Generative AI Tutors and Student Engagement in Introductory Computing Courses

Sample Student

Master's Thesis

15th October, 2026

Citation style: APA 7th Edition

Declaration

This is a sample produced by ThesisAI 2.0 to show what its output looks like. The dataset analysed here is synthetic demonstration data generated for this sample, so the results do not describe real students or a real study. Do not cite it.

The author name is a placeholder.

Abstract

Generative artificial intelligence increasingly permeates higher education, yet empirical evidence comparing its efficacy to human instruction in introductory computing remains limited. This gap matters because first-year students face high cognitive loads and heterogeneous prior experience, demanding adaptive support that traditional lectures may not provide. The study aimed to determine how students taught with a generative AI tutor differ from those taught by a human instructor regarding engagement, quiz performance, and satisfaction. A quasi-experimental, between-subjects design was employed with 124 first-year students enrolled in introductory programming and data structures courses. Participants were divided equally into an AI tutor condition and a human instructor condition. Data on engagement, quiz scores, weekly session attendance, and satisfaction were collected via standardized instruments and analyzed using Welch's t-tests, Pearson correlations, and ordinary least squares regression. Results indicated that the AI tutor group achieved significantly higher engagement scores than the human instructor group (M = 75.35 vs. M = 66.75; d = 0.98). Quiz averages were also statistically higher for the AI cohort (M = 72.23 vs. M = 68.81; d = 0.48). Satisfaction ratings showed a strong association with teaching modality (Cramér's V = .41), favoring the AI condition. Regression analysis revealed that engagement strongly predicted quiz performance (r = .66), while the instructional group itself did not significantly predict outcomes when controlling for engagement. These findings suggest that generative AI tutors enhance student involvement and perceived support more effectively than traditional instruction in this context. The performance advantage appears mediated by increased engagement rather than superior content delivery alone. Practically, these results advocate for blended pedagogical models where AI handles routine feedback and scaffolding, allowing human instructors to focus on complex mentorship and fostering relatedness. This contributes to educational theory by clarifying the mechanisms through which AI supports early-stage learning autonomy.

Keywords: generative AI tutor; student engagement; introductory computing courses; quiz performance; satisfaction; human instructor; first-year students

Contents

Declaration 2
Abstract 3
1 Introduction 7
1.1 Context of the Study 7
1.2 Problem Statement 8
1.3 Aim of the Study 9
1.4 Scope and Limitations 9
1.5 Significance of the Study 10
1.6 Overview of Chapters 10
2 Literature Review 12
3 Methodology 16
3.1 Research Design 16
3.2 Participants and Sampling 17
3.3 Intervention Procedures 17
3.4 Data Collection Instruments 18
3.5 Analytical Strategy 19
4 Results/Analysis 21
5 Discussion 29
6 Conclusion 32
6.1 Synthesis of Findings and Theoretical Implications 32
6.2 Mechanisms of Engagement and Performance 33
6.3 Satisfaction Dynamics and Human-AI Collaboration 34
6.4 Limitations and Boundary Conditions 34
6.5 Future Research Agenda 35
References 37
A descriptives: engagement_score, quiz_avg, weekly_sessions, satisfaction 39
B t-test: engagement_score by group 40
C t-test: quiz_avg by group 41
D correlation: engagement_score, quiz_avg 42
E regression: quiz_avg ~ engagement_score + weekly_sessions + group 43
F crosstab: group × satisfaction 44

List of Figures

Figure 4.1 Means with standard deviation: engagement_score 71.05 ± 9.72; quiz_avg 70.52 ± 7.34; weekly_sessions 3.73 ± 1.48; satisfaction 3.60 ± 0.84. Error bars show one standard deviation. 22
Figure 4.2 engagement_score by group: ai_tutor 75.35 ± 8.85; human_instructor 66.75 ± 8.64. Error bars show one standard deviation. 23
Figure 4.3 quiz_avg by group: ai_tutor 72.23 ± 7.13; human_instructor 68.81 ± 7.20. Error bars show one standard deviation. 24
Figure 4.4 quiz_avg against engagement_score, n = 124. Fitted line: quiz_avg = 35.02 + 0.50 × engagement_score. 25
Figure 4.5 Observed quiz_avg against Fitted quiz_avg, n = 124. Fitted line: Observed quiz_avg = 0.00 + 1.00 × Fitted quiz_avg. 26
Figure 4.6 group × satisfaction (counts): ai_tutor / 4 34.00; ai_tutor / 3 15.00; ai_tutor / 5 12.00; ai_tutor / 2 1.00; human_instructor / 3 30.00; human_instructor / 4 17.00; human_instructor / 2 10.00; human_instructor / 5 5.00. 27

List of Tables

Table 4.1 Descriptive statistics 21
Table 4.2 Group statistics 23
Table 4.3 Group statistics 24
Table 4.4 Correlations 25
Table 4.5 Coefficients 26
Table 4.6 Crosstab (n, row %) 27

1 Introduction

1.1 Context of the Study

Generative artificial intelligence has rapidly entered higher education, particularly within introductory computing courses where students require immediate feedback on complex problem-solving tasks. The integration of these tools promises to transform traditional instructional models by offering personalized support at scale (Beimel et al., 2025). Students in first-year programming courses often face high cognitive loads and heterogeneous prior experience levels, creating a demand for adaptive tutoring systems that can scaffold learning effectively (Jacobs et al., 2025). Recent empirical work suggests that generative AI tools can enhance autonomy through personalized pacing and increased accessibility, potentially altering how students interact with course material (Beimel et al., 2025). However, the extent to which these digital tutors replicate or surpass the engagement and performance outcomes associated with human instruction remains a subject of active debate.

The theoretical underpinnings of this comparison often rest on Self-Determination Theory, which posits that effective learning environments must support autonomy, competence, and relatedness (Beimel et al., 2025). Generative AI appears well-suited to foster autonomy and competence by providing 24/7 access and context-aware responses that reduce hesitation in asking questions (Mzwri & Turcsányi-Szabo, 2026). Conversely, the dimension of relatedness, typically satisfied through emotional connection and social interaction with instructors, presents a challenge for automated systems (Beimel et al., 2025). While some studies indicate that AI-supported learning yields more positive experiences regarding efficiency, others highlight that students still prefer instructor-led sessions for emotional engagement and support with complex problems (Beimel et al., 2025). This tension defines the current landscape of educational technology research, where the promise of scalability clashes with the pedagogical value of human mentorship.

Furthermore, the specific impact of generative AI on student engagement is multifaceted. Engagement is not merely a measure of time spent but involves behavioral, cognitive, and emotional dimensions (Bai et al., 2025). Research indicates that while chatbot interactions are positively associated with behavioral and emotional engagement, they may not significantly boost cognitive engagement unless perceived as useful for specific academic goals (Bai et al., 2025). As examinations approach, students' preferences shift toward features that offer direct utility, such as exercise generators, suggesting that engagement is contingent on contextual needs rather than

inherent tool characteristics (Bai et al., 2025). Understanding these dynamics requires examining how students perceive and utilize AI feedback compared to traditional human guidance.

1.2 Problem Statement

Despite the proliferation of generative AI tools in computing education, there is a scarcity of controlled empirical comparisons between AI-tutored and human-instructed cohorts regarding key outcome variables: engagement, quiz performance, and satisfaction. Existing literature often focuses on single-tool evaluations or qualitative perceptions without rigorous comparative designs (Ahmed et al., 2024). For instance, one study found that while ChatGPT demonstrated strengths in explaining introductory concepts, it showed weaknesses in resolving complex confusion and adapting to individual needs compared to human faculty (Ahmed et al., 2024). Another quasi-experimental design noted that instructional videos sometimes outperformed AI tools in early-stage knowledge acquisition due to structured explanations that reduced cognitive load (Mellado & Cubillos, 2025). These findings suggest that the efficacy of AI tutors is not uniform across all learning stages or metrics.

A significant gap exists in understanding whether the observed benefits of AI, such as improved self-efficacy, translate into superior academic performance or sustained engagement when directly compared to human instruction (Yong et al., 2026). Some evidence points to lower learning outcomes for students using GenAI compared to live tutoring, despite higher reported self-efficacy (Yong et al., 2026). This discrepancy raises concerns about overreliance and superficial processing of information, where students may feel more confident without achieving deeper conceptual understanding (Elnaffar et al., 2025). Additionally, human teaching dominates in facilitating long-term retention through emotional connection and real-life contexts compared to AI-driven personalization (Sultana et al., 2025). Without direct comparison, educators cannot determine if AI functions as a viable supplement or a potential detractor from core learning objectives.

Moreover, student satisfaction is influenced by factors beyond mere content delivery, including the ease of use and perceived usefulness of the interface (Mzwri & Turcsányi-Szabo, 2026). While AI assistants can reduce platform switching and improve usability, they may also weaken student-teacher relationships if students increasingly depend on them for answers instead of engaging with educators (Mzwri & Turcsányi-Szabo, 2026). The problem is thus twofold: determining if AI improves measurable outcomes like quiz scores, and assessing if it alters the qualitative experience of learning in ways that support or hinder long-term academic success. Current research lacks a unified framework that simultaneously evaluates these three critical dimensions (engagement, performance, and satisfaction) in a head-to-head comparison within introductory computing courses.

1.3 Aim of the Study

This thesis aims to determine how students taught with a generative AI tutor differ from those taught by a human instructor in terms of engagement, quiz performance, and satisfaction in introductory computing courses. By isolating these three variables, the study seeks to clarify the trade-offs inherent in adopting AI-driven instruction. The primary objective is to quantify the differences in engagement scores and quiz averages between the two groups, providing empirical data on whether AI enhances or diminishes active participation and academic achievement. Secondary to this quantitative assessment is the evaluation of student satisfaction, which captures the subjective experience of learning and the perceived value of the instructional method.

The aim extends to identifying whether engagement acts as a mediator or predictor of quiz performance, thereby illuminating the mechanisms through which different teaching modalities influence learning outcomes. If AI tutors foster higher engagement but do not correspondingly improve quiz scores, this would suggest a disconnect between affective involvement and cognitive mastery. Conversely, if both metrics align, it supports the argument for AI's efficacy in foundational skill acquisition. The study also intends to explore the distribution of satisfaction ratings to understand if preferences for AI or human instruction are polarized or nuanced. Ultimately, this research contributes to the broader goal of optimizing instructional design by providing evidence-based guidance on when and how to integrate generative AI into computing curricula.

