose-built AI study assistants can support academic learning when integrated into regular classroom instruction. However, short-term exposure and predominantly task-focused interactions may limit their influence on motivational and engagement-related outcomes. The study contributes classroom-based, cross-national evidence on educational AI tools and highlights the importance of instructional design and teacher mediation in shaping both usage patterns and learning outcomes.

**Keywords**
artificial intelligence in education; conversational agents; academic achievement; student motivation; learning engagement




This article is included in the Erasmus+ gateway.

**Corresponding authors:** Theodoros Karafyllidis (karafyllidis.theodoros@ucy.ac.cy), Anna Vacalopoulou (avacalop@athenarc.gr), Spyridoula Stamouli (pstam@athenarc.gr)

**Author roles: Karafyllidis T**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Vacalopoulou A**: Conceptualization, Data Curation, Writing – Original Draft Preparation, Writing – Review & Editing; **Stamouli S**: Conceptualization, Data Curation, Funding Acquisition, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Papaevripidou M**: Conceptualization, Investigation, Methodology, Supervision, Writing – Review & Editing; **Paraskevopoulos G**: Data Curation, Software, Writing – Review & Editing; **Dimitrakopoulou D**: Investigation, Writing – Review & Editing; **Sotiriou S**: Resources, Supervision, Writing – Review & Editing; **Katsouros V**: Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This project has received funding from the European Union's Erasmus+ programme under grant agreement No [101087451] (AI4EDU – ERASMUS-EDU-2022-PI-FORWARD-LOT1).
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2026 Karafyllidis T *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Karafyllidis T, Vacalopoulou A, Stamouli S *et al.* **Artificial intelligence in everyday classrooms: Effects of a purpose-built AI study assistant on secondary students' learning, motivation, and engagement [version 1; peer review: 1 approved, 1 not approved]** Open Research Europe 2026, **6**:267 https://doi.org/10.12688/openreseurope.23295.1

**First published:** 29 Jul 2026, **6**:267 https://doi.org/10.12688/openreseurope.23295.1

## 1. Introduction

Artificial intelligence (AI) is advancing at an unprecedented pace and is increasingly positioned as a transformative force across multiple sectors of contemporary society, including education (Adiguzel et al., 2023). Recent developments in Machine Learning (ML), Natural Language Processing (NLP), and generative AI have enabled the creation of technologies capable of supporting learning, teaching, assessment, and educational decision-making in increasingly sophisticated ways (Holmes & Tuomi, 2022). Large-scale reviews of AI use in K–12 and higher education contexts consistently highlight its potential to personalize learning, provide timely feedback, and foster more adaptive and responsive educational environments (Chen et al., 2020a; Zafari et al., 2022). As a result, AI is widely portrayed as a key driver of educational innovation, capable of addressing long-standing challenges related to learner diversity, instructional efficiency, and student support (Kamalov et al., 2023).

Over the past decade, AI-powered educational tools have evolved from relatively narrow adaptive systems, such as automated assessment platforms, towards more interactive and conversational applications, including Intelligent Tutoring Systems (ITS) and conversational agents that engage learners through dialogue, explanation, and guided inquiry (Fernández-Herrero, 2024; Labadze et al., 2023; Pérez et al., 2020). This evolution has shifted research attention away from purely technical performance towards questions of pedagogical value and educational impact (Chiu et al., 2023). In particular, scholars increasingly examine how AI tools may influence students' learning experiences by supporting self-regulation, fostering active engagement, and enabling more individualized learning trajectories (Ayeni et al., 2024; Heeg & Avraamidou, 2023; Jia et al., 2024).

Empirical research suggests that AI-supported learning environments can influence students' understanding of subject content (Alneyadi & Wardat, 2023), motivation, and engagement with learning tasks (Badarudin et al., 2024), and overall learning experiences (Lee et al., 2022). Conversational agents and chatbot-based systems, in particular, have been associated with increased learner participation, perceived support, and motivation, as they offer accessible, non-judgmental interaction and immediate responses to student input (Kuhail et al., 2023). However, the magnitude and consistency of these effects vary considerably across studies, with outcomes strongly shaped by instructional design, duration of use, and the role of the teacher in mediating AI-supported activities.

Despite growing interest and experimentation with AI tools among students and educators, their systematic integration into everyday school practice remains limited. Multiple reviews point to a persistent gap between technological innovation and pedagogical implementation (Alasadi & Baiz, 2023; Chen et al., 2020b), noting that many AI tools are introduced without sufficient alignment with curriculum goals, learning theory, or classroom realities (Kavitha & Joshith, 2024). Moreover, a substantial proportion of existing studies are conducted in controlled or short-term experimental settings, offering limited insight into how AI functions under authentic classroom conditions characterized by time constraints, curricular pressures, and heterogeneous learner needs.

Three interrelated gaps in the literature are particularly salient. First, despite the growing availability of generative AI technologies, there remains limited empirical evidence on the educational impact of AI applications that are specifically designed for pedagogical purposes and aligned with curricular and instructional objectives. Much existing research focuses on general-purpose platforms that were not originally developed for learning contexts, making it difficult to draw robust conclusions about their educational value. Second, many studies continue to rely on controlled or short-term implementations, providing insufficient evidence on how AI-supported tools operate within everyday classroom practices and established instructional routines. Third, there is a notable lack of comparative, cross-national research examining how AI-supported learning tools function across different educational systems and institutional contexts. This constrains the generalizability of existing findings and limits understanding of how contextual factors shape educational outcomes.

In response to these gaps, the present study investigates the educational impact of Study Buddy (SB), a purpose-built AI study assistant designed to support students' learning activities. Unlike general-purpose AI platforms, SB was specifically developed for educational use and aligned with pedagogical objectives. The study examines its integration within regular classroom instruction, thereby capturing patterns of use and learning outcomes under authentic school conditions. Furthermore, by adopting a shared framework, involving secondary schools in Cyprus and Greece, the study offers a comparative perspective on the contextual stability of observed effects.

Guided by this framework, the study seeks to contribute empirically grounded evidence on how purpose-built AI-supported learning tools relate to key aspects of students' learning experiences. Specifically, it addresses the following research questions:

RQ1: To what extent does the use of an AI-supported study assistant influence students' academic performance in secondary education?

