OPEN PRAXIS

An Experimental Study of the Effect of AI Course Assistants on Student Grade Outcomes

RESEARCH ARTICLE

GEORGE HANSHAW
ETHAN WICKER
KRISTEN DENLINGER
*Author affiliations can be found in the back matter of this article

INTERNATIONAL COUNCIL FOR OPEN AND DISTANCE EDUCATION

ABSTRACT

This study investigates the effects of utilizing the AI course assistant, Spark, on student academic performance at Los Angeles Pacific University (LAPU). Focusing on grade outcomes, the research reveals that students who engaged with Spark three or more times achieved significantly higher grades compared to those who did not use the AI assistant. Through nonparametric statistical methods, including the Mann-Whitney U-test and Wilcoxon signed-rank test, we found a moderate-to-strong positive relationship between Spark utilization and improved academic performance. Propensity score matching was used to control for confounding factors such as age, gender, and ethnicity, ensuring robust comparisons between groups. The results underscore the potential of AI tools like Spark to enhance student success, particularly in online and asynchronous learning environments. These findings suggest that frequent interactions with AI assistants contribute to improved academic outcomes, offering valuable insights for institutions seeking to integrate AI technology to support student learning.

CORRESPONDING AUTHOR:
George Hanshaw
Los Angeles Pacific University, US
ghanshaw@lapu.edu

KEYWORDS:
AI course assistants; online education; student grade outcomes; academic performance; artificial intelligence; grade outcomes; engagement; education technology; personalized support; educational equity

TO CITE THIS ARTICLE:
Hanshaw, G., Wicker, E., & Denlinger, K. (2025). An Experimental Study of the Effect of AI Course Assistants on Student Grade Outcomes. *Open Praxis*, *17*(2), pp. 270–285. DOI: https://doi.org/10.55982/openpraxis.17.2.772

The recent proliferation of Generative Artificial Intelligence (GenAI) tools in educational settings has led to the inevitable debate on the efficacy of such tools. Questions continue to arise on whether or not institutions of higher education should endorse GenAI usage on campus. While many colleges and universities have chosen to discourage student AI usage for coursework, other such institutions have embraced the current trend, going so far as to embed GenAI tools into the digital classroom.

With the vast amount of educational technology (EdTech) tools available, including GenAI, it can be a challenge for institutions to decide which tools to adopt. This is especially true, as EdTech tools can be costly in terms of both monetary outlay and in terms of time and resources for implementation. As Wang and Zhao. (2024) argue, it is thus important for institutions to adopt AI tools because they significantly improve student's learning outcomes.

With this in mind, we have evaluated the effectiveness of specially trained AI course assistants on student academic success. In exploring the impact of these AI course assistants on academic performance, this study is guided by a theoretical framework designed to assess the effective of course-imbedded AI classroom assistants. This study aims to determine whether there is a measurable difference in grade point average (GPA) outcomes between students who utilized the AI course assistant and those who did not, providing insight into the overall impact of the technology on academic achievement. The authors acknowledge that GPA is just one measure of classroom success, but is here used as a proxy to quantify learning.

BACKGROUND

The focus of this study is on student usage of the GenAI course assistant, "Spark," at Los Angeles Pacific University (LAPU). LAPU is a fully online, asynchronous university with a total student population of just under 1,900 students. LAPU caters to full-time working adults, with approximately 61.2% of its student population being part-time. Spark was first piloted in the Spring 2 term of 2024, with a total of 113 students (77 students in BIBL 230: Biblical Literature and 36 students in PSYC 105: Introduction to Psychology). These courses are General Education requirements for students at LAPU. In the Summer 1 term of 2024, Spark was fully implemented in all courses across the university at both undergraduate and graduate levels.

For the purpose of our study, "Spark" is the name given to the GenAI course assistants used in various courses at both the undergraduate and graduate levels for both traditional and competency-based courses. Spark is provided via Nectir, an educational technology company that specializes in providing AI infrastructure to higher educational institutions. Nectir's platform enabled LAPU to create multiple Sparks, specific to individual courses across the LAPU course catalog. Students and instructors access these Sparks typically via the learning management system (LMS), where a hyperlink to the GenAI course assistants infrastructure is provided in each weekly overview. Authentication to Nectir's platform is also managed via the LMS integration. All courses have a course-specific Spark created for use by students enrolled in a given course. There is also a campus-wide Writing Tutor Spark available for use by all students.

A process known as Retrieval-Augmented Generation (RAG) is used to provide each Spark with knowledge and content relevant to a given course. RAG refers to an evolving family of techniques used to improve the generated content produced via a generative Large Language Model (LLM). For example, the specific Spark for BIOL 225: Microbiology/Lab has access to the course syllabus, various documents related to Anatomy and dissection, as well as the lab safety manual. During conversations with the BIOL 225 Spark, the AI responses are primarily generated via these reference documents.

The documents used for the RAG process vary per course. In addition, each Spark is also prompted to not complete assignments for students. Rather, each Spark is prompted to answer student questions Socratically, just as a human teaching assistant might guide a student during a face-to-face tutoring session. This deliberate approach fosters a deeper understanding of the subject matter, encouraging students to actively grapple with concepts rather than passively receiving answers. By prompting students to articulate their thought processes and identify their own knowledge gaps, the Spark aims to cultivate independent learning and problem-solving skills that extend beyond the immediate assignment. Ultimately, the goal is to empower students to become self-directed learners, capable of applying their knowledge to novel situations and developing a lifelong curiosity for learning.

LITERATURE

The integration of Artificial Intelligence (AI) in educational settings has gained significant attention in recent years, with numerous studies highlighting its potential to enhance learning outcomes and personalize educational experiences (Jung, 2024; Luckin et al., 2016; Maphoto et al., 2024; Zawacki-Richter et al., 2019). One of the primary roles AI plays in education is through intelligent tutoring systems (ITS) that provide real-time feedback and support to students, mimicking the assistance typically provided by human tutors (van den Berg, 2024). Our study contributes to this growing body of literature by focusing on the impact of an AI course assistant, Spark, on student academic outcomes in a fully online higher education environment.