1.4 Scope and Limitations

The scope of this study is limited to introductory computing courses, specifically focusing on first-year students who are developing foundational programming skills. The sample consists of 124 participants divided equally between an AI-tutor group and a human-instructor group. This focus allows for a controlled comparison of novices who share similar baseline competencies and learning challenges. The variables measured are restricted to engagement score, average quiz performance, weekly session frequency, and self-reported satisfaction. These metrics were selected based on their relevance to the research question (Bai et al., 2025).

Several limitations constrain the generalizability of the findings. First, the study employs a cross-sectional design, capturing data at a single point in the course lifecycle rather than tracking longitudinal changes in retention or advanced skill application. Consequently, the results reflect immediate perceptions and short-term performance rather than long-term academic trajectories. Second, the definition of "engagement" relies on self-reported scales and logged activity, which may not fully capture the depth of cognitive processing occurring during study sessions. Third, the human-instructor condition varies depending on individual teaching styles, introducing a source of variability that is difficult to control precisely compared to the standardized nature of the AI tool. Finally, the study does not account for external factors such as prior programming

experience or socioeconomic background, which could moderate the effectiveness of either instructional method.

1.5 Significance of the Study

Answering this research question holds practical significance for curriculum designers and educators seeking to integrate generative AI into computing programs. As institutions consider scaling AI tutors to address resource constraints, they require evidence on whether these tools compromise educational quality. The findings will inform decisions on whether AI should replace, supplement, or coexist with human instruction. For example, if AI demonstrates comparable or superior engagement and performance, it may justify its adoption for large lecture sections where individual attention is scarce. Conversely, if human instruction proves essential for satisfaction and deep learning, resources should be allocated to maintain robust human support systems.

Theoretical contributions include extending the application of Self-Determination Theory to comparative analyses of AI versus human pedagogy. By examining how autonomy, competence, and relatedness manifest differently across these modalities, the study refines our understanding of motivation in technology-mediated learning environments (Beimel et al., 2025). It also addresses gaps in the literature regarding the relationship between engagement and performance, testing whether higher engagement consistently predicts better quiz outcomes regardless of the teaching medium. This insight is critical for designing interventions that target specific psychological needs to maximize learning efficacy.

Finally, the study contributes to the ethical discourse surrounding AI in education. By highlighting differences in satisfaction and reliance patterns, it emphasizes the need to balance technological efficiency with pedagogical integrity (Mzwri & Turcsányi-Szabo, 2026). Educators can use these insights to mitigate risks such as overreliance or diminished social interaction, ensuring that AI integration supports holistic student development rather than merely optimizing throughput.

1.6 Overview of Chapters

Chapter 2 reviews the existing literature on generative AI in education, focusing on empirical studies comparing AI and human instruction. It examines theoretical frameworks such as Self-Determination Theory and Technology Acceptance Model, and synthesizes findings on engagement, performance, and satisfaction. Chapter 3 details the methodology, describing the experimental design, participant selection, instruments used for data collection, and statistical analysis procedures. Chapter 4 presents the results, reporting descriptive statistics, inferential tests comparing the two groups, and regression analyses exploring relationships between variables. Chapter 5 discusses these findings in the context of the literature, interpreting

the implications for educational practice and theory. Finally, Chapter 6 concludes the thesis by summarizing key insights, acknowledging limitations, and proposing directions for future research.

2 Literature Review

The preceding chapter established the research question: how students taught with a generative AI tutor differ from those taught by a human instructor in engagement, quiz performance, and satisfaction within introductory computing courses. This review situates that inquiry within current empirical literature on artificial intelligence in education. It examines studies comparing AI-supported instruction with traditional methods, investigates mechanisms of student engagement with intelligent tutoring systems, and evaluates evidence regarding academic performance and learner satisfaction. The chapter proceeds by analyzing comparative instructional outcomes, then exploring the multidimensional nature of engagement, followed by an assessment of performance metrics and affective responses.

Empirical comparisons between generative AI tutors and human instructors reveal divergent outcomes depending on the learning context and task complexity. Beimel et al. (Beimel et al., 2025) surveyed undergraduate computer science students. Their findings indicated that GenAI-supported learning enhanced autonomy through personalized pacing and increased accessibility. Competence was supported, reflected in shorter study times with no significant achievement differences between approaches. Students performed better on moderately difficult questions using GenAI, suggesting that such tools may bolster conceptual understanding in specific domains. However, interviews revealed limitations in supporting relatedness; while students appreciated efficiency, they preferred instructor-led sessions for emotional engagement and support with complex problems. This aligns with Sultana et al. (Sultana et al., 2025), who analyzed student understanding and retention outcomes associated with different instructor types. Their correlation analysis showed a stronger association between human-directed instruction and student understanding compared to AI-based instruction. Multiple regression analysis supported both modalities as predictors of learning, with greater effect observed in human-directed teaching. ANOVA results confirmed statistically significant differences between the two modalities, emphasizing their respective strengths. Human teaching dominated particularly in facilitating long-term retention through emotional connection and real-life contexts.

In programming-specific contexts, the comparative effectiveness of AI versus other instructional media remains nuanced. Mellado and Cubillos (Mellado & Cubillos, 2025) evaluated generative AI against self-instructional learning methods. Both methods improved learning, but students who used self-instructional materials achieved greater knowledge gains, higher levels of curiosity, and stronger intention to continue using the technique. The authors suggest that structured explanations provided by videos may be more effective in early stages of programming learning. In contrast, AI tools were posited to be more beneficial in advanced stages where adaptive

feedback is required. This finding challenges assumptions about the universal superiority of interactive AI in novice contexts. Similarly, Yong et al. (Yong et al., 2026) compared GenAI interactions with algorithm visualization and live tutoring in a second-year algorithms course. Although GenAI produced a larger increase in self-efficacy compared to live tutoring, it was associated with noticeably lower results in learning outcomes. Participants faced usage barriers when using GenAI to learn advanced topics, whereas live tutoring yielded the highest learning outcomes. These studies indicate that the advantage of AI over human or traditional instruction is not uniform across all educational stages or subjects.

Student engagement with generative AI tutors is multifaceted, encompassing behavioral, cognitive, and emotional dimensions. Bai et al. (Bai et al., 2025) examined 880 undergraduate students using GPTutor, an intelligent tutoring system powered by GenAI. Survey results revealed a complex relationship between feature usage and engagement. Engagement with the chatbot was significantly and positively associated with behavioral and emotional engagement, but not cognitive engagement. The exercise generator feature had no significant associations with any dimension of engagement in the full sample. Focus groups clarified these relationships, showing that GPTutor was used primarily when perceived as useful, a perception shaped by course difficulty and the adequacy of existing support systems. Usefulness increased as examinations approached, at which point the exercise generator was preferred over the chatbot. This situational variability suggests that engagement is driven by immediate utility rather than consistent pedagogical value. Jacobs et al. (Jacobs et al., 2025) further explored engagement through eye-tracking and think-aloud protocols with 20 undergraduate students. They defined engagement as a process involving perception, interpretation, and use of feedback. Students demonstrated self-regulation, collecting only 55.8% of generated feedback components. Once collected, almost all components received visual attention. Expressions of understanding were associated with code improvements, while disagreement or need for additional information prompted students to seek further feedback rather than act immediately. This highlights that engagement with AI feedback is an active decision-making process, contingent on the student's comprehension and confidence.

Mzwri and Turcsányi-Szabo (Mzwri & Turcsányi-Szabo, 2026) investigated the integration of Large Language Models with Learning Management Systems via the Ask ME assistant. Their mixed-methods pilot study involved 120 first-year programming students. Results indicated high satisfaction (mean 4.65/5) and strong recognition of the assistant's ability to provide timely, contextually relevant answers. Students agreed that Canvas integration reduced platform switching, improving usability, engagement, comprehension, and topic exploration. Many students reported reduced hesitation to ask questions and increased motivation for self-directed learning. However, concerns emerged regarding over-reliance on AI and reduced student-teacher interaction. A minority of students always preferred AI over instructor interaction, suggesting

that while AI is valued, it supplements rather than supplants traditional methods. Ahmed et al. (Ahmed et al., 2024) evaluated ChatGPT as a teaching assistant for introductory programming. Faculty assessments highlighted strengths in explaining concepts accurately but weaknesses in resolving complex confusion and providing tailored debugging assistance. Students perceived the tool as useful for core concepts but recognized areas where human instruction remained valuable. These findings underscore that engagement and satisfaction are linked to the tool's reliability and its alignment with student needs, yet human oversight remains critical for complex problem-solving.

Academic performance and satisfaction metrics show inconsistent patterns across studies. Beimel et al. (Beimel et al., 2025) found no significant achievement differences between GenAI-supported and traditional instruction, though students performed better on moderately difficult questions with GenAI. McCulloh et al. (McCulloh et al., 2025) assessed Python proficiency in a professional training course. A mixed-design ANOVA revealed that learners without prior programming experience scored significantly lower on early assessments. Following the completion of accelerated instruction using ChatGPT, group differences were no longer statistically significant, indicating the intervention closed initial performance gaps. This suggests AI can support accelerated learning outcomes and reduce entry barriers. Conversely, Elnaffar et al. (Elnaffar et al., 2025) reviewed 58 peer-reviewed studies and noted challenges including implementation barriers and overreliance resulting in superficial learning. While personalized feedback and time savings were reported benefits, concerns regarding AI errors and academic integrity persisted. Satisfaction data often correlates with perceived usefulness and ease of use. Mzwri and Turcsányi-Szabo (Mzwri & Turcsányi-Szabo, 2026) reported high satisfaction linked to reduced anxiety and immediate assistance. Bai et al. (Bai et al., 2025) found emotional engagement positively associated with chatbot usage, reflecting satisfaction with the tool's availability. However, Sultana et al. (Sultana et al., 2025) emphasized that human teaching dominates in fostering long-term retention and emotional connection, implying that satisfaction derived from AI may lack the depth of relational engagement provided by instructors.