RQ2: To what extent does the use of an AI-supported study assistant influence students' motivation and engagement in learning activities?

## 2. Methodology

This study was conducted within the framework of the AI4EDU: Conversational AI Assistant for Teaching and Learning project (AI4EDU, project ID: 101087451, ERASMUS-EDU-2022-PI-FORWARD), a European initiative aimed at the design, development, and empirical evaluation of two conversational AI applications for school education: Study Buddy (SB), a student-facing study assistant, and Teacher Mate (TM), a teacher-facing support platform. The overarching goal of AI4EDU is to investigate how these AI-based applications can support teaching and learning processes under authentic classroom conditions and to generate evidence-informed guidance for their pedagogical integration.

The study adopts a focused, cross-national approach and reports findings from two school-based case studies implemented in Cyprus and Greece, both of which followed a shared evaluation framework and employed comparable instruments, enabling meaningful cross-case comparison. The present paper concentrates exclusively on the student-facing application, and examines its impact on students' academic performance, motivation, and engagement.

### 2.1 Study Buddy: An AI conversational study assistant

SB is an AI-driven application designed to function as a virtual study companion that supports students during independent and guided learning activities. Its primary purpose is to assist learners in reinforcing subject knowledge, organizing learning tasks, and developing effective study strategies through natural language interaction.

SB offers a broad set of pedagogically oriented tools, including:

- **Summarizer**: Condenses lengthy texts into concise summaries, enabling students to focus on key information.
- **Interactive concept exploration**: Facilitates dynamic, guided sessions to deepen students' understanding of complex topics.
- **Chronological order of events:** Generates detailed timelines from inputted events or time periods, complete with explanatory transitions and comprehension questions.
- **Extract key points:** Identifies critical elements in texts, making notetaking and study preparation more efficient.
- **Revise tool:** Reviews student-written assignments, offering feedback and actionable suggestions for improvement.
- **Auto-grade exercise:** Evaluates responses to questions instantly, providing detailed feedback to support continuous learning.
- **Explain term:** Delivers explanations of academic terms and concepts, helping students grasp essential knowledge.
- **Text adaptation:** Modifies text to align with students' reading abilities, skills, and learning objectives.
- Conversations across time: Encourages critical thinking and creativity by allowing students to engage in imaginative dialogues with historical figures, cultural icons, or scientists.
- **Debate tool:** Prepares students for debates by simulating an interactive experience where the AI acts as a debate partner.
- **Self-assessment quiz:** Creates customized quizzes based on students' study materials, promoting mastery of topics through self-assessment.

The application also includes a dashboard environment where students can complete teacher-assigned tests, review feedback, and track learning progress over time. Interaction with SB can take place through text- or voice-based input, allowing flexible and intuitive use. Overall, SB is intended to support self-directed learning while remaining adaptable to teacher guidance and curricular goals.

## 2.2 Evaluation design and implementation procedures

A quasi-experimental design was employed in both case studies. In each participating school, intact classes were assigned to either a control group or an experimental group, following the school's existing classroom structure. Group assignment was conducted by the participating teachers, with attention to maintaining comparable group sizes and broadly similar prior academic profiles.

Prior to the classroom implementation, preparatory training sessions were conducted by the research team with the participating teachers. These sessions aimed to introduce foundational AI concepts, address pedagogical and ethical considerations related to AI use in education, and familiarise teachers with the functionality and intended educational use of both SB and TM.

In both cases, the instructional intervention focused on the teaching of a specific curriculum-aligned topic over a defined instructional period corresponding to the time normally allocated to that unit. All students received regular classroom instruction delivered by their teacher, following national curriculum requirements. The distinguishing factor between the two conditions concerned students' access to SB.

Students in the experimental groups were given access to SB and participated in a familiarization session led by their teacher prior to the intervention. This session introduced the core functions of the application and outlined expectations for appropriate and purposeful use. During the implementation period, experimental-group students were encouraged to use SB to support their learning. Depending on the case, this use occurred either during independent study outside school or during structured classroom activities. In some instances, teachers provided guiding tasks to scaffold students' interaction with the application.

Students in control groups studied the same content using conventional learning resources, such as textbooks, worksheets, notes, and teacher-provided materials, without access to SB. The intervention in both cases lasted approximately three weeks, reflecting typical instructional practice and minimizing disruption to established teaching routines.

## 2.3 Case study contexts

### 2.3.1 Case study 1: Cyprus – Physics (Grade 10)

The first case study was conducted in a public upper-secondary school in Cyprus and involved two Grade 10 Physics classes (N = 47). One class served as the control group (n = 24), while the other served as the experimental group (n = 23). The instructional unit focused on the topic of gravitational force.

This case followed a pre-test–post-test design. Prior to the instructional unit, both groups completed a curriculum-aligned academic performance test and the Motivation and Engagement Survey (MES). During the intervention, experimental-group students used SB independently outside school to review and consolidate their understanding of the topic, while control-group students relied exclusively on traditional study methods. Following the three-week intervention, the same academic test and the MES were re-administered. This design enabled examination of changes in academic performance, motivation, and engagement over time, as well as comparisons between experimental and control conditions.

### 2.3.2 Case study 2: Greece – History (Grade 7)

The second case study was conducted in a lower-secondary school in Greece and involved Grade 7 History classes studying an instructional unit on Byzantine emperors. The participating teacher implemented the intervention across three intact classes: one control group (25 students) and two experimental groups (19 and 26 students, respectively). For analytical purposes, the two experimental classes were treated as a single experimental condition.

The implementation lasted approximately three weeks. Students in the experimental groups used SB primarily during scheduled sessions in the school's computer laboratory to search for and explore information related to Byzantine emperors, with emphasis on governance strategies and historical decision-making. The teacher explicitly encouraged students to critically evaluate AI-generated responses and to cross-reference information with textbooks and other

authoritative sources. Students synthesized their findings and presented them in class, supporting the development of historical thinking and digital literacy.

The control group studied the same content using traditional instructional materials without access to SB. In this case, academic performance was assessed using a post-test only, administered at the end of the instructional unit. Students in both groups completed the MES before and after the intervention.