AI COURSE ASSISTANTS IN HIGHER EDUCATION

The use of AI in higher education has evolved from adaptive learning platforms to more sophisticated tools like AI course assistants, which can support students by offering guidance, answering questions, and facilitating learning through Socratic methods. AI systems such as Spark, designed to provide specific, course-relevant assistance, have the potential to transform the educational landscape by making learning more accessible and personalized, especially in online learning environments (Baker & Smith, 2019). Several studies have demonstrated the positive impact of AI tools on student engagement and motivation, particularly in asynchronous learning contexts where real-time instructor interaction is limited (Jin et al., 2023; Maphoto et al., 2024).

AI's IMPACT ON ACADEMIC OUTCOMES

AI's potential to influence academic performance has been a focal point of recent research. Studies on intelligent tutoring systems and AI-driven feedback mechanisms have shown that these tools can improve student learning outcomes by providing personalized, timely, and targeted feedback (Chen et al., 2020). In particular, the use of AI course assistants has been found to enhance students' academic performance by supporting their understanding of course material and helping them navigate difficult concepts (Holstein et al., 2019). For example, a study by Wang and Zhao (2024) revealed that students using AI-powered tools in an online course demonstrated better academic outcomes compared to their peers who did not have access to such tools. These findings suggest that AI assistants can act as scalable solutions for improving educational outcomes, particularly in resource-constrained environments.

Frequency of AI Interaction and Student Outcomes

An emerging area of interest in the AI education literature is the relationship between the frequency of student interaction with AI tools and academic success. As noted by Holstein et al. (2019), students who frequently engage with AI tutoring systems tend to show greater improvements in performance compared to those with limited interaction. This suggests that the effectiveness of AI tools is not only dependent on their availability but also on how often and how deeply students engage with them. The current study adopts this framework by defining "utilization" of Spark as engaging with the AI assistant three or more times, based on the rationale that multiple interactions are necessary for students to receive meaningful support.

EQUITY AND ACCESSIBILITY CONSIDERATIONS

Another important consideration in the deployment of AI in educational settings is its impact on equity and accessibility. Research by Holmes et al. (2019) has highlighted concerns about the digital divide and whether AI tools can inadvertently reinforce existing disparities in access to educational resources (Jung, 2024; van den Berg, 2024). In online learning environments, where students may already face challenges related to technology access, AI systems must be designed and implemented in ways that promote inclusivity. Studies have shown that when

AI tools are equitably accessible, they can play a crucial role in leveling the playing field by providing all students with consistent and reliable support (Luckin et al., 2016). Therefore, when considering implementation of gen-AI tools, institutional stakeholders must actively engage in efforts to adopt and utilize AI solutions aimed at leveling the educational playingfield and address the unique needs of all students, especially those from marginalized and under-resourced populations.

To promote equitable access to Spark, the implementation process prioritized embedding the tool into courses in a way that would make it easy for students and faculty to utilize. There was also a focus on training both students and instructional faculty on best practice. To this end, Spark was fully integrated into LAPU's learning management system (LMS), eliminating the need for external software or additional logins. To support students in effectively utilizing the AI course assistant, two short instructional videos were developed, providing step-by-step guidance on Spark's functionality and best practices for engagement. By embedding these training resources directly into course materials, we aimed to ensure that students of all technological backgrounds could confidently use Spark to enhance their learning experience.

CHALLENGES AND ETHICAL CONCERNS

While AI-powered educational tools offer significant benefits, including personalized learning and real-time feedback, concerns persist regarding their ethical implications and unintended consequences.

Potential Biases in AI Models

AI-driven systems, including generative AI course assistants, rely on large datasets for training. However, these datasets can reflect biases present in the original sources, potentially leading to unfair or skewed responses (Holmes et al., 2019). Biases in AI-generated feedback may disproportionately affect certain student groups, particularly those from underrepresented backgrounds (Baker & Smith, 2019). Ensuring that AI course assistants provide equitable support requires careful monitoring and continuous refinement of AI training datasets. This necessitates not only technical adjustments to the algorithms, but also a critical examination of the underlying systems that contribute to the embedded biases in the first place.

Privacy and Data Security

Another concern is data privacy. AI tools often collect large amounts of student interaction data to enhance personalization and adapt responses. This raises concerns about how student data is stored, who has access to it, and whether students can opt out (Luckin et al., 2016). Universities must implement strict data governance policies to ensure that student information is protected and that AI-driven tools comply with data privacy regulations such as FERPA and GDPR. Furthermore, the potential for aggregated and anonymized student data to be used for research or commercial purposes necessitates transparent communication with students and clear institutional guidelines on data usage.

The Digital Divide and Access to AI Tools

Despite AI's potential to increase access to educational resources, disparities in access remain a significant challenge. Students from low-income backgrounds or those with limited digital literacy may struggle to engage with AI-powered learning tools (Jung, 2024; van den Berg, 2024). While AI can provide scalable support, it is not a substitute for human instruction, and institutions must ensure that AI tools do not widen existing educational inequalities (Holmes et al., 2019). Therefore, a comprehensive strategy must integrate AI tools with robust human support systems and targeted digital literacy programs to bridge the access gaps and ensure equitable learning outcomes for all students.

By addressing these challenges, institutions can take proactive steps to mitigate risks and maximize the benefits of AI-driven education, ensuring that such technologies serve all students equitably. This necessitates not only the development of robust ethical frameworks but also the continuous evaluation and adaptation of AI systems in response to evolving student needs and societal expectations.