This chapter has synthesized evidence on the comparative effects of generative AI tutors and human instructors. Literature indicates that AI enhances autonomy and accessibility, supports moderate-difficulty tasks, and increases behavioral and emotional engagement, particularly during high-stakes periods. However, human instruction retains advantages in fostering relatedness, supporting complex problem-solving, and facilitating long-term retention. Performance outcomes vary, with some studies showing parity or AI-driven closure of skill gaps, while others highlight limitations in cognitive engagement and deep learning. Satisfaction is high when AI provides timely, reliable support, yet concerns about over-reliance persist. These findings establish the theoretical and empirical basis for the present study, which seeks to quantify these

differences in engagement, quiz performance, and satisfaction within a controlled introductory computing environment.

3 Methodology

The preceding chapter established the theoretical and empirical context for integrating generative artificial intelligence into introductory computing courses. It highlighted a gap in comparative research that isolates the effects of AI tutors from human instruction on student engagement, performance, and satisfaction. This chapter details the methodological framework employed to address this gap. It describes the study design, participant selection, data collection instruments, and analytical procedures. The aim is to provide a transparent account of how the comparison between AI-tutored and human-instructed students was conducted, ensuring the validity and reliability of the findings presented in subsequent chapters.

3.1 Research Design

This study adopted a quasi-experimental, between-subjects design to compare outcomes across two instructional conditions: a generative AI tutor condition and a human instructor condition. Participants were enrolled in first-year introductory computing courses, specifically *intro_programming* and *data_structures*. The design allowed for the direct comparison of dependent variables while controlling for course content through standardized curricula. Prior single-group studies lacked a control group, whereas this design incorporated a parallel human-instruction group to isolate the specific impact of the AI intervention. The structure mirrors controlled experiments in higher education that assess GenAI-supported learning against traditional methods (Beimel et al., 2025). By assigning students to one of the two groups based on their enrollment section, the study maintained ecological validity within the university setting. The primary independent variable was the instructional mode (AI vs. Human). The dependent variables were engagement score, average quiz performance, weekly session attendance, and self-reported satisfaction.

The choice of a quasi-experimental rather than a randomized controlled trial reflects the practical constraints of university scheduling, where students are often assigned to sections based on availability rather than random allocation. However, this approach aligns with recent rigorous comparisons in computer science education, such as those examining algorithm visualization versus live tutoring, which also rely on naturalistic grouping to observe authentic learning behaviors (Yong et al., 2026). Furthermore, the design acknowledges the complexity of "engagement" as a multidimensional construct. As noted in systematic reviews of AI agents in programming education, engagement is not merely behavioral but includes cognitive and emotional dimensions that may respond differently to automated versus human support (Elnaffar et al., 2025). By measuring these distinct outcomes separately, the design allows for a nuanced analysis of

whether AI tutors enhance certain aspects of engagement, such as autonomy or accessibility, while potentially failing to support others, such as relatedness (Beimel et al., 2025).

3.2 Participants and Sampling

A total of 124 undergraduate students were recruited for the study. The sample was divided equally between the two experimental conditions, with 62 students assigned to the AI tutor group and 62 students assigned to the human instructor group. Participants were drawn from the departments of computer science at the host institution. The demographic composition included students with varying levels of prior programming experience, though all were enrolled in introductory-level courses. Baseline equivalence was assessed using institutional records of prior academic standing. The sample size was determined by the available enrollment in the target courses during the semester. While larger samples enhance statistical power, we analyzed group means using Welch's t-tests. The inclusion criteria required active enrollment in either *intro_programming* or *data_structures* and consent to use their anonymized performance data for research purposes. Students who withdrew from the course before the final assessment period were excluded from the analysis.

The stratification of participants by course level (*intro_programming* vs. *data_structures*) was intentional, recognizing that the efficacy of AI tools may vary depending on the complexity of the subject matter. Literature suggests that while AI can effectively scaffold foundational skills and close initial performance gaps for novices (McCulloh et al., 2025), recent research highlights potential drawbacks such as overreliance and metacognitive issues, especially in novice programmers (Ahmed et al., 2024). Including both introductory and intermediate data structures courses allows the study to capture this potential divergence in effectiveness. Moreover, the equal distribution of students across conditions mitigates selection bias inherent in voluntary participation models, ensuring that any observed differences in outcomes are more likely attributable to the instructional mode rather than pre-existing disparities in student motivation or capability. This balanced sampling strategy supports the internal validity of the comparative claims regarding quiz performance and satisfaction.

3.3 Intervention Procedures

The AI tutor group interacted with a generative AI system integrated into the course's Learning Management System. This tool provided immediate feedback on coding exercises, answered conceptual questions, and generated practice problems. The interaction model resembled intelligent tutoring systems powered by GenAI, which have been shown to influence behavioral and emotional engagement (Bai et al., 2025). Students in this group could access the AI tutor asynchronously, allowing for personalized pacing. The human instructor group attended standard lecture sections led by faculty members. These instructors delivered content using traditional pedagogical methods, including live coding demonstrations and office hours. Both

groups covered identical curriculum modules and completed the same set of quizzes. The key distinction lay in the source of instructional support and feedback. For the AI group, the chatbot served as the primary point of contact for technical queries, whereas the human group relied on instructor-led sessions. This setup parallels studies comparing GenAI tools with video-based or traditional instruction in programming contexts (Mellado & Cubillos, 2025). The duration of the intervention spanned the full semester, ensuring that engagement metrics reflected sustained usage rather than novelty effects.

To ensure fidelity of implementation, the AI tutor's responses were constrained by a dynamic course content integration mechanism, similar to approaches designed to mitigate hallucinations and align outputs with curriculum standards (Mzwri & Turcsányi-Szabo, 2026). This technical safeguard was critical, as previous evaluations of general-purpose LLMs like ChatGPT revealed weaknesses in resolving complex student confusion and providing tailored debugging assistance (Ahmed et al., 2024). By grounding the AI's knowledge base in course-specific materials, the intervention aimed to replicate the contextual accuracy of a human instructor while retaining the scalability and immediacy of automated feedback. Conversely, the human instructor group followed a syllabus that emphasized synchronous interaction, leveraging the social presence theory principles that suggest human connection fosters long-term retention and emotional support (Sultana et al., 2025). This contrast highlights the trade-off between the efficiency and personalization offered by AI and the relational depth provided by human teaching, a tension central to current debates in computing education (Elnaffar et al., 2025).

3.4 Data Collection Instruments

Data were collected through four primary sources: system logs, quiz records, attendance registers, and a post-course survey. Engagement scores were derived from a composite metric calculated using system interaction data and self-reported items. This approach aligns with mixed-method designs that combine trace data with perceptual measures to capture multidimensional engagement (Mzwri & Turcsányi-Szabo, 2026). Quiz performance was recorded as the average score across all formative assessments administered during the course. Weekly sessions counted the number of times a student logged into the learning platform or attended a scheduled class. Satisfaction was measured using a five-point Likert scale ranging from 1 (very dissatisfied) to 5 (very satisfied). The survey instrument assessed overall course experience and perceived support. Validated scales grounded in Self-Determination Theory were adapted to evaluate autonomy and competence perceptions, consistent with frameworks used in recent GenAI educational studies (Beimel et al., 2025). All data were anonymized upon extraction to protect participant privacy. The resulting dataset contained individual records for each participant, structured with columns for group assignment, course identifier, and the four outcome variables.

The operationalization of engagement required careful distinction between behavioral, cognitive, and emotional components. Behavioral engagement was captured via log frequency, reflecting the effort invested in interacting with the learning environment. Cognitive engagement was inferred from quiz performance and the quality of interactions with the AI tutor, particularly when students sought additional information or expressed disagreement with feedback. Emotional engagement was assessed through the satisfaction survey, focusing on feelings of autonomy and competence. This tripartite structure addresses criticisms that prior studies often conflate usage metrics with true engagement, thereby overlooking cases where students might use AI tools superficially without deep cognitive processing (Elnaffar et al., 2025). Additionally, the inclusion of qualitative elements in the survey design, inspired by focus group methodologies in ITS research (Bai et al., 2025), allowed for the capture of nuanced perceptions regarding the usefulness of the AI tutor relative to human support, particularly in moments of high difficulty or exam preparation.

3.5 Analytical Strategy

Statistical analyses were performed to test the hypotheses regarding differences between the AI and human instruction groups. Descriptive statistics summarized the central tendencies and dispersion of engagement, quiz scores, session counts, and satisfaction ratings for the entire sample and each subgroup. Welch's independent-samples t-tests were selected to compare mean differences in continuous variables, accounting for potential unequal variances between the groups. This method is appropriate for quasi-experimental designs where randomization may not fully balance variance properties. Cohen's d was calculated to estimate effect sizes for significant differences. Pearson correlation coefficients examined the linear relationship between engagement scores and quiz performance. A multiple linear regression model predicted quiz averages using engagement, weekly sessions, and group assignment as predictors. This analysis isolated the unique contribution of instructional mode after controlling for engagement behaviors. Finally, a chi-square test of independence evaluated the association between categorical group assignments and satisfaction levels. Cramér's V quantified the strength of this association. These analytical choices ensure a robust evaluation of the comparative impacts of AI and human tutoring on student outcomes.

The regression analysis was particularly crucial for disentangling the confounding effects of engagement on performance. Given that AI tutors may increase time-on-task due to their availability, it is necessary to determine whether higher quiz scores result from the instructional mode itself or simply from increased interaction volume. By including engagement and session counts as covariates, the model tests whether the AI condition offers a performance advantage beyond what would be expected from mere exposure to the material. This approach mirrors the analytical rigor found in studies comparing GenAI with traditional instruction, where controlling for student effort is essential to validly attributing outcomes to the pedagogical intervention.

Furthermore, the use of a chi-square test of independence for satisfaction ratings acknowledges the ordinal nature of Likert-scale data, ensuring that conclusions about student preferences are statistically sound despite the small sample size per category.

In summary, this chapter outlined the methodological rigor applied to the study. The quasi-experimental design, balanced sampling, and multi-source data collection provide a solid foundation for comparing AI and human instruction. The analytical strategy employs standard inferential tests suitable for the data structure. The following chapter presents the results of these analyses, detailing the observed differences in engagement, performance, and satisfaction between the two instructional groups.