## 2.4 Evaluation instruments

### 2.4.1 Academic performance tests

In both case studies, students' academic performance was assessed using curriculum-aligned tests aligned with the learning objectives of the respective instructional units. The tests reflected the learning objectives of the respective instructional units and corresponded to assessment practices typically employed in each subject area.

In the Cyprus Physics case, academic performance was measured using a test scored on a 100-point scale, administered both before and after the intervention. The test was based on the Newtonian Gravity Concept Inventory (NGCI) (Williamson, Willoughby, & Prather, 2013), which the participating teacher translated and adapted for use with the specific class context. It consisted of 25 multiple-choice questions on gravitational force–related content requiring justification of the selected answer.

In the Greece History case, academic performance was assessed using a paper-based post-test scored on a 20-point scale. The test consisted of two parts: (a) five true/false questions with justification and (b) five multiple-choice questions with justification. In both cases, identical assessment instruments were used for control and experimental groups.

### 2.4.2 Motivation and engagement survey (MES)

Students' motivation and engagement were assessed using the Motivation and Engagement Survey (MES) developed by Lee and Reeve (2012). The MES is a validated instrument consisting of 38 items rated on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree).

The survey captures multiple dimensions of students' learning-related motivation and engagement, including psychological need satisfaction (items 1–10), self-efficacy (items 11–15), mastery goals (items 16–18), behavioral engagement (items 19–23), emotional engagement (items 24–28), cognitive engagement (items 29–33), and agentic engagement (items 34–38).

- **Motivation** (Items 1–18)
- **Engagement** (Items 19–38)

Negatively worded items were reverse-coded so that higher scores consistently reflected more positive motivational and engagement profiles. The MES was administered to both experimental and control groups before and after the intervention in both case studies.

Responses were anonymized prior to analysis. Motivation and engagement were analyzed at the individual level and reported as group mean scores at each measurement point to examine overall trends over time while ensuring protection of student privacy.

## 2.5 Study Buddy usage data and implementation monitoring

To document the implementation of the intervention and to support interpretation of the study's outcome findings, system-generated usage data from SB were collected and analyzed. The primary purpose of this analysis was to monitor implementation fidelity and student exposure to the AI application under authentic classroom conditions.

SB automatically records anonymized interaction logs at the session level for all user interactions. Each log entry includes a pseudonymized user identifier, session duration (in seconds), number of conversational turns, the conversation title and content, and the specific pedagogical tool activated during the interaction, where applicable. No personally identifiable information was collected or stored.

**Table 1. Academic performance scores in case study 1 (Cyprus).**

| Group | Size | Pre-test mean score/100 | Post-test mean score/100 | Mean gain score/100 |
|---|---|---|---|---|
| Control | 24 | 37.00 | 59.17 | 22.17 |
| Experimental | 23 | 40.87 | 71.83 | 30.96 |

Usage data were collected for the full duration of the instructional intervention and were restricted to students assigned to the experimental groups. Although the interaction content was available to the research team, it was used solely to support contextual understanding of tool use and interaction patterns and was not subjected to systematic qualitative coding or discourse analysis.

Consistent with the exploratory and implementation-focused role of the usage data, analyses were conducted at a descriptive level. Usage indicators were not statistically linked to individual students' academic performance, motivation, or engagement scores. This methodological decision was taken to avoid overinterpretation of log data and to maintain a clear distinction between implementation monitoring and outcome evaluation.

### 2.6 Ethical considerations, data management, and analysis

The study was conducted in accordance with national and institutional ethical guidelines for educational research Ethical approval was obtained prior to the implementation of the study in both participating countries.[1]

Written informed consent was obtained from both students and their parents or legal guardians prior to participation. Participants and their guardians received detailed information sheets explaining the purpose of the study, the nature of the activities involved, the voluntary nature of participation, and the procedures for data handling and privacy protection. Students and parents or legal guardians also provided their assent by signing a corresponding consent form. Participation was voluntary, and students were informed that they could withdraw from the study at any time without academic or other consequences.

All questionnaire and assessment data were anonymized prior to analysis. Individual identifiers were removed, and results were analyzed using coded datasets to ensure confidentiality. Incomplete questionnaires and missing responses were excluded from statistical analyses; therefore, sample sizes in the motivation and engagement analyses reflect only valid responses at both measurement points.

All statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 29.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were computed to summarize students' academic performance, motivation, and engagement scores. Prior to inferential analyses, the Shapiro–Wilk test was employed to assess the normality of score distributions. Depending on the distributional properties of the data, either parametric (independent-samples t-tests) or non-parametric (Mann–Whitney U tests) procedures were applied to compare experimental and control groups. Statistical significance was evaluated at the 0.05 level.

## 3. Results

### 3.1 Impact of Study Buddy on academic performance

This subsection reports findings related to students' academic performance in the Physics and History case studies. Results are presented separately for each case due to differences in subject matter, assessment instruments, and study design.

#### 3.1.1 Cyprus case study: Physics (Grade 10)

Students' academic performance in the Cyprus case study was assessed using curriculum-aligned pre-test and post-test scores. Mean gain scores were calculated for each group by subtracting pre-test scores from post-test scores and were used as the primary indicator of academic progress. Descriptive statistics for pre-test scores, post-test scores, and gain scores are presented in Table 1.

[1]In Greece, the study was approved by the Ethics Committee of the Athena Research Center (reference number: 19–25/05/2023). In Cyprus, approval was granted by the Center for Educational Research and Evaluation, Ministry of Education, Sport and Youth (reference number: 07.30.002.002/140517).