SUMMARY

The literature underscores the potential of AI course assistants like Spark to positively influence student learning outcomes, particularly in online and asynchronous learning environments and among diverse populations. By offering personalized feedback, supporting student engagement, and fostering equitable access to educational resources, AI tools can significantly enhance the learning experience. However, the effectiveness of these tools is closely linked to the frequency of student interaction, highlighting the need for institutions to encourage regular engagement with AI systems to maximize their impact.

RESEARCH QUESTIONS

RQ1: What is the impact of an AI course assistant on a student's GPA?

HYPOTHESES

Null Hypothesis (H_0): There is no significant difference in the GPAs between students who utilized the AI course assistant three or more times and those who utilized it fewer than three times.

Alternative Hypothesis (H_1): Students who utilized the AI course assistant three or more times will have significantly higher GPAs compared to those who utilized it fewer than three times.

METHODS

Ethical approval for this study was obtained from the Institutional Review Board (IRB) at Los Angeles Pacific University. All research procedures adhered to institutional and ethical guidelines for studies involving human participants.

RESEARCH DESIGN MODEL

This study employs a quasi-experimental, non-equivalent-groups design to examine the impact of Spark, an AI course assistant, on student academic outcomes. This design was selected to ensure that all students had access to Spark while still allowing for comparative analysis between users and non-users. The study builds upon an initial randomized control trial (RCT) pilot (Hanshaw et al., 2024) that demonstrated significant differences in GPA outcomes, motivation, and efficacy between students who utilized Spark and those who did not throughout a given course. Given the positive results of the pilot study, a quasi-experimental approach was chosen to allow for equitable access while rigorously evaluating Spark's impact on student learning.

The pilot study conducted by Hanshaw et al. (2024) revealed a significant difference in GPA outcomes between users and non-users of Spark, with a strong effect size, underscoring the potential of this tool to enhance academic achievement. Due to these significant benefits, we decided against limiting access to the AI course assistant in the broader study. Instead, we opted for a quasi-experimental design, allowing us to compare the outcomes of students who voluntarily used the assistant with those who did not. This approach preserves equitable access to this potentially transformative educational resource while rigorously assessing its effectiveness.

Non-Equivalent-Groups Comparison

- **Groups Compared:** The study compares the GPA outcomes of two groups: those who used Spark in a given course (treatment group) and those who did not (control group).
- **Potential Confounds:** Since the groups were not randomly assigned, there may be pre-existing differences between the groups that could influence the outcomes. We control for these confounding variables via the large sample size of n = 2,090 student-course combinations and propensity score matching (PSM).

Implications

- **Internal Validity:** While quasi-experiments can offer valuable insights, they may have lower internal validity compared to true experiments due to the potential for confounding variables. For example, students who chose to use the AI assistant might differ in motivation or academic ability from those who did not, which could influence their GPA independent of the assistant's impact.

- **Analysis Approach:** Given the non-random assignment, propensity score matching is used to control for confounding variables and to better isolate the effect of the AI course assistant

To enhance internal validity in our quasi-experimental design, we utilized permutation analysis. Permutation analysis is a non-parametric statistical method that involves repeatedly shuffling the data to create a distribution of the test statistic under the null hypothesis (Good, 2005). This method helps control for confounding variables by comparing the observed test statistic to the distribution of test statistics generated from the permuted data, thereby providing a more robust assessment of the treatment effect (Anderson, 2001). By using permutation analysis, we can mitigate the impact of pre-existing differences between the treatment and control groups, as it does not rely on the assumption of random assignment (Edgington & Onghena, 2007). This approach enhances the internal validity of our study by ensuring that the observed effects are less likely to be due to confounding variables and more likely to be attributable to the AI course assistant itself.

SETTING AND CONTEXT

The implementation of Spark at LAPU was driven by the institution's commitment to enhancing student engagement and academic success in an asynchronous learning environment. Given the challenges of online education, such as limited real-time interaction and varying levels of student support, LAPU sought to integrate AI course assistants to provide personalized, on-demand academic guidance and improve learning outcomes.

Implementation of Spark

- Spark is a GenAI course assistant integrated into LAPU's learning management system (LMS) via the Nectir platform.
- Each course had a dedicated Spark assistant trained using Retrieval-Augmented Generation (RAG), allowing it to access course-specific documents such as syllabi, lecture materials, and assignments.
- Spark was designed to engage students Socratically, mirroring a human teaching assistant by guiding students rather than completing assignments for them.
- Instructors and students accessed Spark through a single sign-on (SSO) link embedded in each weekly overview of their courses.

SAMPLING

Population

The population sample for this study comprised 2,090 unique student-course combinations and 1,338 unique graduate and undergraduate students enrolled across 99 different courses, which were taught in 225 sections. The gender distribution within this sample is predominantly female, with 81% identifying as female, 17.4% as male, and 1.5% choosing not to state their gender.

For ethnicity, the sample is much more diverse: 46% of the students identified as Hispanic of any race, 17% as Black or African American, 20% as White, 7% as Asian, 5% as two or more races, 2% as race and ethnicity unknown, 1% as Native Hawaiian or Other Pacific Islander, and fewer than 1% as American Indian or Alaska Native. The inclusion of students from a wide range of courses and sections enhances the generalizability of the findings, providing a robust foundation for examining the impact of the AI course assistant across different demographic groups and academic contexts.

The inclusion of a diverse range of courses in this study strengthens the generalizability of the findings. The dataset spans 99 distinct courses across multiple disciplines, ensuring that the impact of Spark is evaluated across different subject areas and instructional formats. This broad representation enhances confidence that the observed effects are not limited to a specific course type or field of study.

DATA COLLECTION TOOLS

Usage Report Provided by Nectir

Nectir supplied a detailed report capturing conversation and message data on a per-user basis. This report included only student-specific usage information and excluded any AI assistant-generated content. Table 1 outlines the columns and their descriptions within the report:

A description of the columns provided in the report is below.