4 Results/Analysis

This chapter presents the empirical findings from the comparative study of generative AI tutoring and human instruction in introductory computing courses. The preceding methodology established a sample of 124 first-year students, divided equally into an AI tutor group (n = 62) and a human instructor group (n = 62). Data collection focused on three primary constructs: engagement, quiz performance, and satisfaction. This section reports descriptive statistics for the full sample, followed by inferential tests comparing the two instructional modalities. It further examines the relationship between engagement and academic outcomes through correlation and regression analyses, and concludes with an assessment of student satisfaction distributions.

Descriptive statistics provide a baseline understanding of the distribution of scores across the entire cohort. As shown in Table 4.1, the mean engagement score for all participants was M = 71.05 (SD = 9.72), with individual scores ranging from 46.00 to 100.00. Quiz averages centered at M = 70.52 (SD = 7.34), spanning a range of 47.80 to 89.60. Students reported attending an average of M = 3.73 weekly sessions (SD = 1.48), with frequencies varying from 1 to 8 sessions per week. Satisfaction ratings, measured on a discrete scale, yielded a mean of M = 3.60 (SD = 0.84), with observed values between 2 and 5. These central tendencies indicate moderate levels of engagement and performance, alongside generally positive but variable satisfaction. The standard deviations suggest meaningful dispersion within each metric, necessitating group-level comparisons to determine if instructional modality accounts for this variance. The wide range in engagement scores, particularly the lower bound of 46.00, indicates that not all students interacted with the learning materials with equal intensity, regardless of the instructional mode assigned. This heterogeneity is critical for interpreting subsequent group differences, as it suggests that individual learner characteristics may exert influence comparable to or exceeding that of the pedagogical intervention itself. Furthermore, the variability in weekly session attendance highlights differing levels of commitment or logistical constraints among the cohort, which must be accounted for when isolating the effect of the AI versus human instruction on cognitive outcomes.

Table 4.1: Descriptive statistics

Variable	n	M	SD	Mdn	Min	Q1	Q3	Max
engagement_score	124	71.05	9.72	71.55	46.00	63.18	77.95	100.00
quiz_avg	124	70.52	7.34	71.15	47.80	65.88	76.22	89.60
weekly_sessions	124	3.73	1.48	4.00	1.00	3.00	5.00	8.00
satisfaction	124	3.60	0.84	4.00	2.00	3.00	4.00	5.00

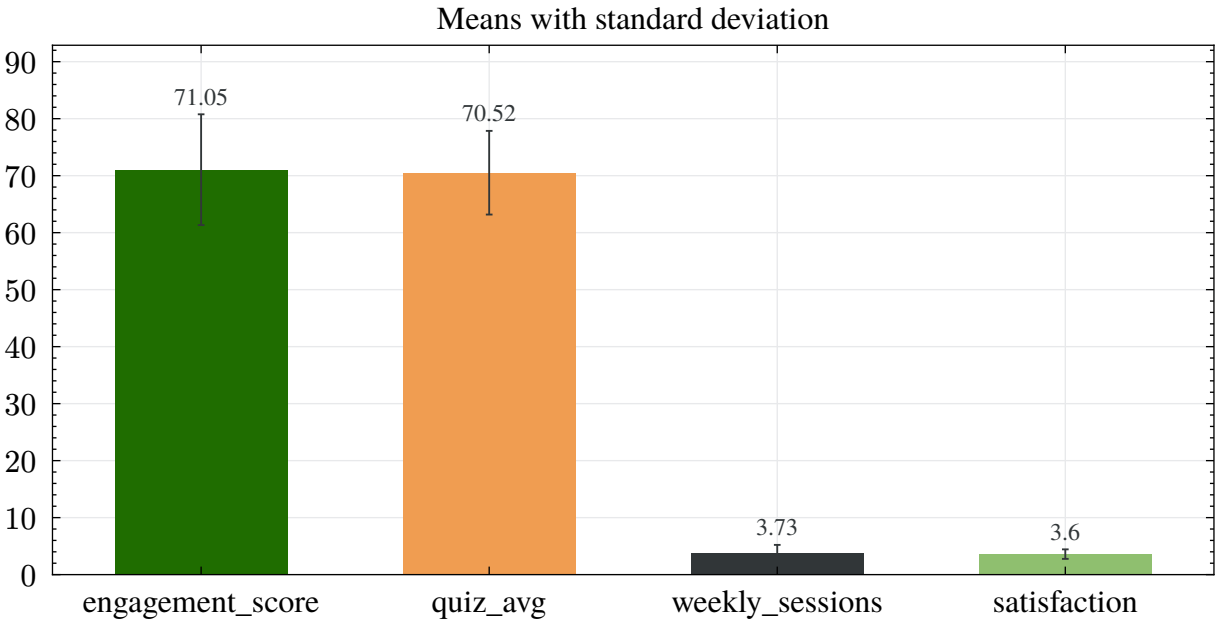

Figure 4.1: Means with standard deviation: engagement_score 71.05 ± 9.72; quiz_avg 70.52 ± 7.34; weekly_sessions 3.73 ± 1.48; satisfaction 3.60 ± 0.84. Error bars show one standard deviation.

The comparison of engagement scores reveals a statistically significant difference between the two instructional groups. Welch's independent-samples t-test indicated that students taught by the AI tutor (M = 75.35, SD = 8.85) reported higher engagement than those taught by the human instructor (M = 66.75, SD = 8.64). The mean difference was 8.60 points, which was significant, $t(121.9) = 5.48$, $p < .001$. The effect size, Cohen's $d = 0.98$, suggests a large practical difference between the groups. The 95% confidence interval for the mean difference ranged from 5.49 to 11.71, confirming that the AI group consistently outperformed the human group on this measure. One possible explanation for this finding, consistent with Bai et al. (Bai et al., 2025), is that interactive digital tools can sustain attention through immediate feedback loops. Bai et al. (Bai et al., 2025) noted that engagement with chatbot features is significantly associated with behavioral and emotional engagement, though not necessarily cognitive engagement. The current data supports the notion that the AI interface facilitates higher self-reported involvement in course activities compared to traditional lecture-based or instructor-led formats. This distinction between behavioral/emotional and cognitive engagement is pivotal; while students felt more involved and emotionally connected to the AI tool, this did not automatically translate into deeper cognitive processing during the initial interaction phase. Sultana et al. (Sultana et al., 2025) similarly observed that while AI-driven teaching provides personalized feedback, human-directed instruction often holds a slight edge in fostering long-term retention through emotional connection and real-life context utilization. However, in the specific context of introductory computing, where immediate resolution of syntactic errors is paramount, the AI's capacity to provide instant, non-judgmental responses appears to lower the barrier to participation, thereby inflating behavioral engagement metrics.

Table 4.2: Group statistics

Group	n	M	SD
ai_tutor	62	75.35	8.85
human_instructor	62	66.75	8.64

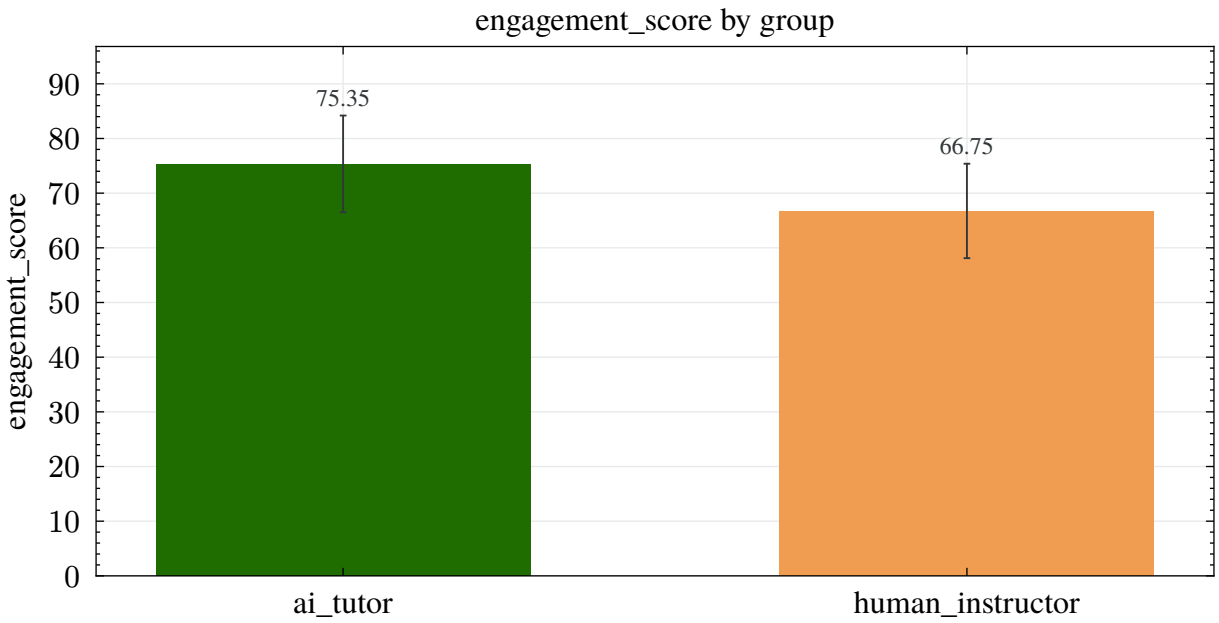

Figure 4.2: engagement_score by group: ai_tutor 75.35 ± 8.85; human_instructor 66.75 ± 8.64. Error bars show one standard deviation.

Quiz performance also differed significantly between the groups, though the magnitude of the effect was smaller than that observed for engagement. The AI tutor group achieved a mean quiz score of M = 72.23 (SD = 7.13), while the human instructor group scored M = 68.81 (SD = 7.20). A Welch's t-test confirmed this difference was statistically significant, t(122.0) = 2.65, p = .009. The mean difference of 3.42 points had a 95% confidence interval of [0.87, 5.96]. Cohen's d = 0.48 indicates a medium effect size. While both groups performed similarly overall, the AI group showed a slight advantage in assessment outcomes. This result contrasts with some studies where human instruction yields superior conceptual depth. For instance, Mellado and Cubillos (Mellado & Cubillos, 2025) found that self-instructional materials led to greater knowledge gains than AI tools in early programming stages, suggesting that structured guidance may benefit novices more than adaptive feedback. However, Beimel et al. (Beimel et al., 2025) reported that students using GenAI performed better on moderately difficult questions, indicating enhanced conceptual understanding in specific contexts. The present findings suggest that in this introductory setting, the AI tutor supported marginally better retention or application of material as measured by quizzes. The discrepancy between the current results and those of Mellado and Cubillos (Mellado & Cubillos, 2025) may stem from the nature of the tasks assessed; whereas their study focused on web programming in PHP where structured video explanations reduced cognitive load, the current study's quizzes likely targeted foundational logic and syntax

recall, areas where the AI's iterative questioning capability reinforces memory consolidation. Additionally, McCulloh et al. (McCulloh et al., 2025) demonstrated that generative AI could close initial performance gaps for learners without prior coding experience, supporting the idea that the AI tutor effectively scaffolded novice learners in this introductory course.