**Table 2. Shapiro-Wilk test of normality for gain scores (Case study 1).**

| Group | Statistic | Degrees of freedom (df) | Significance |
|---|---|---|---|
| Control | 0.979 | 24 | 0.875 |
| Experimental | 0.935 | 23 | 0.140 |

**Table 3. Independent samples t-test results for gain scores (Case study 1).**

| Test | t-value | Degrees of freedom (df) | Significance | Effect size (Cohen's d) |
|---|---|---|---|---|
| Independent Samples t-test | −3.56 | 45 | <0.001 | −1.044 |

**Table 4. Academic performance scores in case study 2 (Greece).**

| Group | Size | Post-test mean score/20 |
|---|---|---|
| Control | 25 | 14.28 |
| Experimental | 45 | 17.09 |

**Table 5. Shapiro-Wilk test of normality (case study 2).**

| Group | Statistic | Degrees of freedom (df) | Significance |
|---|---|---|---|
| Control | 0.905 | 25 | 0.023 |
| Experimental | 0.878 | 45 | <0.001 |

**Table 6. Mann–Whitney U test results (case study 2).**

| Test | Mann-Whitney U value | Z | Significance | Effect size (r) |
|---|---|---|---|---|
| Mann-Whitney U test | 869.00 | 3.823 | <0.001 | 0.46 |

Both groups demonstrated improvement from pre-test to post-test. The experimental group showed a higher mean gain score than the control group. To examine whether this difference was statistically significant, the Shapiro-Wilk test was first conducted to assess the normality of gain score distributions. As shown in Table 2, gain scores in both groups were normally distributed ($p > 0.05$), supporting the use of parametric statistical analysis.

An independent samples t-test was then conducted to compare mean gain scores between the two groups. As shown in Table 3, the difference was statistically significant at the 0.05 level, indicating higher learning gains for students in the experimental group ($t = -3.56$, $p < 0.001$). The magnitude of this difference was large, as indicated by effect size estimates: Cohen's $d = -1.044$, reflecting substantially greater gain scores for the experimental group compared to the control group.

### 3.1.2 Greece case study: History (Grade 7)

In the Greek case study, students' academic performance was assessed using a curriculum-aligned post-test administered at the end of the instructional unit. Mean post-test scores were calculated for the control group and the combined experimental group. Descriptive statistics are presented in Table 4.

Students in the experimental groups achieved higher mean post-test scores than students in the control group. Prior to group comparison, the Shapiro-Wilk test was conducted to assess the normality of score distributions. As shown in Table 5, post-test scores in both groups deviated significantly from normality ($p < 0.05$).

Given the violation of normality assumptions, a non-parametric Mann–Whitney U test was conducted to compare post-test scores between the control and experimental groups, as shown in Table 6. The results revealed a statistically significant difference in academic performance, with students in the experimental group outperforming those in the control group ($U = 869.00$, $p < 0.001$).

To estimate the magnitude of this difference, an effect size was calculated using the standardized test statistic ($z = 3.823$) and total sample size ($N = 70$). The resulting effect size ($r = 0.46$) indicates a medium-to-large effect, suggesting that the observed performance difference was not only statistically significant but also educationally meaningful.

The test results revealed a statistically significant difference in post-test performance between the experimental and control groups, with students who used SB outperforming their peers.

## 3.2 Impact of Study Buddy on motivation and engagement

This subsection reports results related to students' motivation and engagement as measured by the MES. In line with the shared evaluation framework adopted across case studies, results are presented by construct (motivation and engagement) rather than by country. For both constructs, pre- and post-intervention mean scores were calculated for control and experimental groups. Mean gain scores (post-test minus pre-test) were then computed and used as the primary indicators of change over time. Prior to inferential analysis, tests of normality were conducted to determine the appropriate statistical procedures for each case.

### 3.2.1 Motivation

Students' motivation was assessed using the motivation subscale of the MES (Items 1–18). Table 7 presents pre-test and post-test mean motivation scores, along with the corresponding mean gain scores, for control and experimental groups in both case studies.

Across both case studies, changes in students' motivation over time were modest. In all groups, mean gain scores were negative, indicating slight decreases in motivation from pre- to post-test, with no substantial differences observed between experimental and control conditions.

To determine the appropriate inferential statistical tests, Shapiro-Wilk tests were conducted on motivation gain score distributions for control and experimental groups in each case study. The results are presented in Table 8.

The results indicated that the assumption of normality was violated for both groups in the Cyprus case ($p < 0.05$), whereas normality was not violated for either group in the Greece case ($p > 0.05$). Accordingly, non-parametric statistical testing was employed for the Cyprus case study, and parametric statistical testing was employed for the Greece case study.

**Table 7. Motivation pre-test, post-test, and mean gain scores.**

| Case study | Group | Size | Pre-test mean | Post-test mean | Mean gain |
|---|---|---|---|---|---|
| Physics (Cyprus) | Control | 24 | 3.60 | 3.53 | −0.07 |
| | Experimental | 21 | 3.48 | 3.47 | −0.02 |
| History (Greece) | Control | 14 | 3.53 | 3.25 | −0.27 |
| | Experimental | 14 | 3.58 | 3.42 | −0.16 |

**Table 8. Shapiro-Wilk test of normality for motivation mean gain scores.**

| Case study | Group | Statistic | Degrees of freedom (df) | Significance |
|---|---|---|---|---|
| Physics (Cyprus) | Control | 0.910 | 24 | 0.036 |
| | Experimental | 0.886 | 21 | 0.019 |
| History (Greece) | Control | 0.940 | 14 | 0.416 |
| | Experimental | 0.954 | 14 | 0.621 |

**Table 9. Statistical comparison of motivation gain scores.**

| Case study | Test | Test statistic | Significance |
|---|---|---|---|
| Physics (Cyprus) | Mann-Whitney U test | U = 245.5 | 0.882 |
| History (Greece) | Independent Samples t-test | t(28) = −0.640 | 0.528 |

**Table 10. Engagement pre-test, post-test, and mean gain scores.**

| Case study | Group | Size | Pre-test mean | Post-test mean | Mean gain |
|---|---|---|---|---|---|
| Physics (Cyprus) | Control | 24 | 3.34 | 3.40 | 0.06 |
| | Experimental | 21 | 3.15 | 3.31 | 0.16 |
| History (Greece) | Control | 14 | 3.38 | 3.02 | −0.36 |
| | Experimental | 14 | 3.35 | 3.22 | −0.13 |

**Table 11. Shapiro-Wilk test of normality for engagement mean gain scores.**

| Case study | Group | Statistic | Degrees of freedom (df) | Significance |
|---|---|---|---|---|
| Physics (Cyprus) | Control | 0.960 | 24 | 0.430 |
| | Experimental | 0.944 | 21 | 0.266 |
| History (Greece) | Control | 0.915 | 14 | 0.189 |
| | Experimental | 0.947 | 14 | 0.522 |

**Table 12. Statistical comparison of engagement gain scores.**

| Case study | Test | Test statistic | Significance |
|---|---|---|---|
| Physics (Cyprus) | Independent Samples t-test | t = −0,975 | 0,167 |
| History (Greece) | Independent Samples t-test | t = −1,434 | 0,163 |

For the Cyprus case study, a Mann–Whitney U test was conducted to examine differences in motivation gain scores between control and experimental groups. For the Greece case study, an independent-samples t-test was conducted to compare motivation gain scores between the two groups. The results are presented in Table 9.