Table 1 Description of columns.

COLUMN	DATA TYPE	DESCRIPTION
User ID	string	Unique ID for each student
User email	string	Student email
Assistant name	string	AI assistant name (specific to a single academic course)
Conversation ID	string	Unique conversation ID
Conversation Created At	timestamp	Timestamp indicating when first message of conversation was created (UTC)
Message ID	string	Unique message ID
Message Created At (UTC)	timestamp	Timestamp indicating when message was created (UTC)
Message Created At (PT)	timestamp	Timestamp indicating when message was created (Pacific Time)

Grade Report

In addition to the usage report provided by Nectir, the Institutional Research (IR) team provided a comprehensive grade report. This report contained detailed data on each student, their respective instructors, the grades they earned, and the courses they were enrolled in. By correlating this grade data with the usage of Spark, we were able to perform a robust analysis of the AI course assistant's impact on academic outcomes.

The grade report enabled us to control for confounding factors, such as teaching methods and course difficulty, by including instructor and course data in our analysis. This integration of data from the Nectir usage report and the IR grade report further strengthened the validity of our findings, offering a clearer understanding of how Spark influenced academic achievement.

IDENTIFYING USE OF AI ASSISTANT

To identify a student's use of Spark, the AI course assistant, the usage report from Nectir was used.

Students were determined to have used Spark if they had three or more interactions with the AI course assistant. This criterion was established due to the nature of Spark's functionality. The first interaction typically involves initiating the assistant, and the second interaction is a response to the student's initial query. Meaningful data and assistance are provided after the second interaction. Therefore, students who abandoned Spark after two or fewer interactions were defined as not having used Spark to assist them in their coursework.

STATISTICAL ANALYSES

Due to the ordinal nature of GPA data and the non-random assignment of students to AI course assistant usage, nonparametric statistical methods were selected for analysis. These methods do not assume normality and are well-suited for comparing groups with potential confounding factors. Additionally, propensity score matching was used to improve group comparability by balancing key covariates, such as age, gender, and ethnicity. The following statistical analyses were conducted to evaluate the impact of Spark usage on student grade outcomes.

Mann-Whitney U-Test

- Used to compare GPA distributions between Spark users (treatment) and non-users (control).
- Specifically, we conducted the one-side version of the test.
- We measured effect size for the Mann-Whitney U-test using rank-biserial correlation.

Permutation Test for Mean Differences (on all data, before Propensity Score Matching)

- Used to assess whether observed GPA differences between Spark users and non-users could be attributed to chance.
- Conducted by randomly reshuffling group labels 10,000 times to generate a null distribution of mean differences.

Propensity Score Matching (PSM)

- Controls for confounding variables (age, gender, ethnicity) by matching Spark users with non-users on these characteristics.
- Post-matching balance check: Standardized Mean Difference (SMD) for key variables improved to <0.1 after filtering for a 10-year age difference.

Wilcoxon Signed-Rank Test (on PSM Matched Pairs)

- Examines paired GPA differences within PSM-matched student-course pairs.
- We measured effect size for the Wilcoxon signed-rank test using Pearson's correlation coefficient.

Permutation Test for Mean Differences (on PSM Matched Pairs)

- Tests whether GPA differences are due to chance by reshuffling data 10,000 times. Table 2 below summarizes the statistical methods and purposes.

Table 2 Summary of Statistical Methods.

STATISTICAL METHOD	PURPOSE
Mann-Whitney U-Test	Compares GPA distributions between groups
Permutation Test (using all data)	Validates GPA differences
Propensity Score Matching	Controls for confounders
Wilcoxon Signed-Rank Test	Examines GPA changes in matched pairs
Permutation Test (on PSM matched pairs)	Validates GPA differences

Together, these four nonparametric methods provide a rigorous and robust approach to validating the effect of AI assistant usage on student GPA outcomes, offering both statistical significance and meaningful measures of effect size to interpret the results.

RELIABILITY OF MEASURES

Reliability

The usage report from Nectir and grade data from the IR team offer highly reliable system-generated metrics. These automated tools provide consistent and accurate data on both student performance and their usage of Spark, reducing the potential for human error. The study's large sample size of 1,338 students across 225 sections further strengthens the reliability of the findings by minimizing the impact of anomalies or random variation (Good, 2005; Anderson, 2001).

The combined use of objective performance metrics and system-generated usage data ensures that the study's findings are both valid and reliable.

FINDINGS

To assess the impact of AI course assistant usage on GPA outcomes, we analyzed GPA differences between treatment (Spark users) and control (non-users) groups using multiple statistical methods. These tests confirmed whether Spark usage was associated with significantly higher GPAs, both before and after controlling for potential confounders.

DESCRIPTIVE STATISTICS

The dataset consists of 2,090 student-course combinations from 1,338 unique students across 99 distinct courses. The control group (students who used Spark fewer than three times in a course) includes 1,777 student-course combinations with a mean GPA of 3.05 (SD = 1.22). The treatment group (students who used Spark at least three times in a course) comprises 313 student-course combinations, with a mean GPA of 3.28 (SD = 1.08).

Across the entire dataset, the overall mean GPA is 3.08, with a standard deviation of 1.21. This serves as a baseline for comparing the treatment and control groups, helping to assess the impact of AI assistant usage on student performance across different courses.

As shown in Figure 1, the GPA distribution for the treatment group is more concentrated around higher values, with the 75th percentile aligning with the maximum GPA of 4.0, highlighting a shift toward stronger academic performance among Spark users.