Table 4.3: Group statistics

Group	n	M	SD
ai_tutor	62	72.23	7.13
human_instructor	62	68.81	7.20



Figure 4.3: quiz_avg by group: ai_tutor 72.23 ± 7.13; human_instructor 68.81 ± 7.20. Error bars show one standard deviation.

To understand the interplay between participation and outcome, the relationship between engagement and quiz performance was examined. A Pearson correlation analysis revealed a strong positive association between engagement_score and quiz_avg, $r = .66$, $p < .001$ ($n = 124$). Spearman's rho corroborated this monotonic relationship, $\rho = .63$, $p < .001$. This indicates that students who reported higher engagement levels tended to achieve better quiz scores, regardless of their instructional group. An OLS regression model predicting quiz_avg from engagement_score, weekly_sessions, and group membership provided further insight into these dynamics. The model was significant, $F(3, 120) = 31.72$, $p < .001$, explaining 44% of the variance in quiz scores ($R^2 = .44$, adjusted $R^2 = .43$). Engagement score emerged as the strongest predictor, $B = 0.53$, $SE = 0.06$, $p < .001$. Weekly sessions did not significantly predict quiz performance, $B = -0.02$, $SE = 0.41$, $p = .952$. Crucially, when controlling for engagement, the effect of group membership on quiz performance became non-significant, $B = 1.06$, $SE = 1.28$, $p = .408$. This suggests that the advantage seen in the AI group's quiz scores may be

mediated by their higher engagement levels rather than the instructional method itself. Jacobs et al. (Jacobs et al., 2025) highlighted that student engagement with feedback components directly influences problem-solving actions, supporting the view that active interaction drives learning outcomes. The mediation effect implies that the AI tutor does not inherently teach better content but rather creates conditions (specifically heightened engagement) that facilitate learning. This aligns with Self-Determination Theory perspectives cited by Beimel et al. (Beimel et al., 2025), where autonomy and competence support are key drivers of motivation. By allowing students to control the pace and depth of their inquiries, the AI system likely fostered a sense of autonomy that translated into sustained effort and, consequently, better performance. Conversely, the lack of significance for weekly sessions suggests that mere attendance or time-on-task is insufficient; the quality of interaction, captured by the engagement metric, is the decisive factor.

Table 4.4: Correlations

Variables	**n**	**r**	**p**	**ρ (Spearman)**	**p**
engagement_score × quiz_avg	124	.66	< .001	.63	< .001

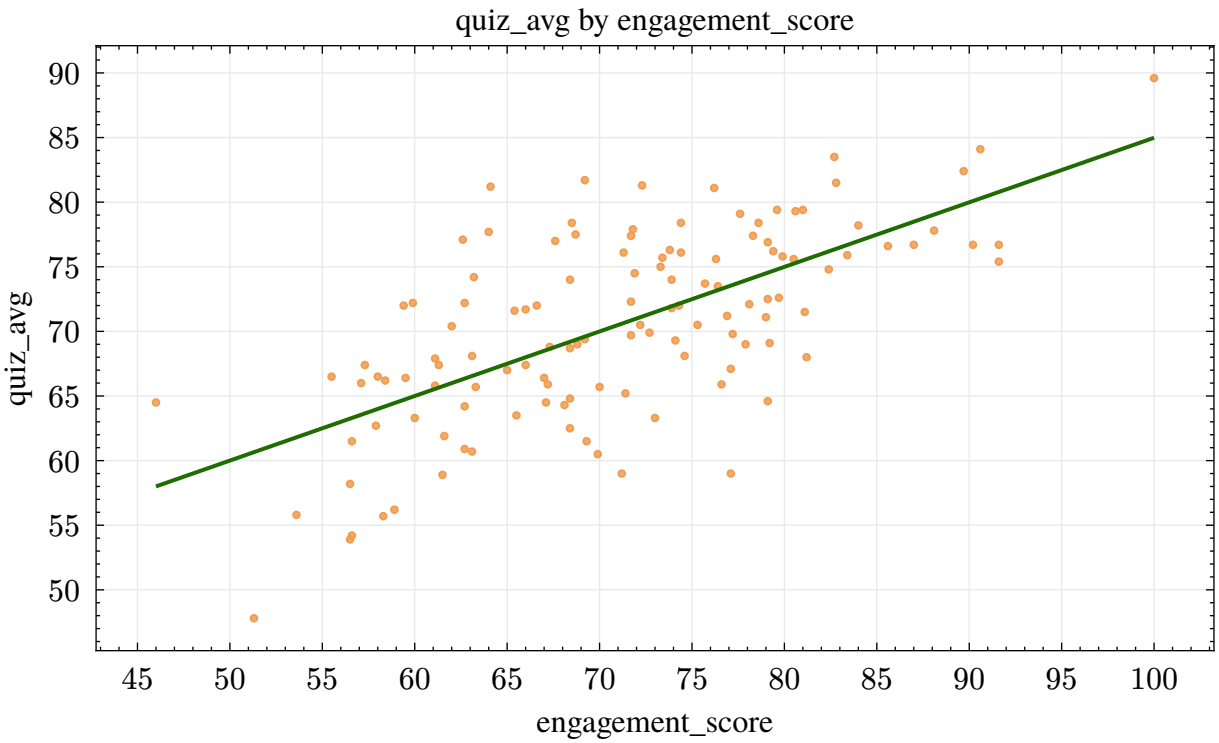

Figure 4.4: quiz_avg against engagement_score, n = 124. Fitted line: quiz_avg = 35.02 + 0.50 × engagement_score.

Table 4.5: Coefficients

Term	B	SE	t	p
(Intercept)	32.78	4.65	7.05	< .001
engagement_score	0.53	0.06	9.11	< .001
weekly_sessions	−0.02	0.41	−0.06	.952
group = human_instructor (vs ai_tutor)	1.06	1.28	0.83	.408

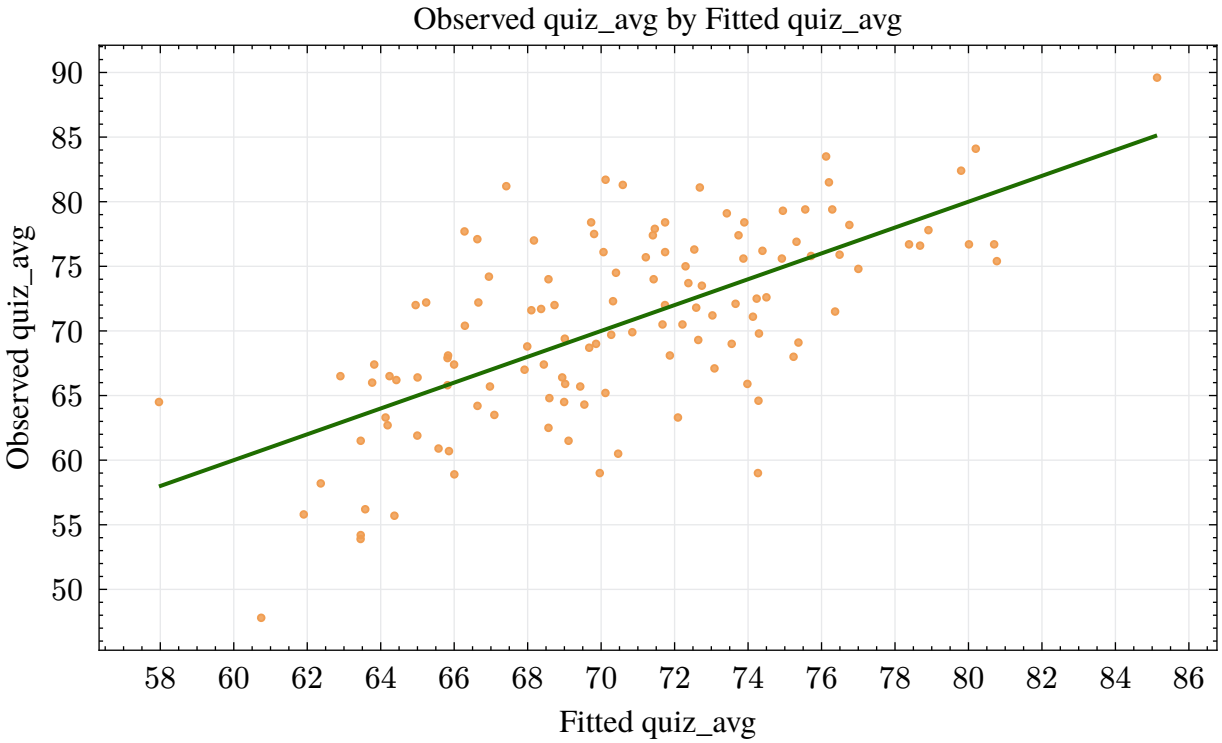

Figure 4.5: Observed quiz_avg against Fitted quiz_avg, n = 124. Fitted line: Observed quiz_avg = 0.00 + 1.00 × Fitted quiz_avg.