The analyses indicated that differences in motivation gain scores between experimental and control groups were not statistically significant in either case study at the 0.05 level.

### 3.2.2 Engagement

Students' engagement was assessed using the engagement subscale of the MES (Items 19–38). Table 10 presents pre-test and post-test mean engagement scores, together with the corresponding mean gain scores, for control and experimental groups in both case studies.

Overall, changes in students' engagement over time were limited. In the Cyprus case, both groups showed small increases in engagement, with slightly higher gains observed in the experimental group. In contrast, in the Greece case, engagement scores decreased from pre- to post-test in both groups, with larger declines observed in the control group.

To determine the appropriate inferential statistical tests, Shapiro-Wilk tests were conducted on engagement gain score distributions for control and experimental groups in each case study. The results are presented in Table 11.

The results indicated that the assumption of normality was not violated for engagement gain scores in either group in both case studies ($p > 0.05$). Accordingly, parametric statistical testing was employed in both cases. Independent-samples

t-tests were conducted to compare engagement gain scores between control and experimental groups in the Cyprus and Greece case studies. The results are presented in Table 12.

The analyses indicated that differences in engagement gain scores between experimental and control groups were not statistically significant in either case study at the 0.05 level.

### 3.3 Analysis of Study Buddy usage logs

This subsection presents descriptive results from the analysis of SB usage logs for the experimental groups in both case studies. The analysis focuses on student access to the application, intensity of use as reflected in session duration and conversational turns, and the pedagogical tools activated during interactions.

In the Cyprus Physics case, 20 of the 23 students assigned to the experimental group accessed SB at least once during the three-week instructional period, corresponding to an adoption rate of approximately 87%. In the Greece History case, usage logs indicate that 27 of the 45 students in the experimental condition accessed SB during the intervention, yielding an adoption rate of approximately 60%. Taken together, these figures indicate that SB was used by a substantial proportion of students in both contexts, although uptake was higher in the Cyprus case.

Across both case studies, students' interactions with SB exhibited considerable variability in terms of session duration and number of conversational turns. In the Cyprus case, usage logs show a combination of brief and extended interaction sessions. Some students engaged in short sessions involving one or two conversational turns, while others interacted with the application repeatedly across the intervention period, including sessions of extended duration. This pattern is consistent with the primarily independent, out-of-school use of SB in this context and reflects differences in students' self-directed study behaviors.

In the Greece case, interaction sessions were often longer and, in several instances, substantially extended in duration. Multiple sessions involved sustained interaction and higher numbers of conversational turns, indicating prolonged engagement with the application. This pattern reflects the structured, classroom-based implementation of SB, where students used the application during scheduled computer-lab sessions to explore historical content, elaborate ideas, and refine understanding under teacher guidance.

Analysis of tool usage patterns is aligned with the instructional focus of each case study. In the Cyprus Physics case, the most frequently used pedagogical tool was "Explain term", suggesting that students primarily engaged with SB to clarify concepts and reinforce subject-specific understanding. Students also used a range of additional tools, including "Interactive concept exploration", "Conversation across time", "Debate tool", "Revision tool", "Extract key points", "Text adaptation", and "Self-assessment quiz". A notable number of interaction sessions were logged without an explicitly selected tool, indicating free-form conversational exchanges aimed at clarification or follow-up questions ("Chat with SB" functionality).

In the Greece History case, the dominant mode of interaction was the "Conversation across time" tool, which was used extensively and often across multiple sessions. Other tools used included "Chronological order of events", "Extract key points", "Explain term", "Auto-grade exercise", "Summarizer", and "Self-assessment quiz". As in the Cyprus case, some interactions relied on the open-ended dialogue with the AI assistant.

Overall, the usage logs indicate that SB was actively used in both experimental contexts, with differences in adoption levels, interaction intensity, and tool-use patterns reflecting the instructional planning and pedagogical objectives of each case study.

## 4. Discussion

### 4.1 Summary of findings

This study examined the educational impact of a purpose-built AI conversational study assistant on students' academic performance, motivation, and engagement in two secondary school contexts in Cyprus and Greece. The findings indicate that the use of SB was associated with statistically significant improvements in academic performance in both case studies, whereas no statistically significant effects were observed for students' motivation and engagement.

In both contexts, usage log analysis confirmed that students in experimental groups engaged with the application during the intervention period. Although adoption and interaction patterns differed across cases, the logs indicate that SB was used in ways consistent with the instructional design and pedagogical aims of each setting.

### 4.2 Interpretation of academic performance outcomes

The positive association between SB use and academic performance aligns with prior research highlighting the potential of AI-supported learning environments to enhance students' understanding and practice of curricular content (Alneyadi & Wardat, 2023; Dahri et al., 2024; Lee et al., 2022). Features such as interactive explanations, structured prompts, and opportunities for self-assessment likely supported retrieval practice, elaboration, and metacognitive monitoring.

The usage analytics provide important contextual insight into these outcomes. In the Cyprus Physics case, high adoption, and predominant use of concept-focused tools, particularly "Explain term", suggest that SB functioned mainly as a revision and clarification aid during independent study. In the Greece History case, students primarily engaged through dialogic and exploratory interactions, especially via "Conversation across time", reflecting the inquiry-oriented nature of the classroom activities. These patterns suggest that the observed learning benefits were supported not only by access to the AI tool, but by how its functionalities were enacted within specific instructional designs.

These findings underscore the importance of instructional mediation in AI-supported learning. Rather than operating as autonomous learning agents, AI applications appear most effective when embedded within coherent pedagogical frameworks that align the tool with curricular goals and teacher guidance.