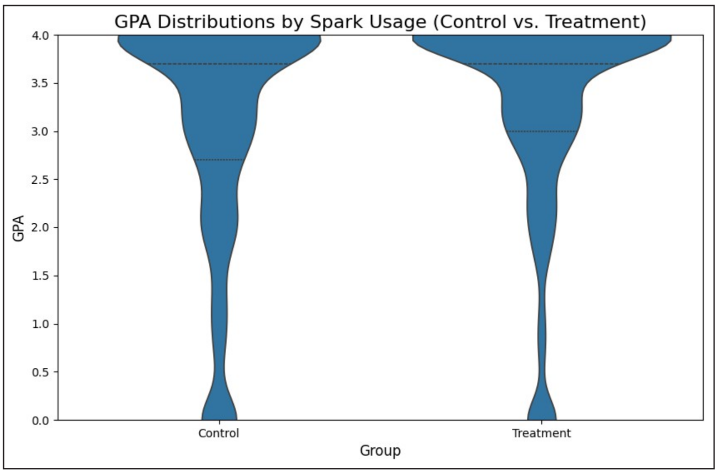

Figure 1 GPA Distributions for Control (Did Not Use Spark) and Treatment (Did Use Spark) Groups.

GPA DIFFERENCES BEFORE PROPENSITY SCORE MATCHING

To evaluate the difference in GPA outcomes between groups, we conducted the Mann-Whitney U-Test and a permutation test for mean differences.

- Mann-Whitney U-Test: Spark users had significantly higher GPAs than non-users (U = 307562.5, p = 0.00185), with a moderate-to-strong effect size of 0.553.
- Permutation Test for Mean Differences: The observed mean difference was 0.226 GPA points (p = 0.001), reinforcing the U-test findings.

Both tests indicate that student-course combinations where Spark was utilized had, on average, significantly higher GPAs than those in the control group.

GPA DIFFERENCES AFTER PROPENSITY SCORE MATCHING

To account for potential confounding variables, we performed propensity score matching (PSM) to create comparable student-course pairs based on age, gender, and ethnicity within the same course and term. The final analysis included 229 matched pairs after filtering for large age differences. Full details of the matching process, including standardized mean differences (SMDs) before and after matching, are provided in Appendix B.

With a well-balanced dataset, we assessed GPA differences using the Wilcoxon signed-rank test and a permutation test for paired mean differences.

- Wilcoxon Signed-Rank Test: Spark users had significantly higher GPAs than their matched counterparts (W = 10027.5, p = 0.0002583), with an effect size of r = 0.262, indicating a small-to-moderate positive relationship between Spark usage and GPA.
- Permutation Test for Matched Mean Differences: The observed mean difference was 0.377 GPA points (p = 0.0005), further confirming that GPA improvements among Spark users were unlikely due to chance.

Figure 2 visually represents the permutation test results, illustrating the null distribution of GPA differences from 10,000 randomized permutations. The observed difference of 0.377 GPA points lies in the extreme tail of the distribution, underscoring its statistical significance.

DOI: 10.55982/openpraxis.17.2.772

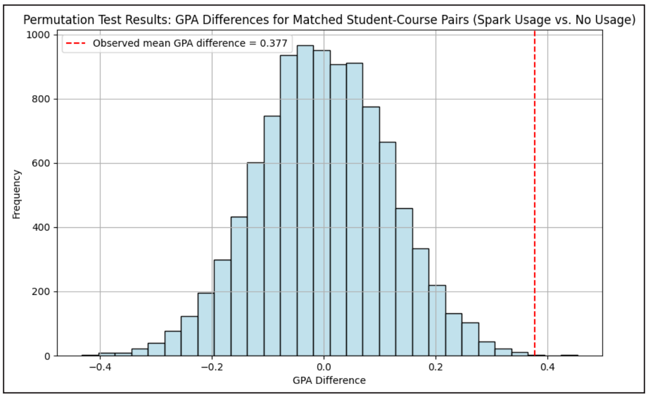

Figure 2 Distribution of Permuted GPA Differences for Matched Student-Course Pairs.

HYPOTHESIS TESTING

To assess the relationship between AI course assistant usage and student GPA, we examined whether students who interacted with Spark three or more times demonstrated significantly higher GPAs compared to those who used it fewer than three times. The null hypothesis (H_{01}) posited that no significant difference existed in GPA outcomes between these two groups, while the alternative hypothesis (H_{11}) proposed that frequent Spark users would achieve significantly higher GPAs.

A series of statistical tests were conducted to evaluate this hypothesis. First, a Mann-Whitney U-test was performed to compare the GPA distributions of Spark users and non-users. The results indicated a statistically significant difference (U = 307562.5, p = 0.00185, effect size = 0.553), suggesting that students who engaged with Spark more frequently attained higher GPAs than those who did not.

To further validate this finding, we conducted a permutation test for mean differences before applying propensity score matching. The test revealed a significant difference in GPA between the two groups (mean difference = 0.226, p = 0.001), reinforcing the conclusion that the observed GPA differences were unlikely to be due to random variation.

To control for potential confounding factors such as age, gender, and ethnicity, propensity score matching (PSM) was applied to create statistically comparable groups. A subsequent Wilcoxon signed-rank test performed on the matched dataset confirmed that Spark users maintained significantly higher GPAs than their non-user counterparts (W = 10027.5, p = 0.0002583, effect size = 0.262).

Finally, a permutation test for matched mean differences further substantiated these findings, yielding an observed mean GPA difference of 0.377 (p = 0.0005). Even after accounting for demographic and academic factors, the GPA advantage for Spark users remained statistically significant.

Taken together, these results provide strong evidence that interaction with the AI course assistant positively impacts student GPA. Given the consistent findings across multiple statistical analyses, we reject the null hypothesis (H_{01}) and conclude that students who engage with Spark at least three times experience measurable improvements in academic performance.

SUMMARY OF FINDINGS

Across multiple statistical approaches, our results consistently indicate that student-course combinations where Spark was utilized had significantly higher GPAs compared to those where it was not. This relationship persists even after controlling for potential confounders through propensity score matching. These findings suggest that AI course assistants may contribute to improved academic performance.