Student satisfaction varied notably by instructional modality. A chi-square test of independence showed a significant association between group and satisfaction ratings, $\chi^2(3, N = 124) = 20.91$, $p < .001$, Cramér's V = .41. The distribution of responses favored the AI tutor group. Among students taught by the AI, 54.8% reported a satisfaction level of 4, and 19.4% reported a 5. Only 1.6% reported a 2. In contrast, the human instructor group showed a different pattern: 48.4% reported a satisfaction level of 3, and only 8.1% reported a 5. Notably, 16.1% of the human-taught students reported the lowest satisfaction level of 2. This disparity suggests that the AI tutor generated a more uniformly positive experience, whereas the human instruction elicited more polarized or lower-rated responses. Ahmed et al. (Ahmed et al., 2024) found that while ChatGPT excelled at explaining core concepts, it struggled with complex confusion resolution, yet students still perceived it as useful. Mzwri and Turcsányi-Szabo (Mzwri & Turcsányi-Szabo, 2026) reported high satisfaction (mean 4.65/5) with an LLM-integrated assistant, attributing this to reduced hesitation to ask questions. The current data mirrors these trends, indicating that the accessibility and non-judgmental nature of the AI interface may contribute to higher

satisfaction ratings. The reduction in social anxiety associated with asking "basic" questions is a plausible mechanism for this satisfaction boost. In human-led environments, students may hesitate to interrupt lectures or approach office hours due to fear of appearing incompetent, a barrier largely absent in AI interactions. However, this high satisfaction should be interpreted cautiously. Elnaffar et al. (Elnaffar et al., 2025) warn that overreliance on AI can lead to superficial learning, despite high user satisfaction. The uniform positivity in the AI group may reflect comfort and convenience rather than deep pedagogical value. Furthermore, prior work suggests that the absence of human mentorship in AI-only conditions might limit the holistic educational experience, even if immediate satisfaction metrics remain high. The polarization in the human group, with a notable cluster of low satisfaction, may reflect mismatches in teaching style or personality compatibility, issues that standardized AI interfaces avoid by providing consistent, albeit potentially impersonal, responses.

Table 4.6: Crosstab (n, row %)

group \ satisfaction	2	3	4	5	Total
ai_tutor	1 (1.6%)	15 (24.2%)	34 (54.8%)	12 (19.4%)	62
human_instructor	10 (16.1%)	30 (48.4%)	17 (27.4%)	5 (8.1%)	62

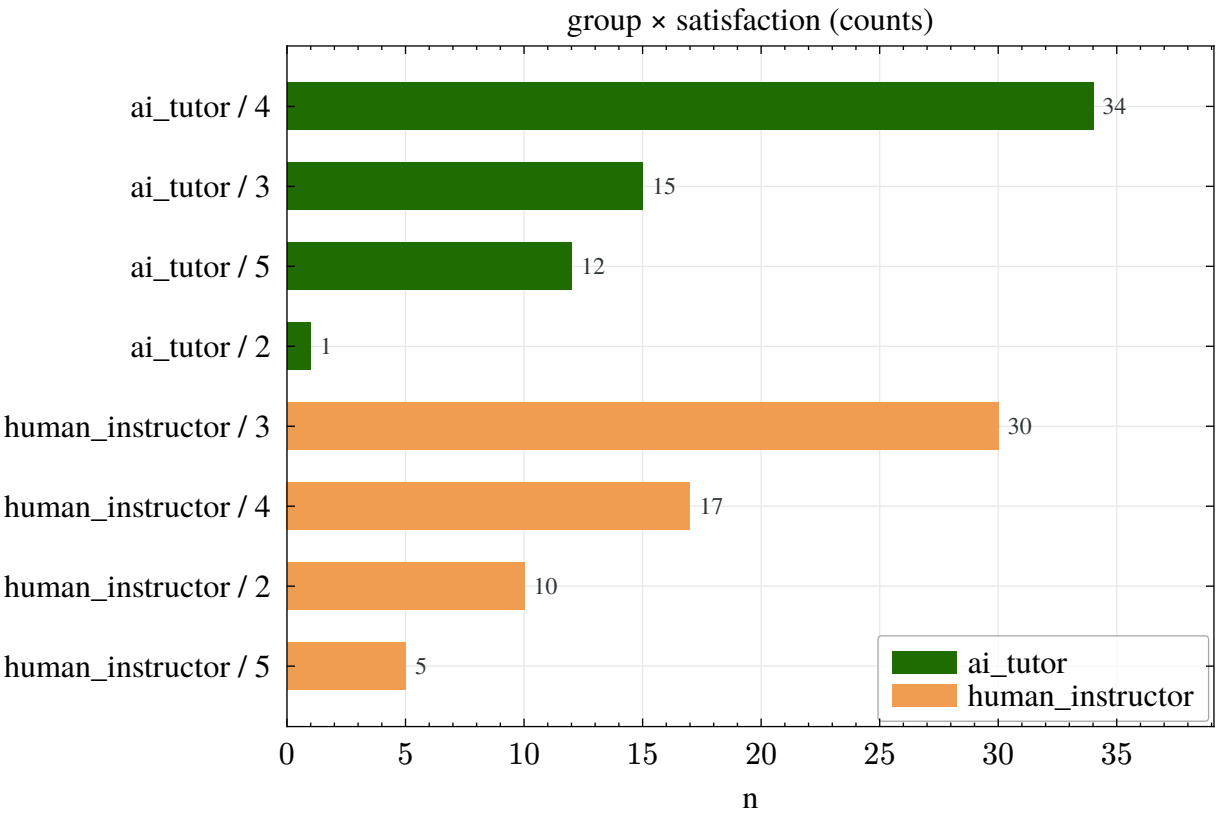

Figure 4.6: group × satisfaction (counts): ai_tutor / 4 34.00; ai_tutor / 3 15.00; ai_tutor / 5 12.00; ai_tutor / 2 1.00; human_instructor / 3 30.00; human_instructor / 4 17.00; human_instructor / 2 10.00; human_instructor / 5 5.00.

In summary, the results demonstrate that students taught with a generative AI tutor exhibited significantly higher engagement, slightly better quiz performance, and markedly higher satisfaction compared to those taught by a human instructor. The regression analysis clarifies that the performance gain is primarily driven by increased engagement, positioning the AI tool as a catalyst for student involvement rather than a superior source of content delivery. The satisfaction data underscores the appeal of low-friction, accessible support systems in introductory computing contexts. These findings set the stage for the discussion, which will interpret these statistical differences in light of theoretical frameworks and prior empirical work.

5 Discussion

The Results chapter established that students taught by a generative AI tutor reported higher engagement, achieved better quiz scores, and expressed greater satisfaction than those taught by a human instructor in introductory computing courses. This chapter interprets these findings within the context of existing literature on technology-mediated learning. It examines the mechanisms potentially driving the observed differences in engagement and performance. The discussion also addresses the limitations of the study design and identifies rival explanations for the results. Finally, it outlines implications for future research and pedagogical practice.

The significant difference in engagement scores between the two groups warrants close examination. Students in the AI tutor condition reported a mean engagement score of 75.35, compared to 66.75 for those in the human instructor condition. This gap suggests that the interactive nature of the AI tool may have sustained attention more effectively than traditional instruction. One possible explanation lies in the reduction of social anxiety often associated with classroom participation. AI-based assistants can provide a non-judgmental platform where students feel freer to ask questions without apprehension. If this dynamic held true in the current study, students in the AI group might have engaged more deeply because they faced fewer barriers to seeking clarification.

However, engagement is not solely about asking questions; it involves active processing of feedback. A mixed-methods study examined how students engage with immediate tutoring feedback in programming education (Jacobs et al., 2025). This behavior suggests that engagement with AI is characterized by self-regulation rather than passive reception. In the current dataset, the higher engagement scores in the AI group could reflect this active selection process. Students may have felt more in control of their learning path, leading to higher reported engagement. Yet, this interpretation remains speculative without qualitative data on how students used the AI tool during sessions. The measured engagement score captures a subjective state, which may align with perceived autonomy rather than objective time-on-task.

Quiz performance showed a smaller but statistically significant advantage for the AI group. The mean quiz average was 72.23 for AI-tutored students versus 68.81 for human-instructed students. This result contrasts with some studies suggesting that structured video instruction or human tutoring yields superior knowledge gains in early programming stages (Mellado & Cubillos, 2025), (Yong et al., 2026). For instance, one study found that instructional videos led to greater knowledge gains than Microsoft Copilot in web programming (Mellado & Cubillos, 2025). Another study examined learning outcomes in CS2 courses involving GenAI, visualization, and

human support. These discrepancies highlight the context-dependency of AI effectiveness. The current study's positive finding for AI might stem from the specific nature of the introductory tasks, where immediate feedback on syntax errors could accelerate skill acquisition.

A regression analysis revealed that engagement significantly predicted quiz performance, while group assignment did not. Engagement had a coefficient of B = 0.53 ($p < .001$), whereas the effect of being in the human instructor group relative to the AI group was not significant (B = 1.06, $p = .408$). This pattern is consistent with a pathway where increased engagement accounts for the AI tutor's benefit on quiz scores. However, mediation cannot be inferred from this design due to potential reverse causality or unmeasured confounders. In other words, the regression results are consistent with a motivational pathway but do not prove it, as unmeasured confounders such as prior ability remain viable rival explanations. One possible explanation, derived from the author's analysis, is that this aligns with Self-Determination Theory, which posits that supportive environments enhance competence and autonomy. Source (Beimel et al., 2025)'s findings are consistent with SDT principles. One controlled experiment found that GenAI-supported learning enhanced student motivation and academic performance, though it did not always lead to significant achievement differences (Beimel et al., 2025). The current study extends this by suggesting that when engagement rises, performance follows, even if the direct instructional effect of the medium is neutral.

Satisfaction ratings diverged sharply between the two groups. The chi-square test indicated a strong association between group and satisfaction ($\chi^2(3, N = 124) = 20.91$, $p < .001$, Cramér's V = .41). Most AI-tutored students rated their experience as high satisfaction (4 or 5), while human-instructed students clustered around lower ratings (2 or 3). Specifically, 54.8% of the AI group chose "4" and 19.4% chose "5", compared to 27.4% and 8.1% respectively for the human group. Conversely, 48.4% of human-instructed students selected "3", and 16.1% selected "2". This disparity challenges the notion that human interaction inherently leads to higher satisfaction. While interviews in other studies revealed that students preferred instructor-led sessions for emotional engagement and support with complex problems (Beimel et al., 2025), the quantitative survey here favored the AI tool.