### 4.3 Interpretation of motivation and engagement findings

In contrast to the academic performance results, no significant effects on motivation and engagement were observed. Although SB was actively used in both contexts during the pilot implementation, changes in motivational and engagement-related measures were modest and did not differ significantly between experimental and control groups. This finding diverges from a growing body of research reporting positive associations between AI-supported learning environments and students' motivational and affective outcomes (Badarudin et al., 2024; Kuhail et al., 2023; Lee et al., 2022).

One possible explanation is that students' interactions with SB were predominantly task- and content-focused, as indicated by the usage logs. While such interactions may support cognitive outcomes, they may be insufficient on their own to influence broader motivational dispositions, particularly over a short intervention period. Additionally, contextual factors such as the fixed three-weeks period of pilots, curricular demands, assessment pressures, and established classroom routines may have moderated the potential motivational impact of AI-supported activities. These findings suggest that motivational and engagement benefits may require longer-term exposure and more deliberate integration of autonomy-supportive and collaborative pedagogical strategies.

### 4.4 Interpretation of usage logs

The usage log analysis demonstrates that SB was adopted and used in both experimental contexts, but in different ways aligned with subject matter and implementation mode. In Physics, usage was primarily oriented toward concept clarification and revision during independent study, whereas in History, usage emphasized dialogic exploration, narrative engagement, and synthesis within structured classroom activities.

These findings highlight the value of incorporating usage analytics into evaluations of educational AI tools. By documenting not only whether students used the application but also how they engaged with its functionalities, usage data strengthen the interpretation of outcome measures and provide insight into the pedagogical mechanisms through which AI-supported tools may influence learning.

## 5. Conclusion

This study examined the educational impact of a purpose-built AI conversational study assistant in secondary school classrooms in Cyprus and Greece, with particular focus on students' academic performance, motivation, and engagement. The findings indicate that SB use was consistently associated with significantly higher academic outcomes in both case studies. Students in the experimental groups demonstrated greater learning gains in Physics and achieved higher post-test scores in History compared to their peers in control groups.

Analysis of usage logs confirms that these outcomes occurred in contexts where students actively engaged with the application, albeit in different ways aligned with instructional design. In Physics, SB was primarily used for concept clarification and revision during independent study, while in History it supported dialogic, inquiry-oriented exploration under teacher guidance. These patterns suggest that the educational value of AI-supported study assistants lies not only in their availability, but in how they are integrated into subject-specific pedagogical practices.

With respect to motivation and engagement, no statistically significant differences were observed between experimental and control groups. Although SB was used across both contexts, short-term exposure and predominantly task-focused interactions may have limited its capacity to influence broader affective and behavioral outcomes. This finding underscores the importance of implementation duration, pedagogical design, and contextual alignment when aiming to support motivational dimensions of learning through AI tools.

Overall, the study suggests that purpose-built AI study assistants can function as effective cognitive supports for academic learning when embedded within regular instruction, while their impact on motivation and engagement may be more context-dependent and require sustained, intentionally designed pedagogical integration.

Several limitations should be considered when interpreting these results. First, the study employed a quasi-experimental design with intact classes, which limits causal inference. Although efforts were made to ensure group comparability, unmeasured pre-existing differences may have influenced outcomes. Second, sample sizes were relatively small, particularly in the motivation and engagement analyses, reducing statistical power and limiting generalizability. Third, the intervention duration was short; longer-term studies are needed to examine whether observed academic benefits persist over time and whether motivational and engagement-related effects emerge with sustained exposure. Fourth, although usage logs provided valuable insight into student access and interaction patterns, usage indicators were analyzed descriptively and were not linked to individual learning, motivation, or engagement outcomes. Future research should examine how variation in intensity and type of AI relates to learning processes and outcomes, while maintaining appropriate ethical safeguards. Finally, the study focused on two countries with relatively similar educational systems, which may constrain generalizability to other contexts. Future research should extend comparative analyses to more diverse educational settings and explore how different instructional models shape both AI usage patterns and educational impact.

## Ethics and consent

Ethical approval for this study was obtained prior to the implementation of research in both participating countries. In Greece, the study was approved by the Ethics Committee of the Athena Research Center (reference number: 19–25/05/2023). In Cyprus, approval was granted by the Center for Educational Research and Evaluation (CERE), Ministry of Education, Sport and Youth (reference number: 07.30.002.002/140517).

Participation in the study was voluntary and written informed consent was obtained from participants and from the parents or legal guardians of minor participants prior to data collection.

## Data and software availability

### Source data

Zenodo. *Dataset supporting "Artificial Intelligence in Everyday Classrooms: Effects of a Purpose-Built AI Study Assistant on Secondary Students' Learning, Motivation, and Engagement".* https://doi.org/10.5281/zenodo.19232684 (Karafyllidis et al., 2026).

This project contains the following source data:

paper_data.xlsx (Anonymised dataset containing students' academic test scores for Physics and History (pre- and post-test; control and experimental groups), motivation and engagement survey scores (pre- and post-test; control and experimental groups), and outputs of the statistical analyses conducted in SPSS.

This work contains the following underlying data:

1. Motivation and Engagement Survey (pdf file, English version): administered pre- and post-intervention to measure students' motivation and engagement

2. History academic test (pdf file, in Greek and English): used as both pre-test and post-test to assess academic performance of both experimental and control group

3. Anonymised log data from Study Buddy usage by students in Greece and Cyprus (xlsx files) in compliance with GDPR and institutional ethical approvals.