ANALYSIS	TEST STATISTIC	p-VALUE	EFFECT SIZE	MEAN DIFFERENCE
Mann-Whitney U-Test (Before PSM)	U = 307562.5	0.00185	0.553	–
Permutation Test (Before PSM)	–	0.001	–	0.226
Wilcoxon Signed-Rank Test (After PSM)	W = 10027.5	0.0002583	0.262	–
Permutation Test (After PSM)	–	0.0005	–	0.377

DISCUSSION

This study demonstrates that students who engaged with Spark, the AI course assistant, achieved significantly higher GPAs than those who did not. These findings align with prior research on AI-driven learning tools (Hanshaw et al., 2024; Wang & Zhoa, 2024), reinforcing the potential for AI assistants to enhance academic performance in online learning environments. The statistical analyses, including propensity score matching and permutation testing, provide robust evidence of this relationship.

These findings further validate the role of AI course assistants in facilitating meaningful student learning experiences. Prior research has similarly demonstrated that AI-enhanced academic support improves student success (Maphoto et al., 2024; van den Berg, 2024). Because these assistants are designed to employ a Socratic inquiry method rather than merely providing direct answers, they actively engage students in critical thinking and deeper cognitive processing. This interactive approach fosters a more reflective and inquiry-driven learning process, reinforcing conceptual understanding and promoting self-directed learning. The results suggest that AI-driven Socratic engagement contributes to improved academic outcomes by guiding students toward constructing their own knowledge rather than passively receiving information.

One key takeaway is that frequent interaction with Spark correlated with improved student outcomes, supporting prior findings on the importance of active engagement with AI tutors (Holstein et al., 2019). However, the study also highlights equity considerations, as students who did not use Spark may have faced barriers unrelated to academic ability. Future research should explore the reasons for non-use and develop strategies to increase engagement with AI course assistants.

While the findings of this study highlight the positive impact of AI course assistants on student academic performance, it is important to acknowledge potential challenges associated with AI in education. Issues such as AI bias, privacy concerns, and disparities in access (Holmes et al., 2019; Jung, 2024) require continued attention to ensure that AI tools serve all students equitably. Future research should explore strategies for addressing these concerns, including improving AI transparency, ensuring data privacy compliance, and providing additional support for students who may struggle to engage with AI-powered learning tools.

INTERPRETATIONS OF FINDINGS

One key takeaway is that frequent interaction with Spark correlated with improved student outcomes, supporting previous findings on active engagement with AI tutors (Holstein et al., 2019). The structured, Socratic-style feedback provided by Spark may contribute to deeper learning, enabling students to refine their understanding of course concepts and improve their academic performance.

Additionally, the frequency of AI interactions played a role in academic performance, reinforcing the importance of consistent engagement with AI-based tools. This suggests that students who actively seek AI support may be developing stronger self-regulated learning strategies, a key factor in academic success.

However, while these results are statistically significant, they do not establish causation. The positive outcomes observed may be influenced by student motivation, prior academic performance, or self-selection bias. Students who are already proactive in their studies may be more likely to engage with Spark, meaning that their improved GPA may not be solely attributable to AI assistance.

Moreover, the study highlights equity considerations, as students who did not use Spark may have faced barriers unrelated to academic ability. These barriers could include technological access, AI literacy, or personal learning preferences. Understanding why some students chose not to engage with Spark is critical for ensuring that AI-driven learning tools are equitable and accessible. Future research should investigate these barriers and develop strategies to increase engagement with AI course assistants, particularly for underserved student populations.

LIMITATIONS

Despite the promising results, this study has several limitations. First, while the quasi-experimental design controlled for confounding factors through propensity score matching, it does not establish causal relationships. Future research could employ randomized controlled trials (RCTs) to strengthen causal claims regarding AI's impact on academic performance.

Second, the study relies on GPA as the primary outcome measure. While GPA provides a useful benchmark for academic performance, it does not capture deeper learning aspects, such as critical thinking, problem-solving, or conceptual mastery. Future studies should incorporate qualitative assessments to gain a richer understanding of students' experiences with AI course assistants. Furthermore, the study's focus on quantitative data limits its ability to explore the nuanced impact of AI on student motivation and engagement, which are crucial factors in effective learning. Investigating these factors through surveys, interviews, and observational studies could provide valuable insights into the holistic effects of AI integration.

Addressing these limitations in future research will provide a more comprehensive perspective on how AI tools influence learning beyond grade outcomes.

Moving forward, institutions implementing AI course assistants should focus on encouraging student interaction with these tools. Further research is needed to explore the long-term impact of AI-assisted learning on retention, motivation, and skill development. Additionally, creating feedback loops that allow students to provide input on the AI's effectiveness and usability is crucial for continuous improvement and adaptation. Moreover, longitudinal studies that track student progress across multiple semesters or academic years are necessary to fully understand the sustained effects of AI integration on educational outcomes.

FUTURE RESEARCH DIRECTIONS

Based on the findings, several areas for future research emerge. First, further exploration is needed into why some students chose not to use Spark. Understanding these barriers to adoption could inform strategies for increasing engagement with AI tools. Additionally, future studies should examine the long-term impact of AI course assistants, assessing their effects on student retention, motivation, and learning behaviors beyond GPA. Moreover, further research may delve into the efficacy of different pedagogical approaches when integrating AI, determining which teaching strategies best leverage AI's capabilities while maintaining human connection. Investigating the impact of AI on the development of metacognitive skills and self-regulated learning is also crucial, as these are foundational for lifelong learning. Finally, exploring the ethical implications of AI-driven personalization, particularly regarding data privacy and potential algorithmic biases, should be a central focus of future investigations.