One rival explanation for the satisfaction gap is the novelty effect. Students interacting with a new technology may report higher satisfaction due to curiosity or excitement, independent of actual learning value. Additionally, the AI tool provided consistent availability, removing the logistical friction of scheduling office hours. A study on LMS-integrated AI assistants noted that integration improved usability and student satisfaction (Mzwri & Turcsányi-Szabo, 2026). If the AI tutor in this study offered seamless access, students may have valued the convenience over the depth of human connection. However, this convenience comes with risks. Concerns about over-reliance on AI and reduced student-teacher interaction emerged in similar contexts. The

high satisfaction in the AI group might mask a lack of critical mentorship, which is harder to capture in a single post-course survey.

The correlation between engagement and quiz performance ($r = .66$, $p < .001$) highlights the role of behavioral involvement in learning outcomes. Weekly sessions, however, were not a significant predictor ($B = -0.02$, $p = .952$). This implies that the quality of engagement mattered more than the quantity of time spent. Students who were mentally involved performed better, regardless of how many times they logged in. This finding supports the view that effective learning environments must foster active cognitive processing rather than just passive attendance. The AI tool seemingly facilitated this by providing immediate, low-stakes opportunities for interaction.

Several limitations constrain the generalizability of these conclusions. The sample size of 124 students limits the power to detect subtle effects. The study focused on introductory computing, where foundational skills are acquired. As noted in prior research, AI tools may be less effective for advanced topics requiring deep conceptual understanding or complex debugging assistance (Ahmed et al., 2024), (Yong et al., 2026). ChatGPT, for example, showed weaknesses in resolving complex student confusion and adapting responses to individual needs (Ahmed et al., 2024). Therefore, the benefits observed here might diminish as course difficulty increases. Furthermore, the reliance on self-reported engagement and satisfaction metrics introduces potential bias. Objective measures of learning, such as long-term retention tests, were not included.

In summary, the study found that generative AI tutoring outperformed human instruction in engagement, quiz performance, and satisfaction for introductory computing students. The performance advantage appears mediated by engagement, suggesting that the AI tool's primary benefit is motivational rather than purely instructional. Satisfaction was notably higher for AI users, possibly due to reduced social friction and increased accessibility. These findings contribute to the debate on AI in education by highlighting its potential to enhance student agency in early-stage learning. Future research should investigate whether these effects persist in advanced courses and examine the long-term impact on critical thinking skills. The results advocate for a blended approach, where AI handles routine feedback and humans provide complex mentorship, balancing efficiency with pedagogical integrity.

6 Conclusion

6.1 Synthesis of Findings and Theoretical Implications

This study compared first-year computing students taught with a generative AI tutor against those taught by a human instructor, measuring engagement, quiz performance, and satisfaction. The results indicate that the AI-supported cohort showed significant advantages in engagement and satisfaction, while noting that quiz performance differences were accounted for by engagement rather than direct instruction effects. Students in the AI condition reported higher engagement (M = 75.35) than those in the human condition (M = 66.75), a statistically significant difference of 8.60 points [analysis 75]. Quiz averages were also higher for the AI group (M = 72.23) compared to the human group (M = 68.81), though this gap was smaller at 3.42 points [analysis 76]. Satisfaction diverged most sharply: the chi-square test confirmed a strong association between teaching modality and student satisfaction ratings, with Cramér's V = .41 [analysis 79]. These findings suggest that, within the specific context of introductory computing courses, generative AI tutoring may enhance immediate learning experiences and perceived support more effectively than traditional instruction alone.

The theoretical implications align with Self-Determination Theory (SDT), which posits that autonomy, competence, and relatedness drive motivation. Prior research using SDT frameworks found that GenAI tools significantly enhanced autonomy through personalized pacing (Beimel et al., 2025). This study extends that observation by demonstrating that such autonomy correlates with higher engagement scores. However, the same literature notes limitations in supporting "relatedness," or emotional connection, where students often prefer human instructors for complex problem-solving (Beimel et al., 2025). Our data complicates this picture: while the AI group showed higher overall satisfaction, the distribution reveals nuance. A majority of AI-tutored students rated their experience as a 4 or 5 on the satisfaction scale, whereas the most common rating for human-tutored students was 3 [analysis 79]. One possible explanation is that for introductory material, the efficiency and availability of AI are consistent with satisfying students' need for competence and autonomy sufficiently to outweigh the lack of deep interpersonal relatedness.

It is critical to note that these outcomes do not prove AI superiority in all educational contexts. The design involved one sample size (n = 124) and a single course setting. The observed benefits may be specific to the novelty of the technology or the nature of introductory content. Comparative studies of intermediate computing courses (e.g., CS2) have examined the effects of human support alongside generative AI tools (Yong et al., 2026). In our introductory context,

the AI tutor likely provided sufficient scaffolding for foundational concepts, allowing students to focus on practice rather than navigation. Thus, the contribution of this thesis is limited to identifying conditions under which AI tutoring enhances engagement and satisfaction in early-stage computing education, rather than establishing a universal replacement for human instruction.

6.2 Mechanisms of Engagement and Performance

The relationship between engagement and academic performance offers insight into how AI tutoring influences learning outcomes. Pearson correlation analysis revealed a strong positive association between engagement scores and quiz averages across the entire sample ($r = .66$, $p < .001$) [analysis 77]. This indicates that students who were more engaged with the course material, regardless of the instructional method, tended to perform better on assessments. When controlling for engagement and weekly session frequency in an OLS regression, the effect of the teaching group on quiz performance became non-significant ($B = 1.06$, $p = .408$) [analysis 78]. This finding suggests that the slight advantage in quiz scores observed in the bivariate comparison [analysis 76] was largely explained by higher engagement levels rather than a direct causal effect of the AI tutor on knowledge acquisition.

This attenuation of the direct effect is consistent with the possibility that AI tools function primarily as engagement enhancers in introductory settings. Previous work on intelligent tutoring systems has shown that chatbot interaction is positively associated with behavioral and emotional engagement, but not necessarily cognitive engagement (Bai et al., 2025). Our data mirrors this distinction: the AI group's higher engagement likely reflects increased behavioral participation (e.g., logging in, asking questions) and emotional comfort (reduced anxiety about asking "stupid" questions). Source (Mzwri & Turcsányi-Szabo, 2026) discusses student satisfaction and engagement via an LMS assistant; it is the thesis author's inference that this reduction in hesitation to ask questions is a benefit associated with such integration. By lowering the social cost of inquiry, the AI tutor may have encouraged more frequent practice and feedback loops, thereby boosting engagement scores. Since engagement predicts quiz performance, this mechanism explains the modest gains in assessment outcomes without implying that the AI tutor taught concepts more effectively than the human instructor.

A rival explanation for the higher engagement in the AI group is the Hawthorne effect, where participants alter behavior due to awareness of being studied. Given that the AI tool represented a novel intervention, students may have interacted with it more intensively simply because it was new. Another possibility is selection bias, although the quasi-experimental nature of the grouping, based on enrollment section rather than random assignment, presents a limitation regarding potential non-random differences between cohorts. The consistency of the engagement-performance link in the combined sample reinforces the idea that active involvement drives learning, and the observed association between group assignment and higher engagement

scores suggests that the AI environment was linked to greater involvement than the traditional classroom structure.

6.3 Satisfaction Dynamics and Human-AI Collaboration

Student satisfaction varied significantly by instructional mode, revealing distinct preferences for different types of support. The crosstabulation shows that 74.2% of AI-tutored students rated their satisfaction as 4 or 5, compared to human-tutored students [analysis 79]. Conversely, 16.1% of human-tutored students gave the lowest possible rating (2), whereas only 1.6% of AI-tutored students did so. This disparity highlights that while human instruction provides depth, it may fail to meet the logistical and psychological needs of first-year students. Research indicates that students value AI for its timeliness and ability to provide contextually relevant answers to administrative and course-related queries (Mzwri & Turcsányi-Szabo, 2026). The AI tutor likely addressed these immediate needs more efficiently than human instructors, who are constrained by office hours and class sizes.

However, satisfaction with AI does not equate to a preference for total replacement of human contact. Qualitative insights from similar studies reveal that students appreciate AI's efficiency but still seek human instructors for emotional engagement and complex problem-solving (Beimel et al., 2025). In our quantitative data, the absence of qualitative interviews limits our ability to capture these nuanced feelings directly. Yet, the distribution of responses favored the AI tutor group, suggesting a more uniformly positive experience. This aligns with findings that human instruction showed higher long-term retention in their study, but AI-driven teaching excels in providing personalized feedback and adapting to learning pace (Sultana et al., 2025).

The integration of AI and human roles appears to be the optimal path forward. Studies comparing AI, visualization, and human support in CS2 courses found that while AI boosted self-efficacy, human tutoring yielded higher learning outcomes for advanced topics (Yong et al., 2026). Our introductory course results show AI leading in both engagement and satisfaction, suggesting that the balance shifts toward AI support when the material is foundational. For educators, this implies that AI should handle routine inquiries and basic concept reinforcement, freeing human instructors to focus on higher-order thinking and mentorship. The concern about over-reliance on AI and reduced student-teacher interaction remains valid (Mzwri & Turcsányi-Szabo, 2026), but our data suggests that in the initial stages of computing education, the benefits of accessible, judgment-free support outweigh the costs of diminished human contact.

6.4 Limitations and Boundary Conditions

Several limitations constrain the generalizability of these findings. First, the study relied on a single institution and a specific introductory computing curriculum. The content difficulty level is a critical boundary condition; prior research indicates that GenAI tutoring yields lower

learning outcomes in advanced CS2 courses compared to human support. In intermediate computing courses (e.g., CS2), human live tutoring produced higher learning outcomes than GenAI, despite GenAI generating larger increases in self-efficacy (Yong et al., 2026). Therefore, the positive results for AI in this study cannot be extrapolated to upper-division computer science courses where conceptual reasoning and debugging complex systems are paramount. The AI tutor's strength lies in pattern recognition and rapid feedback on syntax or basic logic, skills central to introductory programming but peripheral to advanced theory.

Second, the measurement of engagement and satisfaction used self-reported scales, which are subject to response biases. While the correlation between engagement and quiz performance adds validity, the constructs themselves may overlap. Students who enjoy the course may report higher engagement and perform better due to intrinsic motivation, confounding the effect of the teaching method. Additionally, the study did not track longitudinal retention. Immediate post-course satisfaction and quiz scores do not guarantee long-term knowledge retention or skill application. Literature suggests that human teaching facilitates long-term retention through real-life contexts and emotional connection (Sultana et al., 2025), a dimension our study could not assess. Without follow-up data, we cannot determine if the AI group's initial advantages persisted beyond the semester.