Source and underlying data are available under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

## References


Adiguzel T, Kaya MH, Cansu FK: **Revolutionizing education with AI: Exploring the transformative potential of ChatGPT.** *Contemp. Educ. Technol.* 2023; **15**(3): ep429.
**Publisher Full Text**

Alasadi EA, Baiz CR: **Generative AI in Education and Research: Opportunities, Concerns, and Solutions.** *J. Chem. Educ.* 2023; **100**(8): 2965–2971.
**Publisher Full Text**

Alneyadi S, Wardat Y: **ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools.** *Contemp. Educ. Technol.* 2023; **15**(4): ep448.
**Publisher Full Text**

Ayeni OO, Al Hamad NM, Chisom ON, *et al.*: **AI in education: A review of personalized learning and educational technology.** *GSC Advanced Research and Reviews* 2024; **18**(2): 261–271.
**Publisher Full Text**

Badarudin LP, Ahmad T, Muhammad S: **The Effect of AI (Artificial Intelligence) in Education on Student Motivation: A Systematic Literature Review.** *Journal for Lesson and Learning Studies* 2024; **8**(1): 1–10.
**Publisher Full Text**

Chen L, Chen P, Lin Z: **Artificial Intelligence in Education: A Review.** *IEEE Access* 2020a; **8**: 75264–75278.
**Publisher Full Text**

Chen X, Xie H, Zou D, *et al.*: **Application and theory gaps during the rise of Artificial Intelligence in Education.** *Computers and Education: Artificial Intelligence* 2020b; **1**: 100002.
**Publisher Full Text**

Chiu TKF, Xia Q, Zhou X, *et al.*: **Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education.** *Computers and Education: Artificial Intelligence* 2023; **4**: 100118.
**Publisher Full Text**

Dahri NA, Yahaya N, Al-Rahmi WM, *et al.*: **Investigating AI-based academic support acceptance and its impact on students' performance in Malaysian and Pakistani higher education institutions.** *Educ. Inf. Technol.* 2024; **29**: 18695–18744.
**Publisher Full Text**

Fernández-Herrero J: **Evaluating Recent Advances in Affective Intelligent Tutoring Systems: A Scoping Review of Educational Impacts and Future Prospects.** *Educ. Sci.* 2024; **14**(8).
**Publisher Full Text**

Heeg DM, Avraamidou L: **The use of Artificial intelligence in school science: a systematic literature review.** *Educ. Media Int.* 2023; **60**(2): 125–150.
**Publisher Full Text**

Holmes W, Tuomi I: **State of the art and practice in AI in education.** *Eur. J. Educ.* 2022; **57**(4): 542–570.
**Publisher Full Text**

Jia F, Sun D, Looi, *et al.*: **Artificial Intelligence in Science Education (2013–2023): Research Trends in Ten Years.** *J. Sci. Educ. Technol.* 2024; **33**(1): 94–117.
**Publisher Full Text**

Karafyllidis T, Stamouli S, Vacalopoulou A: Supporting Data for paper "Artificial Intelligence in Everyday Classrooms: Effects of a Purpose-Built AI Study Assistant on Secondary Students' Learning, Motivation, and Engagement". [Data set]. *Zenodo* 2026.
**Publisher Full Text**

Kamalov F, Santandreu Calonge D, Gurrib I: **New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution.** *Sustainability* 2023; **15**(16).
**Publisher Full Text**

Kavitha K, Joshith VP: **Pedagogical incorporation of artificial intelligence in K-12 science education: A decadal bibliometric mapping and systematic literature review (2013-2023).** *Journal of Pedagogical Research* 2024; **8**(4): 437–465.
**Publisher Full Text**

Kuhail MA, Alturki N, Alramlawi S, *et al.*: **Interacting with educational chatbots: A systematic review.** *Educ. Inf. Technol.* 2023; **28**(1): 973–1018.
**Publisher Full Text**

Labadze L, Grigolia M, Machaidze L: **Role of AI chatbots in education: systematic literature review.** *Int. J. Educ. Technol. High. Educ.* 2023; **20**(1): 56–72.
**Publisher Full Text**

Lee YF, Hwang GJ, Chen PY: **Impacts of an AI-based chabot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation.** *Educ. Technol. Res. Dev.* 2022; **70**(5): 1843–1865.
**Publisher Full Text**

Lee W, Reeve J: **Teachers' estimates of their students' motivation and engagement: Being in synch with students.** *Educ. Psychol.* 2012; **32**(6): 727–747.
**Publisher Full Text**

Pérez JQ, Daradoumis T, Puig JMM: **Rediscovering the use of chatbots in education: A systematic literature review.** *Comput. Appl. Eng. Educ.* 2020; **28**(6): 1549–1565.
**Publisher Full Text**

Williamson KE, Willoughby SD, Prather EE: **Development of the Newtonian Gravity Concept Inventory.** *Astron. Educ. Rev.* 2013; **12**(1): 1–20.
**Publisher Full Text**

Zafari M, Bazargani JS, Sadeghi-Niaraki A, *et al.*: **Artificial Intelligence Applications in K-12 Education: A Systematic Literature Review.** *IEEE Access* 2022; **10**: 61905–61921.
**Publisher Full Text**

# Open Peer Review

## Current Peer Review Status: ✔ ✘

**Version 1**

Reviewer Report 03 September 2026

https://doi.org/10.21956/openreseurope.25202.r78464

© **2026 Nguyen A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✘ **Andy Nguyen**
University of Oulu, Oulu, Finland

This manuscript examines the use of a purpose-built conversational AI assistant in two authentic secondary-school settings. The study has several strengths, including its classroom-based implementation, attention to both learning and motivational outcomes, and the publication of outcome and usage data. I also appreciate that the authors report nonsignificant findings regarding motivation rather than focusing selectively on favorable results. Overall, the study addresses a timely topic and has the potential to contribute useful empirical evidence on conversational AI in school education. Nevertheless, there are several substantial concerns regarding the current framing of the study, methodological rigor, theoretical grounding, and interpretation of the findings.

My primary concern is that the current research design and analyses do not support the causal language used throughout the title, research questions, abstract, discussion, and conclusion. The intervention was assigned at the class level, whereas the inferential analyses appear to treat individual students as independently assigned observations. This creates a fundamental unit-of-analysis problem because students within the same class share teachers, instructional practices, peer interactions, schedules, and broader classroom conditions.

In Cyprus, one class received the intervention and one class served as the control. In Greece, two experimental classes were compared with one control class. Nevertheless, the analyses treat the 47 and 70 students, respectively, as independent observations. In the Cyprus study, intervention condition is therefore completely confounded with class membership. The Greece study faces a similar problem, particularly because there is only one control class. Consequently, the reported $p$-values cannot meaningfully distinguish an intervention effect from ordinary between-class differences. Although multilevel modelling or cluster-robust standard errors would ordinarily be considered when observations are nested within classes, these approaches cannot reliably address the present problem given the very small number of clusters. Statistical inference based on only two or three clusters would itself be highly unstable.