CONCLUSION

The findings from this study demonstrate that the utilization of AI course assistants, specifically Spark, is associated with improved academic outcomes, as measured by GPAs earned in courses by students who utilized the tool. Students who engaged with Spark three or more times exhibited significantly higher grade outcomes compared to those who did not, even after controlling for potential confounding variables through propensity score matching. These results align with the growing body of literature supporting the efficacy of AI-driven educational tools in enhancing student performance (Hanshaw et al., 2024). The use of nonparametric statistical methods, such as the Mann-Whitney U-test and Wilcoxon signed-rank test, ensured that the analyses were robust against violations of normality assumptions, further validating the positive impact of AI assistants like Spark on academic achievement.

The implications of this study are significant for institutions of higher education, particularly as AI continues to shape personalized education and expand access to academic support (Jung, 2024; Maphoto et al., 2024; van den Berg, 2024). By demonstrating that AI course assistants can positively influence GPA, this research suggests that integrating such tools into the digital classroom can offer substantial benefits in terms of academic success. Furthermore, the study highlights that increased engagement with the AI assistant—operationalized as three or more interactions—is key to realizing these benefits. This finding implies that institutions should not only make AI tools available but also actively encourage students to engage with them consistently to maximize their effectiveness.

From a pedagogical perspective, the results suggest that AI assistants can play a pivotal role in providing personalized and timely feedback, which may be particularly useful in asynchronous learning environments where instructor interaction is limited. Additionally, the ability of AI tools like Spark to function as a Socratic guide, rather than merely providing answers, supports deeper learning and critical thinking among students.

AI course assistants like Spark have the potential to significantly enhance student learning outcomes, particularly in online and asynchronous environments. As educational technology continues to evolve, institutions should consider AI tools as an integral part of their pedagogical toolkit, leveraging them not only to improve academic performance but also to foster a more personalized and supportive learning experience.

APPENDIX A

CODE AND DOI ARCHIVE

Code availability

The code supporting the findings of this study is openly available in a public GitHub repository. This repository contains all scripts and resources required to replicate the analyses presented in this paper, including data preprocessing, statistical testing, and visualizations. Detailed instructions for setting up the environment and running the analyses are provided in the repository's README.md file.

Github repository

https://github.com/Nectir/nectir-ai-research

DOI archive

To facilitate reproducibility and long-term access, the code associated with this study has been archived through Zenodo. The archived version, which corresponds to the version used in this research, can be accessed via the following DOI:

Zenodo archive

https://doi.org/10.5281/zenodo.13883503

APPENDIX B

PROPENSITY SCORE MATCHING (PSM) AND COVARIATE BALANCE

Overview of propensity score matching

To control for potential confounding variables, we applied propensity score matching (PSM) to create comparable student-course pairs. Matching was conducted within the same course and academic term based on the following covariates:

- Age at entry
- Gender
- Ethnicity

A one-to-one nearest neighbor matching algorithm was used, with a caliper restriction to ensure close matches. Students without gender data (n = 33) were excluded from the PSM process.

After matching, we performed a covariate balance check by calculating standardized mean differences (SMDs) between treatment and control groups. A SMD below 0.1 is generally considered an indicator of good balance.

Covariate balance before and after matching

Table 3 presents the SMDs for each covariate before matching, after matching, and after applying an additional refinement that filtered out matched pairs where the age difference exceeded 10 years.

COVARIATE	SMD BEFORE MATCHING	SMD AFTER MATCHING	SMD AFTER AGE DIFFERENCE FILTERING
Age	0.3886	0.4333	0.0749
Gender (Male)	–0.1332	0.0193	–0.041
Ethnicity (Asian)	0.0214	–0.0125	–0.0185
Ethnicity (Black or African American)	–0.0532	0.0453	0.0125
Ethnicity (Hispanics of any race)	0.0868	0.0195	–0.044
Ethnicity (Native Hawaiian or Other Pacific Islander)	0.0137	–0.0271	–0.0315
Ethnicity (Race and Ethnicity unknown)	–0.0786	–0.0514	0.0939
Ethnicity (Two or more races)	–0.108	–0.0969	0.1426
Ethnicity (White)	0.0171	0.0326	0

Table 3 Standardized Mean Differences at Different Stages of Propensity Score Matching.

Key Takeaways from Table 3

- Before matching, age at entry had the largest imbalance (SMD = 0.3886), while ethnicity and gender were closer to balance.
- After matching, age imbalance slightly increased (SMD = 0.4333), necessitating further refinement.
- After filtering out pairs with an age difference >10 years, the age imbalance was resolved (SMD = 0.0749), and all other covariates remained within an acceptable range (SMD < 0.1), except for Ethnicity (Two or More Races) = 0.1426.
- Given that ethnicity (Two or More Races) represents a small portion of the dataset, the overall balance was deemed acceptable for further analysis.

Visual Representation of Covariant Balance

To complement Table 3, Figure 3 presents the absolute SMDs at different stages of matching, highlighting the improvements in balance after filtering for age differences.

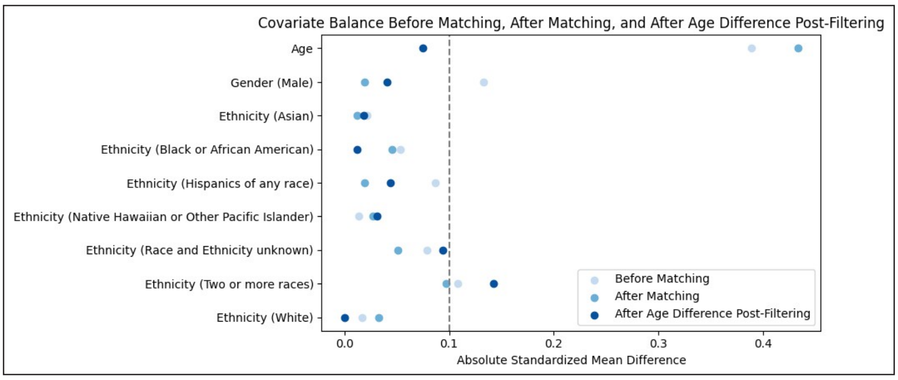

Figure 3 Absolute Standardized Mean Differences at Different Stages of Propensity Score Matching.