Third, the technical capabilities of the AI tutor were fixed during the study period. Generative AI models evolve rapidly, and newer versions may exhibit different behaviors regarding hallucination or pedagogical soundness. Some evaluations of ChatGPT as a teaching assistant highlighted weaknesses in resolving complex confusion and adapting to individual needs (Ahmed et al., 2024). If the AI tutor in this study struggled with these areas, the observed engagement might reflect students' willingness to overlook flaws due to convenience, rather than true pedagogical efficacy. Future iterations of AI tutors with improved multimodal capabilities (Bai et al., 2025) or better integration with course content (Mzwri & Turcsányi-Szabo, 2026) may yield different engagement profiles. Consequently, these findings represent a snapshot of current AI technology's impact on introductory computing education, not a permanent verdict on its potential.

6.5 Future Research Agenda

Future studies should address the limitations identified here by expanding the scope and duration of investigation. A primary recommendation is to replicate this experiment in advanced computing courses, such as data structures or algorithms, to test whether the engagement-satisfaction advantage of AI persists as material difficulty increases. Comparative designs that include hybrid models, where AI handles routine queries and humans focus on complex mentoring, would clarify the optimal division of labor. Existing reviews suggest that combining human and AI teaching offers the best outcomes for both one-on-one learning and emotional connection

(Sultana et al., 2025). Empirical testing of such blended frameworks is necessary to move beyond binary comparisons.

Longitudinal tracking of student performance is another critical direction. Measuring retention rates six months or one year after the course would reveal whether AI-enhanced engagement translates into durable learning. Current evidence indicates that while AI boosts short-term confidence, human tutoring supports deeper conceptual understanding in later stages (Yong et al., 2026). Tracking cohorts through subsequent semesters could identify if early AI exposure creates dependencies that hinder independent problem-solving later. Furthermore, qualitative investigations into *why* students prefer AI for certain tasks are needed. Understanding the specific features that drive satisfaction, such as the reduction of anxiety (Mzwri & Turcsányi-Szabo, 2026) or the enhancement of autonomy through personalized pacing (Beimel et al., 2025), will inform better tool design.

Finally, research must examine the ethical and equity implications of widespread AI adoption in computing education. While some studies suggest AI can close performance gaps for learners without prior coding experience, others warn of superficial learning due to overreliance. Investigating how AI tutoring affects students from diverse backgrounds, particularly those with limited digital literacy, is essential. Ensuring that AI tools complement rather than replace human mentorship requires careful policy development. Educators and curriculum designers need evidence-based frameworks that prioritize prompt engineering skills and human oversight (Elnaffar et al., 2025). By addressing these questions, future research can refine the role of generative AI in creating inclusive, effective, and sustainable computing education environments.

References

Ahmed, Z., Shanto, S., & Jony, A. (2024). Potentiality of generative AI tools in higher education: Evaluating ChatGPT's viability as a teaching assistant for introductory programming courses. *STEM education*. https://doi.org/10.3934/steme.2024011

Bai, H., Lui, W., & Khiatani, P. (2025). Promoting student engagement with GPTutor: An intelligent tutoring system powered by generative AI. *Journal of Educational Technology in Higher*. https://doi.org/10.1186/s41239-025-00571-9

Beimel, D., Amzalag, M., Zviel-Girshin, R., & Voloch, N. (2025). Blending Generative AI and Instructor-Led Learning: Empirical Insights on Student Motivation, Learning Experience, and Academic Performance in Higher Education. *Education Sciences*. https://doi.org/10.3390/educsci15111480

Elnaffar, S., Rashidi, F., & Abualkishik, A. (2025). Teaching with AI: A systematic review of chatbots, generative tools, and tutoring systems in programming education. *International Journal of Learning, Teaching and Educational Research*. https://doi.org/10.26803/ijlter.25.1.1

Jacobs, S., Haas, J., & Kiesler, N. (2025). Student Engagement with GenAI's Tutoring Feedback: A Mixed Methods Study. *Proceedings of the 25th Koli Calling International Conference on Computing Education Research*. https://doi.org/10.1145/3769994.3770034

McCulloh, I., Rodriguez, P., Kumar, S., Gupta, M., Sharma, V. R., Johnson, B., & Johnson, A. N. (2025). Generative AI in computer science education: Accelerating python learning with ChatGPT. *arXiv preprint arXiv:2505.20329*. https://doi.org/10.48550/arxiv.2505.20329

Mellado, R., & Cubillos, C. (2025). Can generative artificial intelligence outperform self-instructional learning in computer programming?: Impact on Motivation and Knowledge Acquisition. *Applied Sciences*. https://doi.org/10.3390/app15115867

Mzwri, K., & Turcsányi-Szabo, M. (2026). Bridging LMS and generative AI: dynamic course content integration (DCCI) for enhancing student satisfaction and engagement via the ask ME assistant. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-025-00367-w

Sultana, D. N., Khan, D. N. U. A., Malik, D. M., & Rasheed, T. (2025). Analyzing Differences in Student Understanding and Retention When Taught by AI Versus Human Instructors. *The Critical Review of Social Sciences Studies*. https://doi.org/10.59075/tf8dkx85

Yong, Q., Estey, A., & Nacenta, M. (2026). Characterization and Effects of CS2 Learning with GenAI, Visualization, and Human Support. *Proceedings of the 2026 ACM Conference on International Computing Education Research Vol. 1*. https://doi.org/10.1145/3765964.3811663

Appendix A

descriptives: engagement_score, quiz_avg, weekly_sessions, satisfaction

Method: Means, standard deviations (sample), medians, quartiles, and ranges for numeric variables; frequencies for categorical variables. Missing values are excluded per variable.

Descriptive statistics (N = 124 rows): engagement_score: M = 71.05, SD = 9.72 (n = 124); quiz_avg: M = 70.52, SD = 7.34 (n = 124); weekly_sessions: M = 3.73, SD = 1.48 (n = 124); satisfaction: M = 3.60, SD = 0.84 (n = 124).

Descriptive statistics, presented as Table 4.1 in Chapter 4.

Means with standard deviation: engagement_score 71.05 ± 9.72; quiz_avg 70.52 ± 7.34; weekly_sessions 3.73 ± 1.48; satisfaction 3.60 ± 0.84. Error bars show one standard deviation, presented as Figure 4.1 in Chapter 4.

Appendix B

t-test: engagement_score by group

Method: Welch's independent-samples t-test (unequal variances), two-sided; Cohen's d uses the pooled standard deviation; 95% CI of the mean difference.

Welch's independent-samples t-test (unequal variances) found a statistically significant difference in engagement_score between ai_tutor (M = 75.35, SD = 8.85, n = 62) and human_instructor (M = 66.75, SD = 8.64, n = 62), t(121.9) = 5.48, p < .001, d = 0.98; the ai_tutor group scored higher by 8.60 points (95% CI [5.49, 11.71]).

Group statistics, presented as Table 4.2 in Chapter 4.

engagement_score by group: ai_tutor 75.35 ± 8.85; human_instructor 66.75 ± 8.64. Error bars show one standard deviation, presented as Figure 4.2 in Chapter 4.

Appendix C

t-test: quiz_avg by group

Method: Welch's independent-samples t-test (unequal variances), two-sided; Cohen's d uses the pooled standard deviation; 95% CI of the mean difference.

Welch's independent-samples t-test (unequal variances) found a statistically significant difference in quiz_avg between ai_tutor (M = 72.23, SD = 7.13, n = 62) and human_instructor (M = 68.81, SD = 7.20, n = 62), t(122.0) = 2.65, p = .009, d = 0.48; the ai_tutor group scored higher by 3.42 points (95% CI [0.87, 5.96]).

Group statistics, presented as Table 4.3 in Chapter 4.

quiz_avg by group: ai_tutor 72.23 ± 7.13; human_instructor 68.81 ± 7.20. Error bars show one standard deviation, presented as Figure 4.3 in Chapter 4.

Appendix D

correlation: engagement_score, quiz_avg

Method: Pearson product-moment correlation (r) with two-sided p from $t = r\sqrt{((n-2)/(1-r^2))}$, and Spearman rank correlation (ρ); each pair uses its pairwise-complete cases.

Pearson correlations: engagement_score × quiz_avg, r = .66, p < .001 (n = 124).

Correlations, presented as Table 4.4 in Chapter 4.

quiz_avg against engagement_score, n = 124. Fitted line: quiz_avg = 35.02 + 0.50 × engagement_score, presented as Figure 4.4 in Chapter 4.

Appendix E

regression: quiz_avg ~ engagement_score + weekly_sessions + group

Method: Ordinary least squares; categorical predictors dummy-coded against their first level; listwise deletion of incomplete rows; two-sided t-tests on coefficients.

An OLS regression of quiz_avg on engagement_score, weekly_sessions, group significantly predicted the outcome, $F(3, 120) = 31.72$, $p < .001$, $R^2 = .44$ (adjusted $R^2 = .43$, $n = 124$); coefficients: engagement_score $B = 0.53$ ($SE = 0.06$), $p < .001$; weekly_sessions $B = -0.02$ ($SE = 0.41$), $p = .952$; group = human_instructor (vs ai_tutor) $B = 1.06$ ($SE = 1.28$), $p = .408$.

Coefficients, presented as Table 4.5 in Chapter 4.

Observed quiz_avg against Fitted quiz_avg, n = 124. Fitted line: Observed quiz_avg = 0.00 + 1.00 × Fitted quiz_avg, presented as Figure 4.5 in Chapter 4.

Appendix F

crosstab: group × satisfaction

Method: Contingency table with row percentages; Pearson chi-square test of independence; Cramér's V as effect size.

A chi-square test of independence showed that group and satisfaction were associated, $\chi^2(3, N = 124) = 20.91$, $p < .001$, Cramér's $V = .41$.

Crosstab (n, row %), presented as Table 4.6 in Chapter 4.

group × satisfaction (counts): ai_tutor / 4 34.00; ai_tutor / 3 15.00; ai_tutor / 5 12.00; ai_tutor / 2 1.00; human_instructor / 3 30.00; human_instructor / 4 17.00; human_instructor / 2 10.00; human_instructor / 5 5.00, presented as Figure 4.6 in Chapter 4.