This issue requires substantial reconsideration of the study design and claims. The authors should

either (a) collect data from a sufficient number of classes within each condition to permit defensible inference regarding intervention effects, or (b) substantially reframe the current study as two exploratory classroom implementation case studies. Under the latter approach, causal terms such as “effects,” “influence,” and similar formulations should be removed or substantially qualified throughout the manuscript, and the findings should instead be interpreted as descriptive or exploratory class-level associations. The manuscript should also clearly report the number of teachers involved, how classes were allocated to conditions, and whether the same teacher taught the experimental and comparison classes.

There are additional concerns regarding baseline equivalence. In the Cyprus Physics study, the experimental class already performed better at pre-test. This difference should not be treated as negligible, nor does it support the claim that the two classes had broadly similar prior achievement profiles. Comparing gain scores does not adequately address baseline imbalance or potential selection bias in a nonrandomized design. For baseline–follow-up data, an ANCOVA or regression model predicting post-test achievement while controlling for pre-test performance would generally provide a more appropriate analysis than comparing gain scores. However, even such an analysis would not resolve the more fundamental confounding between intervention condition and class membership in the present design.

The Greece History study raises a related concern because no clearly defined achievement pre-test is reported. The accompanying workbook contains a variable labelled “Term_grade,” but the manuscript does not adequately define or analyze this variable. The authors should explain what this measure represents, when it was collected, how it was assessed, and whether it predates the intervention. If it represents prior achievement, its potential use as a baseline covariate should be considered. Without an appropriate baseline measure, the History findings demonstrate a post-intervention difference between classes, rather than evidence of greater learning or learning gains attributable to the conversational AI intervention.

Beyond these methodological issues, I also have concerns about the conceptual framing of the intervention. The manuscript seeks to investigate causal relationships and intervention effects, yet the actual design, functionality, usability, and utility of the conversational AI assistant are likely to play a central role in determining any observed outcomes. These characteristics are not sufficiently conceptualized, measured, or empirically examined. It is therefore difficult to determine whether the reported findings reflect the presence of a conversational AI assistant per se, the particular pedagogical and technological design of this specific system, differences in how students engaged with it, or other classroom-level factors. The manuscript would benefit from a clearer theoretical account of how specific AI functionalities and patterns of student–AI interaction is expected to produce the proposed learning and motivational outcomes, together with empirical evidence concerning whether the system was actually used in the intended ways.

Finally, the theoretical and literature grounding should be updated substantially. Conversational and generative AI in education is developing rapidly, yet the manuscript does not sufficiently engage with recent state-of-the-art research. In particular, there appears to be limited engagement with relevant literature published in 2025 and more recently. The authors should update the literature review and discussion to position the study against recent empirical and conceptual developments concerning conversational and generative AI in education, including research on pedagogical design, learner–AI interaction, AI-supported learning processes, and the conditions under which such systems may or may not contribute to learning outcomes.

Taken together, these concerns affect not only individual statistical analyses but also the level of inference that can reasonably be drawn from the study. The authentic classroom implementation and openly reported data provide a potentially useful basis for an exploratory contribution. **However, substantial revisions to the methodological treatment, theoretical grounding, study framing, and interpretation of the findings are needed before the manuscript can support its current claims.**

**Is the work clearly and accurately presented and does it engage with the current literature?**
No

**Is the study design appropriate and is the work technically sound?**
Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**
No

**Are all the source data and materials underlying the results available?**
Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**
Partly

**Are the conclusions drawn adequately supported by the results?**
Partly

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** AI in education, Edtech, learning sciences, design science

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.**

Reviewer Report 31 August 2026

https://doi.org/10.21956/openreseurope.25202.r79730

**© 2026 Ocen S.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Samuel Ocen**
Mountains of the Moon University, Fort Portal, Western Region, Uganda

Review Report: "Artificial Intelligence in Everyday Classrooms" (Karafyllidis et al.)

**Summary:** This cross-national quasi-experimental study evaluates "Study Buddy," a purpose-built AI conversational assistant, across two secondary classrooms (Cyprus Physics, N=47; Greece History, N=70). Findings show statistically significant academic gains in both cases (Cohen's d = -1.044 in Cyprus; r = 0.46 in Greece), but no significant effects on motivation or engagement.

**Strengths:**

- Authentic classroom implementation rather than artificial lab conditions strengthens ecological validity.
- Cross-national design (Cyprus/Greece) with a shared evaluation framework enables meaningful comparison.
- Appropriate statistical rigor: normality testing (Shapiro-Wilk) correctly guided the choice between parametric and non-parametric tests.
- Usage-log analysis adds valuable triangulation, showing how tool-use patterns (e.g., "Explain term" vs. "Conversation across time") aligned with subject-specific pedagogy.
- Transparent reporting of null findings on motivation/engagement avoids overselling the intervention.

**Weaknesses:**

- Small, unequal sample sizes (especially Cyprus, n=23–24 per group) limit statistical power and generalizability.
- Quasi-experimental design with intact classes and teacher-led group assignment raises concerns about selection bias and unmeasured confounds.
- Short intervention window (~3 weeks) constrains conclusions about durability of academic gains and precludes fair assessment of motivation/engagement, which plausibly need longer exposure.
- Usage data were not statistically linked to individual outcomes, so the mechanism connecting tool use to learning gains remains inferential rather than demonstrated.
- Reliance on teacher-adapted/translated instruments (e.g., NGCI translation) introduces potential validity threats not fully addressed.

**Overall assessment:** A methodologically sound, policy-relevant contribution to classroom-based AI-in-education research, with promising evidence for cognitive/academic benefits of purpose-built (vs. general-purpose) AI tools. Conclusions on motivation and engagement should be treated as preliminary given design and duration limitations. Recommend replication with larger samples, randomized assignment where feasible, and longer intervention periods before broader claims are made.

**Is the work clearly and accurately presented and does it engage with the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**

Partly

**Are all the source data and materials underlying the results available?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Artificial Intelligence, Machine Learning and Natural Language Programming but as well interested in AI for education

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**