Final matched sample size

After PSM and age-difference filtering, the final dataset used in subsequent analyses consisted of:

- 229 matched student-course pairs (458 total observations).
- Each pair included one Spark user and one non-user from the same course and term.
- Key covariates were balanced, mitigating the risk of confounding effects.

Conclusion

This refinement process ensured that treatment and control groups were comparable, allowing for a more rigorous assessment of GPA differences while minimizing biases from demographic or course-related factors.

DOI: 10.55982/openpraxis.17.2.772

DATA ACCESSIBILITY STATEMENT

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

SUSTAINABLE DEVELOPMENT GOALS (SDGs)

This study is linked to the following SDG(s): Quality education (SDG 4).

ETHICS AND CONSENT

Ethical approval was obtained through the Institutional Review Board (IRB) at Los Angeles Pacific University.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS (CRediT)

George Hanshaw: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Supervision, Validation, Visualization. Writing – original draft, Writing – review & editing. Ethan Wicker: Conceptualization, Data curation, Formal Analysis, Investigation, Project administration, Writing – original draft, Writing – review & editing. Kristen Denlinger: Conceptualization, Data curation, Formal Analysis, Investigation, Project administration, Writing – original draft, Writing – review & editing. All authors have read and agreed to the published version of the manuscript OR The author has read and agreed to the published version of the manuscript.

AUTHOR NOTES

Based on *Academic Integrity and Transparency in AI-assisted Research and Specification Framework* (Bozkurt, 2024), the authors of this paper acknowledge that the paper was partially reviewed and edited by Nectir AI Writing assistant (September 2024), ChatGPT (September 2024), and Grammarly (September 2024), complementing the human editorial process. The human authors critically assessed and validated the content to maintain academic rigor. The authors also assessed and addressed potential biases inherent in the AI-generated content. The final version of the paper is the sole responsibility of the human authors.

AUTHOR AFFILIATIONS

George Hanshaw orcid.org/0000-0002-5986-3965
Los Angeles Pacific University, US

Ethan Wicker orcid.org/0009-0007-2327-1590
Nectir, US

Kristen Denlinger orcid.org/0009-0007-3271-8489
Los Angeles Pacific University, US

REFERENCES

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences, 58*(3), 626–639. https://doi.org/10.1139/f01-004

Baker, T., & **Smith, L.** (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges.* Nesta. https://www.nesta.org.uk/report/education-rebooted/

Bozkurt, A. (2024). GenAI et al.: Cocreation, authorship, ownership, academic ethics and integrity in a time of generative AI. *Open Praxis, 16*(1), 1–10. https://doi.org/10.55982/openpraxis.16.1.654

Chen, X., Xie, H., Zou, D., & **Hwang, G. J.** (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers & Education, 146*, 103751. https://doi.org/10.1016/j.caeai.2020.100002

Edgington, E. S., & **Onghena, P.** (2007). *Randomization tests* (4th ed.). CRC Press. https://doi.org/10.1201/9781420011814

Good, P. I. (2005). *Permutation, parametric, and bootstrap tests of hypotheses* (3rd ed.). Springer.

Hanshaw, G., Vance, J., & **Brewer, C.** (2024). Exploring the effectiveness of AI course assistants on the student learning experience. *Open Praxis, 16*(4), 627–644. https://doi.org/10.55982/openpraxis.16.4.719

Holmes, W., Bialik, M., & **Fadel, C.** (2019). *Artificial intelligence in education: Promises and implications for teaching and learning.* Center for Curriculum Redesign.

Holstein, K., McLaren, B. M., & **Aleven, V.** (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *Journal of Learning Analytics, 6*(2), 27–52. https://doi.org/10.18608/jla.2019.62.3

Jin, S.-H., Im, K., Yoo, M., Roll, I., & **Seo, K.** (2023). Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education, 20*, 37. https://doi.org/10.1186/s41239-023-00406-5

Jung, I. (2024). Personalized education for all: The future of open universities. *Open Praxis, 16*(1), 24–36. https://doi.org/10.55982/openpraxis.16.1.612

Luckin, R., Holmes, W., Griffiths, M., & **Forcier, L. B.** (2016). *Intelligence unleashed: An argument for AI in education.* Pearson Education.

Maphoto, K. B., Sevnarayan, K., Mohale, N. E., Suliman, Z., Ntsopi, T. J., & **Mokoena, D.** (2024). Advancing students' academic excellence in distance education: Exploring the potential of generative AI integration to improve academic writing skills. *Open Praxis, 16*(2), 142–159. https://doi.org/10.55982/openpraxis.16.2.649

van den Berg, G. (2024). Generative AI and Educators: Partnering in Using Open Digital Content for Transforming Education. *Open Praxis, 16*(2), 130–141. https://doi.org/10.55982/openpraxis.16.2.640

Wang, L., & **Zhao, M.** (2024). Can artificial intelligence technology promote the improvement of student learning outcomes? A meta-analysis based on 50 experimental and quasi-experimental studies. *Proceedings of the 3rd International Conference on Educational Innovation and Multimedia Technology (EIMT 2024)*, 29–31. https://doi.org/10.4108/eai.29-3-2024.2347685

Zawacki-Richter, O., Marín, V. I., Bond, M., & **Gouverneur, F.** (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

TO CITE THIS ARTICLE:
Hanshaw, G., Wicker, E., & Denlinger, K. (2025). An Experimental Study of the Effect of AI Course Assistants on Student Grade Outcomes. *Open Praxis*, *17*(2), pp. 270–285. DOI: https://doi.org/10.55982/openpraxis.17.2.772

Submitted: 04 October 2024
Accepted: 05 March 2025
Published: 10 July 2025